人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-01

サンプリングに基づくLODの構造推定に関する基礎的検討

Investigation on LOD Structure Estimation Based on Sampling

矢部彩佳 *　高間康史
Ayaka Yabe , Yasufumi Takama

首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

Abstract: 近年 LOD によるデータ公開が進められており，これらを活用したサービス開発なども期待されている．しかし，他者が公開したデータを利用する場合，データ構造が不明な場合があり，活用を阻害する一要因となっている．本稿では LOD を探索的にブラウズする作業を支援するシステムの実現を目的として，その要素技術となる LOD の構造推定に着目する．SPQRL クエリによるサンプリングに基づく推定方法に関する基礎的な検討を行った結果について報告する．

1　はじめに

本稿では，RDF (Resource Description Framework) で記述された LOD を探索的にブラウズする作業を支援するシステムの要素技術として，SPARQL クエリによるサンプリングに基づく LOD 構造の推定手法に関する基礎的な検定を行った結果について報告する．

近年，計算機で処理しやすい形式でデータを公開･共有する仕組みとして LOD (Linked Open Data) が注目されている．LOD は自分の手元にない外部リソースを扱えることが利点だが，他者が公開したデータを使用する場合，データ構造が不明という問題点がある．このため，探索的に LOD をブラウズし，その構造を把握する必要があると考える．

探索作業を支援するために，探索の起点として有効なリソースの抽出･提示を行う．起点として有効なノードを発見するためには，LOD のデータ構造を分析する必要があるが，全データを取得して分析を行うのでは，外部リソースの活用という LOD の利点が生かせないと考える．そこで本稿では，LOD データを SPARQL クエリを用いてサンプリングし，LOD の構造推定を試みる．

RDF とは，リソースの関係を主語 (subject)・述語 (predicate)・目的語 (object) の 3 つの要素 (トリプル) を用いて表現するデータモデルであり，データセットは，主語と目的語をノード，述語をエッジとするグラフ構造で表現される．本稿ではこの構造を用いてサンプリングを行う．

現在日本で公開されている LOD データを調査したところ，Excel 等のテーブルデータを RDF データに変換したものが多く発見された．そこで本稿では LOD を表構造を持つもの（テーブル型）とそれ以外に分類し，サンプリングにより両者を区別可能であるかを検証し，その結果に基づきサンプリングによる構造推定の可能性について考察する．

*連絡先：首都大学東京大学院 システムデザイン研究科
〒 191-0065 東京都日野市旭ヶ丘 6-6
E-mail: yabe-ayaka@ed.tmu.ac.jp

2　関連研究

2.1　表データの RDF データ化ツール

現在日本で公開されている LOD データには表形式のものが多くある．その理由として，すでに表形式で管理していたデータを公開する場合が多いことと，表形式データの RDF データ変換ツール・サービスが整備されていることが挙げられる．前者は，総務省や市町村が公開しているデータが該当する．後者に関しては，オープンデータ活用支援プラットホーム LinkData.org[1] などが存在する．

LinkData.org は，データ・アプリ・アイディアの作成と公開を行う 4 つの Web サイトを提供しており，その中の LinkData[2] ではテーブルデータを RDF データに変換するサービスを提供している．RDF 変換用のテーブルデータの雛型を Web サイト上で作成．ダウンロードし，RDF データの主語・目的語にあたる部分を埋めてアップロードすることで RDF データに変換が可能である．このサービスにより手持ちのテーブル型データを気軽に RDF 化することができる．これらのツールを利用することにより，今後さらにテーブル型 RDF データが増加していくと推察できる．

[1] http://linkdata.org/
[2] http://linkdata.org/home

2.2 グラフ構造データの分析

近年，世界中に広く普及したSNS(ソーシャルネットワークサービス) や生体科学における遺伝子構造など，グラフ構造をもつデータの分析に関する研究がされている [3][4].

分析対象となるデータが巨大な場合，すべてのデータを分析することはコストや時間面から難しいという問題が存在する．この問題に対し仲前ら [1][2] は，巨大グラフデータから部分的にグラフを抽出する手法として，ランダムウォークサンプリングを改良したサンプリング方法を提案している．入次数の大きいノードを訪れやすいというランダムウォークの性質を考慮したIRW(In-Degree Weighted Random Walk)[1] は，入次数に偏らないサンプリングが可能となる．IRW は入次数がわかることを前提としているが，巨大なグラフデータに対して事前に入次数を調べることは現実的ではないことから，Reservoir を用いたIRW の改善版であるIRRW を提案し，入次数を前提条件としないランダムサンプリングを可能にしている [2].

2.3 RDF データ分析

RDF データは，通常 SPARQL[3] と呼ばれるクエリ言語によって検索が行われるが，適切なクエリを作成する為にはデータ構造の理解が不可欠となる．そこで，後藤ら [5] は探索的検索アプローチによって LOD を理解・利用する DashSearchLD というシステムを提案している．探索的検索とは，探索目的を少しずつ明確化しながら新しい知識を獲得していく学習や調査のような情報検索である．探索的閲覧によって検索空間を遷移しつつ，絞込み検索によって検索絞り込むという行為を繰り返すことにより，検索空間の理解と情報要求の具体化を行い，データ集合の理解に繋がるとしている．DashSearchLD には，SPARQL Endpoint 機能を持つエンドポイントウィジェットと，RDF データのプロパティとその値を表示するメタデータウィジェットがあり，ユーザはこれらのウィジェットをマウスによって操作することで，SPARQL クエリを用いずにデータの探索的検索やプロパティ情報の獲得が可能となる．また，田代ら [6] は，RDF の特徴を考慮したデータ分析支援ツールとして，(1) 共通の述語を持つ主語の抽出・テーブル作成を行うツール．(2) 複数エンドポイント間の共通リソースの抽出を行うツール．(3) 時間情報に基づくデータ分析支援ツールを提案している．(1) は，最大公約数的に共通の述語を持つ主語の抽出を行い，行を主語，列を述語としたテーブルの出力を行う．(2) は，2 つの SPARQL エンドポイント間で共通するリソースを抽出することで，異なる LOD の連結可能性を検討する作業を支援する．(3) は，統計データやログデータのような RDF データから時系列データを抽出し，ヒストグラムとして可視化を行う．

RDF データを活用する上で必要となるのが重要リソースの把握である．SPARQL 検索によって，RDF データの一部を簡単に抽出することができるが，SPARQL は抽出したリソースを重要度の高い順にランク付けする機能を持っていない．検索結果が大量にあった場合，さらに情報を絞り込むためユーザの要望に応じたリソースのランキングを提供することは有用であるとして，一瀬ら [7][8] は DBpedia を対象に SPARQL 検索によって得られたリソースを，グラフ構造から重要度評価を行う PageRank アルゴリズム用い，ランク付けする方法を提案している．

3 LOD 構造判定方法

3.1 テーブル型と非テーブル型の判定

前述の通り，現在公開されている RDF データには，テーブル型データを RDF データに変換したデータとそうでないデータが存在する．前者は市町村が公開しているデータに多く見られる．一方，DBpedia の様な，多種多様なリソースを含む RDF データの場合には，テーブル型をとらないと仮定する．この仮定に基づき本稿では，RDF データがテーブル型か否かを判別することを目的とする．

グラフ構造を分析する方法としては，ランダムウォークサンプリングなどの方法が知られている [1][2]．このような探索を行う方法は，複雑ネットワークを構成しているデータには有効だが，テーブル型のデータ (図 1) の場合には，ほとんどの目的語がリテラル（数値あるいは文字列）であることが多いことから，ランダムウォーク等の探索方法を用いてもすぐに行き止まるため，有効に機能しないと考える．

テーブル型データの特徴として，本稿では以下の 4 点に着目する．

1. 同じプロパティが複数存在
2. 目的語として，リテラルまたは出次数が 0 のリソースを持つため，探索をしてもすぐに行き止まる
3. 各リソースの出次数が揃いやすい
4. 各プロパティはリソース毎に 1 回ずつ出現する

これらの特徴に基づきテーブル型か否かの判別を行い，テーブル型ではない場合のみ探索を行うことで，各 RDF データに対し効率的に起点ノードを発見することが可能と考える．

[3]http://www.w3.org/TR/rdf-sparql-query

名前	性別	年齢	職業	勤め先HP	住所	電話番号	E-mail
http://…A	M	30	○○	http://…	東京	111-1111	…@..jp
http://…B	F	30	△△	http://…	神奈川	222-2222	…@..jp
http://…C	M	30	□×	http://…	千葉	333-3333	…@..jp
http://…D	F	30	▽◇	http://…	埼玉	444-4444	…@..jp

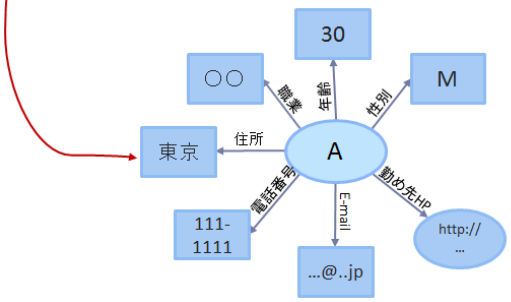


図 1: テーブル型データの RDF グラフ

3.2 データの抽出方法

本稿ではデータ型判断のためのデータ抽出法として，以下の手順をとる．

1. 対象となる RDF データからランダムに主語リソースを抽出し，探索の起点とする．
2. 起点から最良優先探索を行い，取得したノードについて以下の情報を記録する．
   - 出次数
   - 探索の STEP 数
   - プロパティ
3. 起点が持つ各プロパティの出現回数を求める．

ステップ 2 において，最良優先探索に用いるヒューリスティック関数として各リソース出次数を用い，出次数の大きいノードを優先的に探索する．

図 2 に示す例で A を起点とすると，子ノード B, C, D 中で出次数最大（2）の B を選択し，これを主語とするトリプルを SPARQL により求める．

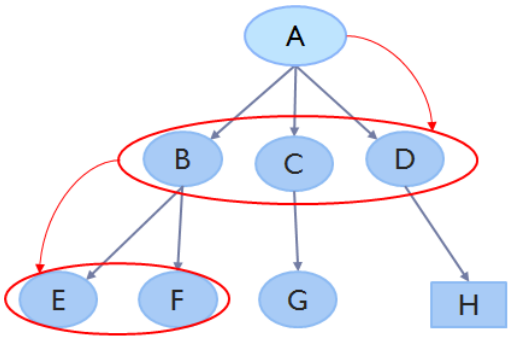


図 2: (例) 最良優先探索

4 型判定に関する予備実験

4.1 実験概要

本稿では，表 1 のデータセットを対象に，以下の 2 点に関する調査を目的として実験を行う．

- 調査 1：起点のプロパティに関する調査
- 調査 2：ステップ数の調査

調査 1 では，プログラム 1 回の試行でデータセットからランダムに起点リソースを 10 個抽出する．各起点に対し 3.2 節で述べた手順でプロパティを取得し，プロパティの出現回数を計算する．各データセットにおける調査回数は表 2 の通りである．非テーブル型はテーブル型に比べ，構造の特徴がわかりづらいため試行を 2 倍行った．

調査 2 では，最良優先探索により各起点から何ステップ進めるかを調査する．DBpedia Japanese 以外の 3 つのデータに関しては，探索可能なノードがなくなるまで探索を続け，DBpedia Japanese に関しては探索の上限を 30 ステップとした．

表 1 に示すデータセットにおいて，テーブル型と想定されるデータとして横手市 AED 設置場所 [4] 及び神奈川名所 LOD データセット [5]，非テーブル型と想定されるものとして横手市 AED 設置場所加工及び DBpedia Japanese を対象データセットしてそれぞれ選んでいる．

横手市 AED 設置場所加工データは，横手市 AED 設置場所データを元に，675 トリプルを削除し各主語リソースの出次数をまばらにした後，人工データ 67 トリプルを追加した．DBpedia Japanese[6] は，2013 年 9 月 4 日の以前に公開されたデータを使用した．

表 1: 使用データセット

データ	総トリプル数	主語リソース数
横手市 AED 設置場所	1,252	113
神奈川名所 LOD	451	45
横手市 AED 設置場所加工	644	140
DBpedia Japanese	32,633,660	3,626,642

表 2: 調査 1

データ	抽出起点数×試行回数
横手市 AED 設置場所	10 × 5
神奈川名所 LOD データセット	10 × 5
横手市 AED 設置場所加工	10 × 10
DBpedia Japanese	10 × 10

[4] 横手市情報政策課:http://linkdata.org/work/rdf1s843i
[5] kamogawa, SayokoShimoyama:http://linkdata.org/work/rdf1s2537i
[6] http://ja.dbpedia.org/

4.2 実験結果

図 3, 4, 5 に実験 1 の結果，表 3 に実験 2 の結果を示す．図は，縦軸がプロパティの種類数，横軸がプロパティの出現回数である．また，表 3 に示す平均及び標準偏差は，標本についてのものであり，母集団の不偏推定量ではない．

テーブル型である横手市 AED 設置場所データ (図 3) と神奈川名所 LOD データセット（図 4）の結果では，両者とも出現回数が 9 回もしくは 10 回のプロパティ数が多くなっている．

横手市 AED 設置場所データに関して，取得したリソースを観察すると，3 回目の試行以外の起点リソースは全て AED に関するリソースであり，共通プロパティが数多く見られた．そのため，10 個の起点に対し 10 回出現したプロパティの種類が多くなっている．これは 3.1 節に示したテーブル型データの特徴 1 に該当する．また，これらのプロパティは各起点リソースに 1 回ずつ出現していたため，特徴 4 にも該当する．3 回目の試行に関しては，10 個の起点の中で 1 つだけ AED に関するものではなく E-mail に関するリソースだったため，AED リソースとは異なるプロパティを持っていたた．そのため，他の試行とは異なり出現回数 9 回のプロパティ数が多くなっている．9 または 10 回出現したプロパティは，AED の名前，設置場所の住所，設置場所の郵便番号などで，すべての AED リソースで出現していた．出現頻度の少なかったプロパティは設置場所施設の開く時間・閉まる時間，外部リソースへのリンクなど，全ての AED リソースが持っているわけではない要素であった．また，今回抽出した全 50 個のリソースそれぞれの出次数は AED リソースで 10〜13，E-mail リソースは 2 であり，AED リソースは特徴 3 を満たしている．

表 3 より，調査 2 に関してはステップ数が全て 1 で終了したことがわかる．このことから 3.1 節で述べた「探索してもすぐ行き止まる」という特徴 2 が満たされていることがわかる．以上より，横手市 AED 設置場所データは，3.1 節で挙げたテーブル型データの特徴を満たすことがわかる．

神奈川名所 LOD データセットに関しては，探索した全 50 個の主語リソース中 49 個は名所に関するリソース，残り 1 個は動画情報を定義するリソースであった．各名所リソースにて，共通のプロパティが多く存在したため，横手市 AED 設置場所データと同様に 10 個の起点に対し，9 回もしくは 10 回出現したプロパティ数が多くなった．また，表 3 より，ステップ数も全て 1 であり，テーブル型データの特徴を満たしていると言える．

図 5 より，横手市 AED 設置場所加工データは，横手市 AED 設置場所データや神奈川名所 LOD データセットとは異なり，1 度しか出現しないプロパティが数多いことがわかる．これは複数のリソースが共通して持つプロパティが少ないということを意味している．複数回出現したプロパティも存在しているが，これは図 3 に示したとおり，加工前のデータが共通プロパティを多く含んでいたためである．また，表 3 より，ステップ数が必ずしも 1 ではないことがわかる．さらに，出次数の標準偏差がテーブル型データと判断した 2 つのデータよりも大きいこともわかる．以上よりこのデータはテーブル型ではないと判断できる．

DBpedia Japanese(図 6) では，共通プロパティが少ない傾向が横手市 AED 加工データよりも顕著に現れている．抽出された起点リソースは，人名や地名，学校名や神社などの施設が主である．同じ人名でも国籍や職業の違いから様々なプロパティが出現したため，出現回数の少ないプロパティが多く現れた．

複数回出現したプロパティは 2 種に大別される．1 つは大抵のリソースが保有するプロパティで,「http://www.w3.org/1999/02/22-rdf-syntax-ns#type(リソースのタイプ)」や「http://dbpedia.org/ontology/wikiPageID (wikipedia のページ ID)」などが該当する．もう 1 つは，ある 1 つのリソースが同じプロパティをいくつも持っている場合である．DBpedia Japanese には後者のパターンが多く見られた．また，表 3 より，出次数の標準偏差が他のデータよりもかなり大きいこともわかる．前述の通り DBpedia は大規模であるため探索の上限を 30 としたが，100 個の起点リソースのうち，ステップ数が 30 以内で終了したものは 7 個であった．このことより，DBpedia Japanese のデータセット内で多くのリンク関係が存在すると言える．以上のことから，本実験結果では DBpedia Japanese はテーブル型ではないと判断できる．

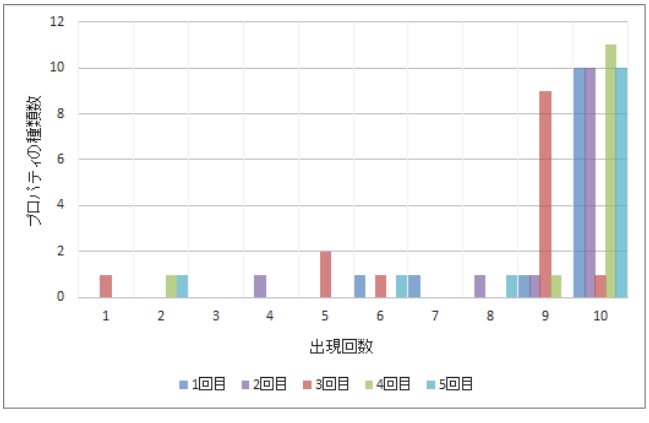


図 3: 調査 1 の結果：横手市 AED 設置場所

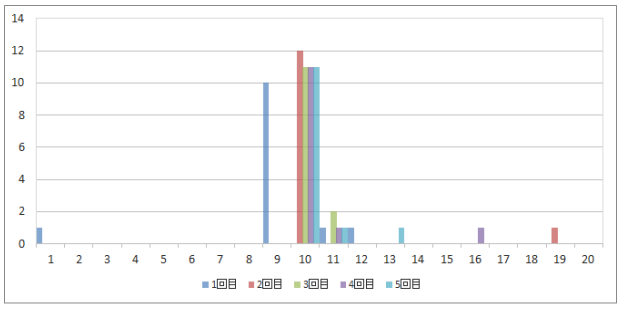


図 4: 調査 1 の結果：神奈川名所 LOD データセット

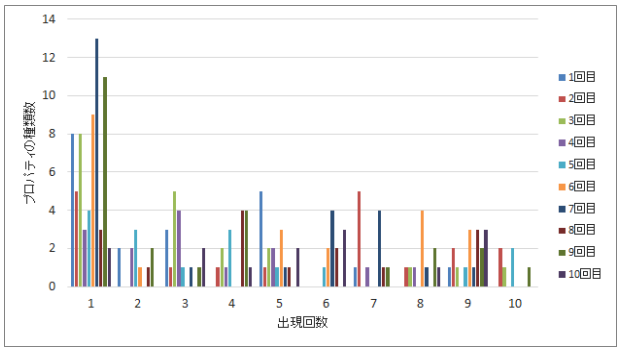


図 5: 調査 1 の結果：横手市 AED 設置場所加工

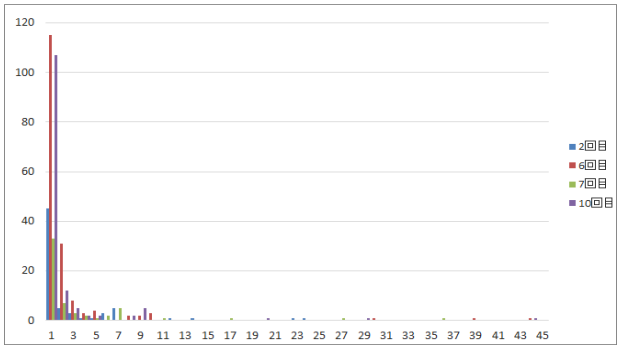


図 6: 調査 1 の結果：DBpedia Japanese

表 3: 調査 2 の結果

	出次数		STEP 数	
データセット	平均	標準偏差	平均	標準偏差
横手市 AED	11.76	1.59	1	0
神奈川名所 LOD	13.32	1.81	1	0
横手市 AED 加工	7.42	3.47	1.29	0.791
DBpedia Japanese	28.5	24.95	-	-

4.3 DBpedia Japanese に関する考察

DBpedia Japanese は日本で公開されている LOD の中で巨大なデータセットの一つであり，各データセットを繋ぐハブのような役割を果たしている[7]．しかし大規模な分，どの様なデータが含まれているかを知ることは困難であるため，その構造を把握することは有用であると考える．本節では，予備実験を通じて観察された DBpedia Japanese の特徴的な構造について考察する．

人名や地名などの様々なリソース“http://ja.dbpedia.org/resource/○○”が持つ，共通プロパティ“http://xmlns.com/foaf/0.1/isPrimaryTopicOf”はプロパティ“http:// xmlns.com/foaf/0.1/primaryTopic”と図 7 のように相互関係を持っていることがわかった．“primaryTopic”の主語リソースは，“http://ja.wikipedia.org/wiki/○○”であり，wikipedia のページである．また，目的語は“http://ja.dbpedia.org/resource/○○”である．両プロパティは同じ関係を逆向きに表現したものであるため，この様な循環的構造をとっている．このように構造が決まっているプロパティを探索過程で発見できれば，DBpedia Japanese の構造理解に役立てることができると考える．

また，本実験過程で得られた，“http://ja.dbpedia.org/resource/Category:○○”が主語として出現する場合，特有のプロパティを持つことがわかった．表 4 に示すプロパティ“core#related”は関連するカテゴリ“Category:○○”が目的語となる．プロパティ“core#broader”は主語リソースを包含する上位カテゴリ“Category:○○”が目的語となる．“pref#Label”は主語リソースのラベル (リテラル) が目的語となる．リソースによって出次数はばらつきがあるものの，プロパティの種類数にはあまり差異がなかったため“Category:○○”を主語としたときのトリプル構造を大雑把に表型と捉えることは可能と考えられる．すなわち，DBpedia Japanese には同種の情報が構成するテーブル型データが複数含まれ，それらの間につながりがあることが想定される．この点については今後調査を行う必要があると考える．

30 ステップ以内に探索が終了しなかったリソースに関して，展開されるリソースにある一定のパターンが存在する事が観察された．よく現れるパターンとしては，学問に関するリソースが連続して展開されるパターン，都道府県に関するリソースが連続するパターン，日本の歴代総理大臣が連続するパターン，欧米の地名から各国の大統領へ遷移するパターンなどが観察された．それらは元の起点リソースが一見全然関係ないものでも現れた．例えば起点リソース“http://ja.dbpedia.org/resource/極道の妻たち_危険な賭け”の場合，俳優や映画といったリソースから世界観，宇宙論・宇宙物理学，物理学…と遷移した．この様なパターンが頻繁に観察された理由として，本稿では出次数の大きいノードを選んでいく探索を行ったため，一度出次数の大きいノードが展

[7] http://linkedopendata.jp/?p=411

開されると，後は毎回同じパスが展開されることが挙げられる．例えば，DBpedia Japanese では,「日本」というリソースを目的語に持つリソースが多いため，このリソースが探索の過程で現れる確率は高く，さらに出次数が 133 と大きいため，展開されやすい．このため，その後に展開されるパターンが類似したものになる場合が多く発生した．DBpedia Japanese の構造を探る上でこのような決まったパターンが多く出現してしまうと，探索の妨げになる可能性があるため，今後対処法を検討する必要がある．

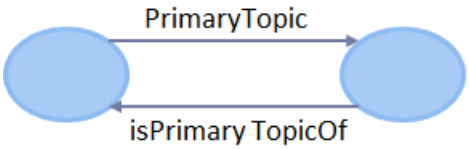


図 7: 相互リンクを持つリソース

表 4: “ Category:○○ ”特有のプロパティ

http://www.w3.org/2004/02/skos/core#related
http://www.w3.org/2004/02/skos/core#broader
http://www.w3.org/2004/02/skos/core#prefLabel

5　まとめ

本稿では，RDF で記述された LOD を探索的にブラウズする作業を支援するシステムの要素技術として，SPARQL クエリによるサンプリングに基づく LOD 構造の推定手法に関する基礎的な検定を行った．

本実験では，3.1 節で述べたテーブル型データの特徴を元にデータがテーブル型であるか否かを判断したが，データセットによっては，ある主語リソースが同じプロパティを複数持つ場合も観測されたため，プロパティの出現回数だけでは型の判定が難しくなる可能性があると考える．よって型判定の精度を高めるためにはプロパティの出現率も考慮する必要がある．今回使用したテーブル型のデータは 1 つのテーブルデータを 1 つの RDF データに変換したものと想定されるため，ステップ数の平均や標準偏差の情報だけでもテーブル型判と断することができた．しかし，複数のテーブルデータを 1 つの RDF データにまとめたようなデータが存在し表と表の間に繋がりある場合，ステップ数が必ずしも 1 ではなくなる．そういったデータセットにおいて表部分をどう発見するか今後の課題である．また，推定精度と実行時間の関係についても調査を行い，サンプリングの起点とするリソース数について検討することも必要と考える．さらに，探索的ブラウズの起点として提示すべきノードについての検討も今後の課題である．

参考文献

[1] 仲前晋太郎, 成凱: Blog における話題分析のためのランダムサンプリング手法の提案, *DEIM Forum*, D3-5, 2010

[2] 仲前晋太郎, 成凱: Reservoir を用いた巨大グラフのランダムサンプリング, *DEIM Forum*,D3-2,2011

[3] 鹿島久嗣: ネットワーク構造予測, 人工知能学会論文誌, Vol.22, No.3, pp.344-351, 2007

[4] 安田雪, 松尾豊, 武田英明: リンクマイニングによる研究者ネットワークの抽出:成長プロセスと国内外からの見え方, 第 *21* 回人工知能学会全国大会, 1B2-8, 2007

[5] 後藤孝行, 濱崎雅弘, 武田英明: DashSearch LD: 探索的検索の Linked Data への適用, 第 *26* 回人工知能学会全国大会, 3C1-OS-13a-3, 2012

[6] 田代航一, 高間康史: RDF データベースを対象としたデータ分析支援ツールの提案, 第 *5* 回情報アクセスと可視化マイニング研究会, SIG-AM-05-02, pp7-12, 2013

[7] 一瀬詩織, 小林一郎, 岩爪道昭, 田中康司: DBpedia における SPARQL 検索結果のランキング手法, 第 *27* 回人工知能学会全国大会, 2N5-OS-21b-4, 2013

[8] 一瀬詩織, 小林一郎, 岩爪道昭, 田中康司: DBpedia を対象にしたリソースのランキング手法における一考察, 情報処理学会第 *75* 回全国大会, 4N-9,2013

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-02

協調的マルチビューに基づくインタラクティブ文書クラスタリングシステムの提案

利根川 拓馬* 高間 康史

Takuma Tonegawa, Yasufumi Takama

首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

Abstract: 本稿では，協調的マルチビューに基づくインタラクティブ文書クラスタリングシステムを提案する．提案システムでは，ユーザフィードバックをクラスタリング結果に反映するために，単語の重み調整に基づく手法を採用し，クラスタや文書，単語と言った異なるレベルの情報を効率的に提示するために協調的マルチビューを採用する．TETDM（Total Environment for Text Data Mining）を用いてプロトタイプシステムを実装し，評価実験を行った結果について示す．

1. はじめに

本稿では，協調的マルチビューに基づくインタラクティブ文書クラスタリングシステムを提案する．学術論文や最新のニュース記事などの文書データは，様々な知識を得るための重要なリソースである[1]．近年，それらの文書データに対して，話題検出・追跡や文書間の関係性の発見などの探索的データ分析を行う必要性が高まっている[9]．そのため，ユーザの探索データ分析を支援し，負担を軽減することを目的としたインタラクティブテキストマイニングシステムの研究が進められている．

探索的データ分析の代表的手法の一つにインタラクティブクラスタリング[2]がある．教師なし学習である通常のクラスタリングとは異なり，インタラクティブクラスタリングではユーザがオブジェクトをグループ化する際にいくつかの制約を与えるため，半教師あり学習と呼ばれる．これにより，ユーザの視点を反映させたクラスタリングを行うことができ，効率よくデータの分析を行うことが可能となる．しかし，インタラクティブクラスタリングシステムを開発する際には,「複数オブジェクトの情報をどのように表示するか」,「異種オブジェクト間の関係性をどのように表示するか」,「制約付きクラスタリングをどのように導入するか」といった問題が挙げられる．

* 連絡先:首都大学東京大学院システムデザイン研究科
〒191-0065 東京都日野市旭が丘 6-6
E-mail:ytakama@sd.tmu.ac.jp

これらの問題に対し有効なアプローチとして，本稿では協調的マルチビュー（Coordinated Multiple Views，CMV）のコンセプトに着目する．提案システムでは，ユーザに提示すべき情報をクラスタレベル，文書レベル，単語集合レベル，単語レベルの 4 レベルに分け，それぞれを別のビューに表示することで，複数種オブジェクトの情報を適切なビューに表示することができる．

また，ビュー間の協調を可能にすることで，異種オブジェクト間の関係性を把握可能とする．制約付きクラスタリング手法としては，類似度計算における単語の重みをユーザが制約として与える手法を採用する．さらに，各文書に対してユーザによるラベル付けを可能とすることで，ユーザの視点とクラスタリング結果の比較を支援する．

提案システムの開発には，テキストデータマイニングのための統合開発環境（TETDM）[3]を採用する．上述のクラスタレベル，文書レベル，単語集合レベル，単語レベルそれぞれに対応したパネル，およびパネル間の連動を実装する．TETDM を用いて実装した提案システムを用いて，言語処理学会[1]年次大会の論文データを対象とした評価実験を行った結果を示す．

2. 関連研究

2.1 テキストデータマイニング

膨大な文書データから有用な情報を発見したり，

[1] http://www.anlp.jp/

文書データ間の関係性を把握したりするといった，ユーザの様々な要求に十分に対応できる情報アクセス手段の重要性が指摘されており[4]，情報抽出，文書検索，文書分類や文書クラスタリングなどのテキストデータマイニングの技術が研究されている．

本研究で利用するテキストデータマイニング統合環境 TETDM は，柔軟な方法で様々なテキストマイニング技術を組み合わせることを目的として開発されている[3]．データ分析ツールを開発し，統合環境内にモジュールとして組み込むことが可能である．モジュールには文書処理を実行する処理モジュールと，処理モジュールの出力を表示する可視化モジュールの 2 種類がある．TETDM はインタフェース画面に設置されたパネルに対し，処理モジュールと可視化モジュールを 1 対 1 で組み合わせることでツールを構成する．また，TETDM は連動処理部によってモジュール間の連携を制御・実施することが可能であり，異なるパネル間の協調の実装を効率的に行うことができる．

2.2 協調的マルチビュー

複数のビューから構成される可視化システムを設計するために，協調的マルチビューのコンセプトが提案されている[5]．複数のビューによって提示された情報が，協調によって相互作用することで，ユーザはデータを効率良く理解することが可能となる．

代表的なマルチビューのタイプの一つに，一方のビューでデータの全体もしくは非常に大きな部分（overview）を表示し，別のビューでデータのより詳細部分（detail view）を表示する，Overview + Detail views がある．Zhang ら[6]は，このタイプのマルチビューを用いてインターネットログの異常を検出するネットワーク管理システムを提案している．

一般的な協調に Brushing と Navigational Slaving がある. Brushing は, あるビューで要素を選択すると，リンクされた他のビューにおける同一の（もしくは関連のある）要素が同時にハイライトされる．Navigational Slaving はユーザがあるビューでスクロールなどのナビゲーション動作を行うと，リンクされた他のビューに自動的に反映される事を指す．Weaver[7]は，Brushing や Navigational Slaving による協調機能を持ったマルチビューをユーザがインタラクティブに構築可能なシステムを提案している．

3. 提案システムの概要

3.1 提案システムのコンセプト

インタラクティブクラスタリングシステムを開発する際,「複数オブジェクトの情報をどのように表示するか」,「異種オブジェクト間の関係性をどのように表示するか」,「制約付きクラスタリングをどのように導入するか」といった点を検討する必要がある．前者の 2 点に関して，本稿では協調的マルチビューのコンセプトを採用する．提案システムを設計するにあたり，以下の 4 点を考慮する．

・文書情報を 4 つのレベルに分けて並列表示：各レベルの情報を適切なビューに表示
・ビュー間の協調：異種オブジェクト間の関係性を提示
・任意の単語の重み変更，再クラスタリングを反復的に実行：ユーザの視点を反映させた制約付きクラスタリングの導入
・任意の文書へのラベル付与：ユーザの視点とクラスタリング結果の比較を支援

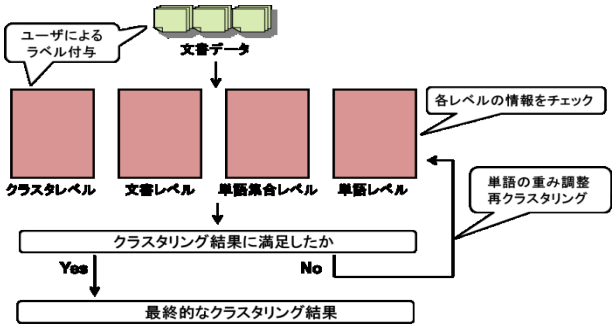


図 1：提案システムのフローチャート

図 1 に提案システムのフローチャートを示す．ユーザは関心を持った文書にラベルを付与し，クラスタリング結果を調べる際に利用する．

提案システムはクラスタリング結果についての情報を 4 つのレベル（クラスタレベル，文書レベル，単語集合レベル，単語レベル）に分け，各レベルの情報を異なるビューに並列表示する．これらのビューを組み合わせることで，ユーザは効率よくクラスタリング結果を確認することができる．表 1 に各レベルに提示する情報及び可能な操作を示す．

文書クラスタリングには k-means アルゴリズムを採用する．文書$d_i = (w_{i1}, \cdots, w_{in})$における単語$t_j$の重み$w_{ij}$は tf-idf 値を用いる．

提案システムは単語の重みによってユーザフィードバックを与えるインタラクティブクラスタリングを採用する[8]．ユーザが単語の重みを調整したい場合, 任意の単語の重みに3^kを掛けることによって単語の重みを調整し，システムにフィードバックを与える．kの値は単語レベルに対応したパネルでインタラクティブに変更することができる．全単語のkの値の初期値は 0 となっており，1 ずつ増減可能である．文書間類似度は式(2)によって計算する．式(2)

はコサイン類似度を元にしており，調整された単語の重みを文書間の類似度に反映させている．

$$sim(d_1,d_2)=\frac{\sum_{j=1}^{n}3^{2kj}w_{1j}w_{2j}}{\sqrt{\sum_{j=1}^{n}(3^{kj}w_{1j})^2}\sqrt{\sum_{j=1}^{n}(3^{kj}w_{2j})^2}} \quad (2)$$

重み調整と再クラスタリングを反復的に行うことによって，ユーザは最終的に満足のいくクラスタリング結果を得ることができる．また同時に，クラスタリング結果に影響を与える単語を理解することや，各クラスタリング結果から新たな知識を得ることが可能となる．

表 1：各レベルに提示する情報及び可能な操作

レベル	提示する情報	可能な操作
クラスタレベル	・クラスタリング結果 ・各文書のタイトル ・重みが調整された単語とその重み	・クリックによる文書番号の指定 ・クラスタ数の変更 ・文書へのラベル付与
文書レベル(指定文書に関する情報)	・指定文書本文 ・指定文書の重要文 ・指定文書の重要単語 ・指定文書の情報(文数，単語数)	
単語集合レベル	・指定文書の単語一覧とその出現頻度 ・指定文書の単語間のコサイン類似度	・クリックによる単語の指定
単語レベル(指定単語に関する情報)	・指定単語を含む文書一覧 ・指定単語の意味 ・指定単語とコサイン類似度の高い単語	・クリックによる文書番号の指定 ・指定単語の重み調整 ・再クラスタリング

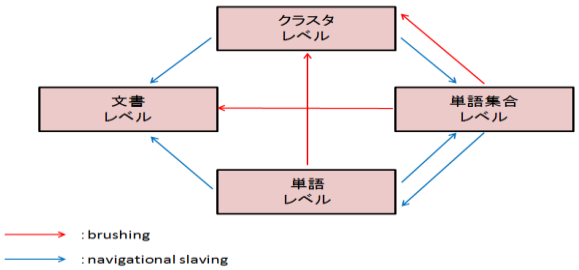


図 2：各レベル間の協調

図 2 に提案システムの各レベル間での協調を示す．図 2 において，赤い矢印は Brushing に対応し，青い矢印は Navigational Slaving に対応している．ユーザがクラスタレベルに対応したビューで文書を指定した場合，その情報が文書レベルと単語集合レベルに対応したビューに表示される．また，ユーザは単語レベルに対応したビューでも文書を指定することができ，指定された文書に関する情報が文書レベルと単語集合レベルに対応したビューに表示される．さらに，単語レベルに対応したビューで文書を指定した場合，その文書を含むクラスタがクラスタレベルでハイライトされる．

ユーザが単語集合レベルに対応したビューで単語を指定すると，その情報が単語レベルに対応したビューに表示される．さらに，指定単語が文書レベルに対応したビューに表示されている本文中に出現している場合，その部分がハイライトされる．同時に，クラスタレベルに対応したビューにおいて指定単語を含む文書番号もハイライトされる．

3.2 提案システムのプロトタイプ

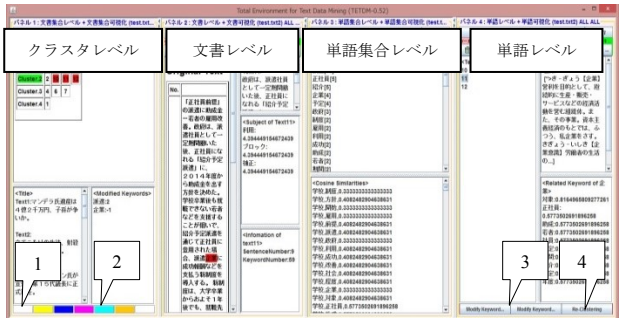


図 3：提案システムのインタフェース

図3に提案システムのインタフェースを示す．提案システムは4つのパネルから構成されており，左端のパネルから順に，クラスタレベル，文書レベル，単語集合レベル，単語レベルにそれぞれ対応している．また，クラスタ数変更テキストフィールド（図中1）でクラスタ数の変更，ラベル付与テキストフィールド（図中2）で文書へのラベル付与，単語の重み調整ボタン（図中3）で単語の重み調整，再クラスタリングボタン(図中4)で再クラスタリングが可能である．

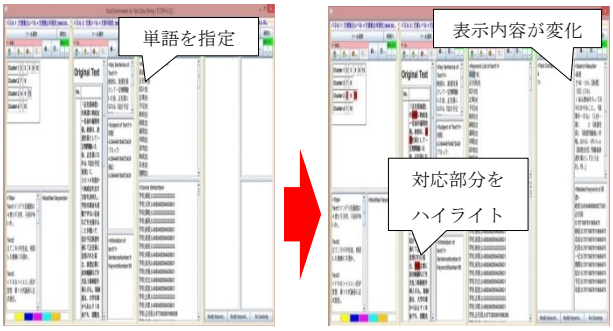


図 4：協調の例

図 4 に協調の例を示す．ユーザが単語集合レベルに対応したパネルにおいて，単語をクリックすると，指定された単語の情報が単語レベルに対応したパネルに表示される．また，クラスタレベルに対応したパネルのクラスタリング結果で，指定された単語を含む文書番号が赤色でハイライトされる．同時に，指定単語が文書レベルに対応したパネルに表示されている文書の本文中に出現している場合，その部分が赤色でハイライトされる．

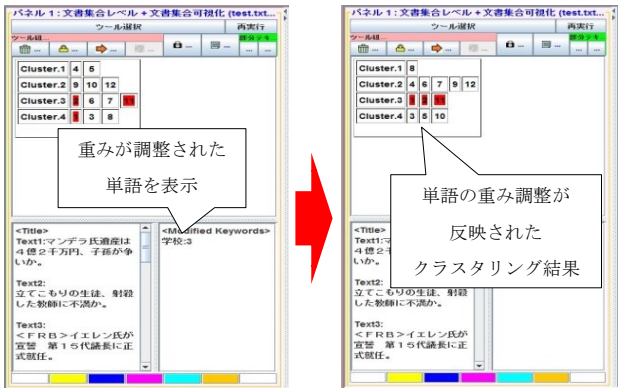


図 5：単語の重み調整と再クラスタリングの例

図 5 に単語の重み調整と再クラスタリングの例を示す．ユーザは各レベルの情報を元に，重みを調整したい単語を決定し，単語レベルに対応したパネルにある単語の重み調整ボタンをクリックする．ユーザが単語の重み調整ボタンをクリックするごとに，前述の単語の重み調整係数である 3^k の k の値が 1 ずつ増減し，指定した単語の重み調整が実行される．同時に，重みが調整された単語とその重みがクラスタレベルに対応したパネルに表示される．ユーザが単語レベルに対応したパネルの再クラスタリングボタンをクリックすると，クラスタリングが再実行され，クラスタレベルに対応したパネルのクラスタリング結果が更新される．

単語の重み調整と再クラスタリングを反復して行うことにより，最終的にユーザの望むクラスタリング結果を得ることができ，ユーザの視点を反映させた制約付きインタラクティブクラスタリングシステムを実現している．

また，クラスタレベルに対応したパネルのラベル付与テキストフィールドにおいて，利用したい色のテキストフィールドを選び，ラベルを付与したい文書番号を入力すると，クラスタレベルに対応したパネルに表示されるクラスタリング結果において，入力した文書番号の色が指定した色に変更される．

4. 評価実験

4.1 実験概要

文書クラスタリングにおける提案システムの有用性や，ユーザにとって有用な情報や協調，また，提示する情報の違いがユーザの分析作業や実験結果に与える影響を分析するために，工学系の大学生および大学院生の男女16名に協力を依頼し評価実験を行った．実験では，言語処理学会年次大会発表論文集の，2002年と2003年におけるポスター発表の予稿データを利用し，「一方の年次に特有の話題」および「両方の年次に共通の話題」を発見するとともに，発見した話題に関するクラスタを生成するタスクを行ってもらった．また，提示する情報の違いが分析作業に与える影響について調査するために，単語集合レベルに対応したパネルの有無によって，実験協力者を8人ずつの2グループに分けて実験を行った．実験終了後には，発見した話題と提案システムの提示情報・機能などの有用度に関する5段階評価のアンケートに回答してもらった．また，視線追跡装置 Tobii X120[2]を用いて記録した，作業中の実験協力者の視線データの分析も行う．

4.2 実験結果

表 2：発見された話題

		単語集合レベルあり	単語集合レベルなし
クラスタが形成された話題	共通話題	・特許(3) ・手話(16) ・翻訳(18)(21)(18) ・日本語の文節解析(21) ・言語の分類(16) ・言語の分析(26) ・話し言葉(6)	・検索(21) ・翻訳(15)(37) ・電子テキスト(29) ・学習支援を目的とした研究(19) ・ユーザを想定した研究(11) ・手話(12) ・テキスト(17) ・インターネット検索(20) ・2言語間の変換(14) ・コーパス(78)
	特有話題	・機械翻訳(35) ・web サイトを利用した研究(6) ・音韻(2) ・外国語の音声翻訳(3) ・品詞(4) ・形態素解析(7) ・話し言葉(6) ・換言(4)	・SVM(3) ・Perl を使用した研究(4) ・テストコレクションの利用(2) ・通訳(2) ・自動的(8) ・発話(4)
クラスタが形成されなかった話題	共通話題	・検索 ・要約支援 ・異なる言語の処理 ・翻訳 ・対話 ・データベース ・談話 ・自然言語処理 ・音声(対話) ・文法	・音声 ・機械翻訳 ・談話 ・言語解析 ・運転 ・音声翻訳 ・音声認識
	特有話題	・形態素解析	・中国語 ・音声解析・利用 ・言葉

表 2 に発見された話題を示す．括弧内の数字はその話題に対応したクラスタのサイズ（論文数）を示している．実験協力者あたりの平均発見話題数を表

[2] http://www.tobii.com/

3 に示す．表 3 より，発見した話題の数は共通話題の方が多い．これはデータセットの年次が近いためと考える．また，単語集合レベルがある場合の方が，発見した話題の数が多くなっており，単語集合レベルのパネルによって，実験協力者が単語に注目しやすく，効率良く話題を発見可能であったと考える．

表 2 より，特有話題のほとんどはクラスタ形成されており，そのクラスタサイズは小さい．これより，関連文書数が少ないほど，重みの調整によるフィードバックでクラスタにまとめやすいと考える．逆に，共通話題はクラスタが形成されなかった話題が多く，クラスタが形成されている話題についてもクラスタサイズが大きくなっている．これより，共通話題の方が関連する論文が多く，対応するクラスタの生成が困難であったと考える．

表 2 において，緑色の話題は「翻訳」と「音声」に関する話題である．「翻訳」と「音声」も両実験に共通して発見されている話題であり，含まれている論文数が多いと考える．また，「翻訳」「外国語の音声翻訳」「音声翻訳」のように，実験協力者が話題をどの程度まで細かく判断するかによって，共通話題であるか特有話題であるかに違いが出ることも観測された．この時，実験協力者が話題の粒度を細かく捉える場合には論文数が少なくなる傾向にあった．従って，話題の粒度がクラスタの形成しやすさに影響を与えたと考える．

表 3：平均発見話題数

	単語集合レベルあり	単語集合レベルなし
共通話題	2.75	2.25
特有話題	1.125	1.125

表 4 に各レベルの有用度の平均をそれぞれ示す．クラスタレベルと単語レベルは単語集合レベルの有無で有用度の平均に差があまり見られないのに対し，文書レベルの有用度は違いが大きくなっている．これは，文書レベルと単語集合レベルにそれぞれ対応したパネルは，単語の指定を行う点で役割が重複しているためと考える．また，単語集合レベルの有用度は高く，単語集合レベルに対応したパネルでの単語指定が行いやすかったと考える．

表 4：各レベルの有用度

	単語集合レベルあり	単語集合レベルなし
クラスタレベル	4	4.375
文書レベル	3.125	4.375
単語集合レベル	4.5	–
単語レベル	3.375	3.5

表 5：有用度の高い情報と協調

	単語集合 レベルあり	単語集合 レベルなし
クラスタリング結果	4.75	4.625
重みが変更された単語	4.375	4.5
文書本文	3.25	4.75
指定文書の単語一覧と出現頻度	4.75	–
指定単語を含む文書一覧	4.125	4.25
クラスタレベルでの表示変更	4.5	4.625
文書レベルでの表示変更	–	4.125
単語集合レベルでの表示変更	4	–
文書番号のハイライト	4.5	4.875
指定単語のハイライト	3.75	4.75

表 5 に有用度が高いと評価された情報と協調を示す．クラスタリング結果やクラスタレベルでの表示変更など，作業を行う際に必須となるものの有用度は非常に高くなっている．また，文書本文，指定単語のハイライトは単語集合レベルの有無により有用度に差が出ている．先述のように単語集合レベルなしの場合は文書レベルの文書本文で単語指定を行うため，これらの有用度が高くなったと考える．

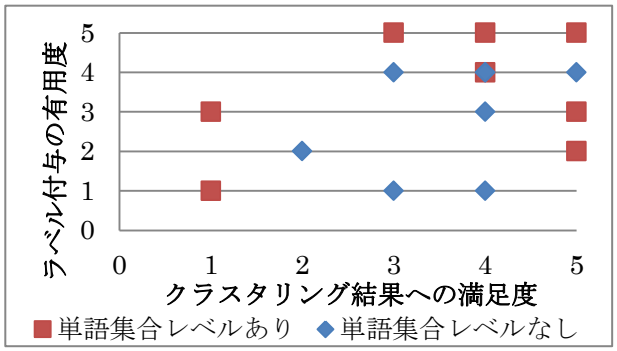


図 6：クラスタリング結果とラベル付与の関係

図 6 に，アンケートにより得られたクラスタリング結果への満足度とラベル付与機能の有用度の関係を示す．図より，ラベル付与が高評価の実験協力者はクラスタリング結果への満足度が高くなる傾向が見られる．これはラベル付与によって，発見した話題を見失うことなく，クラスタに関心のある論文が集まっているかどうかを効率よく確認することができたためと考える．ラベル付与を高評価していないが結果への満足度が高い実験協力者もいたが，その実験協力者は発見話題数が少なく，指定単語を含む文書番号がハイライトされるためラベル付与を高評価しなかったと述べていた．すなわち，比較的少ない話題であればラベル付与を行わなくても，指定単

語のハイライト機能によって話題を見失わず，クラスタリング結果の確認が可能であると考える．

表 6：視線データ

	単語集合レベルあり				単語集合レベルなし		
	クラスタレベル	文書レベル	単語集合レベル	単語レベル	クラスタレベル	文書レベル	単語レベル
見た回数	633	559	585	103	139	186	39
平均時間[秒]	0.75	1.03	1.21	0.3	1.42	2.12	0.68
総時間[秒]	475.58	576.5	709.33	31.3	198.04	394.37	26.33
総クリック数	946	31	280	54	256	114	94
有用度	4	2	5	3	5	5	3

表 6 に，クラスタレベルを有用と判断した 2 名の実験協力者の視線データを分析した結果を示す．各パネルを AOI（Area of Interests）に設定し，パネル内に視線が入った回数･時間を測定している．また，当該実験協力者の gaze plot を図 7 に示す．左側が単語集合レベルのある実験協力者，右側が単語集合レベルのない実験協力者である．どちらもクラスタレベルを見た回数，総時間，総クリック数が多くなっており，作業中に多用した結果有用と判断したと考える．また，単語レベルの平均時間と総時間は比較的少ない．これは単語レベルの情報量が少なく，表示情報を確認するのにそれほど時間がかからなかったためと考える．

単語集合レベルがない実験協力者は，文書レベルの文書本文で単語指定を行うため，単語集合レベルがある場合よりも，文書レベルを利用する必要がある．表 6 を見ても，文書レベルの総時間と平均時間，総クリック数が，単語集合レベルがある場合と比較して多くなっていることがわかる．また，文書レベルの有用度は高いと回答されており，この場合も多用したことが評価につながったと考える．

図 7 において，両実験協力者とも，クラスタレベルから文書レベルといったように，隣接するパネルへの視線移動が多く見られた．このことから，実験協力者が文書の情報をクラスタレベルから単語レベルへと，レベル順に確認する傾向にあったと考える．

以上より，作業中の視認回数・時間やクリック回数と，有用度の間には関係が見られることから，実験協力者による有用度の評価は，作業の実態を反映したものと考える．

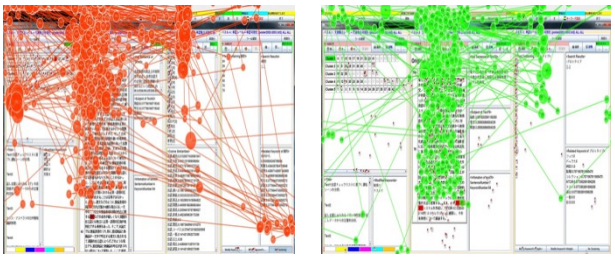


図 7：クラスタリング 1 回あたりの視線の動き（左：単語集号レベルあり，右：なし）

5. おわりに

本稿では，協調的マルチビューに基づくインタラクティブ文書クラスタリングシステムを提案した．ユーザ実験と視線データの分析を行い提案システムの有用性を示すと共に，ユーザにとって有用な情報や協調，提示する情報の違いが分析作業や実験結果に与える影響についても考察を行った．

本稿により得られた知見は，無駄な情報提示や協調の削減や，各情報の最適な提示方法の検討などといった，インタフェースの設計に貢献することが期待できる．

参考文献

[1]那須川 哲哉: テキストマイニングを使う技術/作る技術, 東京電機大学出版局, 2006.

[2]三宅 遼祐, 山田 誠二, 岡部 正幸, 高間康史: インタラクティブクラスタリングのためのマルチタッチインタフェースの提案, 第 25 回人工知能学会全国大会, 1J1-0S9-3, 2011.

[3]砂山 渡, 高間 康史, ダヌシカ ボレガラ, 西原陽子, 徳永 秀和, 串間 宗男, 松下 光範: テキストデータマイニングのための統合環境 TETDM の開発, 人工知能学会論文誌, Vol. 28, No. 1, pp. 1-12, 2013.

[4]市村 由美, 長谷川 隆明, 渡部 勇, 佐藤光弘: テキストマイニング-事例紹介, 人工知能学会誌, Vol.16, No.2, pp.192-200, 2001.

[5]J. C. Roberts: State of the art: Coordinated & multiple views in exploratory visualization, *International Conference on Coordinated and Multiple Views in Exploratory, pp.61-71, 2007.*

[6]T. Zhang, Q. Liao, L. Shi: Bridging the Gap of Network Management and Anomaly Detection through Interactive Visualization, *Pacific Visualization Symposium, pp.253-257, 2014.*

[7]C. Weaver: Building Highly-Coordinated Visualizations in Improvise, *IEEE Symposium on Information Visualization*, *pp. 159-166, 2004.*

[8]岡田 貴史, 石橋 融, 高間 康史: M2VSM を用いたテキストマイニングシステムの構築に関する考察, FSS2006, pp. 203-206, 2006.

[9]那須川 哲哉, 諸橋 正幸, 長尾 徹: テキストマイニング: 膨大な文書データの自動分析による知識発見, 情報処理学会, Vol. 40, No. 4, pp. 358-364, 1999.

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-03

照応解析と動詞シソーラスに基づく ニュース概要把握のための図解生成システム

Generating Illustrated Diagram to Support Understanding of News Summaries Based on Anaphora Resolution and Verb Thesaurus

廣田 暖貴[1*]　白松 俊[1]　岩田 彰[1]

Haruki Hirota[1], Shun Shiramatsu[1], Akira Iwata[1]

[1]名古屋工業大学 大学院工学研究科
[1]Graduate School of Engineering, Nagoya Institute of Technology

Abstract: The present study is aiming to automatically generate news summaries using illustrated diagram that enables users to understand an overview of the news. To generate illustrated diagram, zero anaphora resolution is needed because the elements of the statements are often omitted in Japanese sentences. In this study, we propose a method for zero anaphora resolution focusing on human nouns that appear in diagram based on the centering theory, and a method for hierarchical management of illustrated diagram using the verb thesaurus. We conducted an experiment to compare summaries generated by the system and summaries of existing conventional services. The experimental result indicated that the illustrated diagram is useful to understand the overview.

1　はじめに

本研究ではニュース記事を入力とし，テキストと図を複合的に用いた図解を要約として自動生成する．これにより，ニュースの概要把握を支援するシステムの開発を目指す．しかし，日本語の文章では，主語や目的語など文の要素が省略されることが多く，テキストを図解に変換するためには省略された要素を特定する必要がある．また，直感的な図解を生成するためには，動詞に適した表現をする必要があるが，動詞ひとつひとつに図を登録すると管理コストが膨大になってしまう．

そこで本研究では，図解に出現する格要素および人や組織の名詞に着目したゼロ代名詞補完手法と，動詞シソーラスを用いた図解の階層管理手法を提案する．また，提案手法を用いて生成した要約について評価実験を行い，要約での図解の有用性を確認する．

図解を用いない一般的な自動要約では，情報のソースを受け取り，そこから内容を抽出し，最も重要な内容をユーザに，簡約した形で，かつ，ユーザやアプリケーションの要求に応じた形で掲示する[1]．

*連絡先：　名古屋工業大学大学院工学研究科
創成シミュレーション工学専攻
〒466-8555 愛知県名古屋市昭和区御器所町
E-mail: 25413561@stn.nitech.ac.jp

表 1：テキストと図の性質比較

性質	テキスト	図
概要の一覧性	△	◎
詳細な記述力	◎	△
理解形態	ボトムアップ処理	トップダウン処理

原文書に含まれる情報を短時間に理解できることが求められるため，読み手にとってわかりやすく表現することが重要である．既存の自動要約サービスで用いられているメディアであるテキストは，詳細な意味や種々の抽象概念を表現できるという利点をもつ．しかし，構成される要素から全体へと理解していくため，直感性・概覧性に欠け，理解に時間を要するといった欠点を持っている．表 1 に，テキストと図の性質を示す．この問題は，表現している情報の内容概略を直感的に把握することができる図を複合的に用いることによって克服できると考えられる．

2　関連研究

すでに，テキストから図的メディアを生成する研究[2]，[3]が行われている．関連研究[2]では，物語テキストに含まれる各発話文についての話し手と聞き手を同定し，その会話の中身から登場キャラクタ同士の関係を推定し，関係図の自動構築を行っている．本研究では物語中に含まれる会話に限定することな

く，Web ニュースのテキストを対象とし，構文・格解析により人物の関係を図にしている点で異なる．関連研究[3]では，会議などの会話内容のテキストを入力として，ノードとエッジからなる DT-MAP と呼ばれるグラフを作成し，会話内容を表現している．本研究では，テキストを並べたグラフの作成ではなく，図とテキストを複合的に用いることで，要約を生成している点で異なる．

本研究では，これらの関連研究とは異なる着眼点として，図解のタイプを定義して動詞シソーラス上で継承させて管理できるようにした点と，ゼロ代名詞の補完の際に図解生成に必要な格を考慮した点に注力して研究した．

3 図解生成手法

3.1 図解の定義

テキストと図を複合的に用いることによって文書内の情報を効率的に提示することができるが，図を用いた表現には大きな自由度があり，内容が十分に伝わる形式を選ぶことが必要となる．そのために，図解としてどのようなものを生成するべきかを考慮する．ニュースでは「首相が記者団に発表する」や「王子が日本を訪問」など,「人物」が「対象」に「何かをした」といった記事が多く見られる．そこで，動作の内容や動作の主体と客体を図化すると，テキストのみによる表現よりも読みやすい要約となり，概要把握の補助になると考えた．

そのためには,「誰が」,「何を」,「誰に」,「どうした」といった図解を生成する必要がある．ここで,「誰に」という動作の客体が存在しない場合もあるため，客体をもたない図解と，客体をもつ図解に分けて生成する.図 1 のように,それぞれの図解の type を single，pair と定義する．

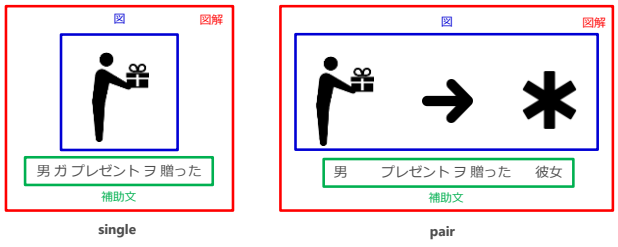


図 1：図解の type と各部の名称

図解に用いる図は，補助文に含まれる動詞に適したピクトグラムを人手で選択している．

ニュースの内容を表現する図解を生成するためには，次のような流れの処理が必要となる．

(1) ニュースを端的に表す補助文の生成
(2) 動詞に適した図の生成

3.2 補助文の生成

補助文を生成するには，ニュース記事から動詞を抜き出し，その動詞を基に図を説明する補助文を生成する．動詞を中心とした文の構造を把握するために，述語項構造解析を行う．述語項構造は，述語と項（述語と格関係にある単語）を同定するものである．動詞・形容詞などの「述語」は，文の中心で動作・状態を表す要素である．そして「項」(名詞＋格助詞）は述語が表す事態に関係する人，ものを表現する要素である．述語項構造を用いることで文中の各述語が表す意味を補う働きをする項を同定し，文の意味の骨格を表すことが可能となる[4]．本研究では既存の解析ツールである KNP[5]を使用することで，文の構造を取得している．表 2 は「太郎は学校へ行ってサッカーをした。」という文の解析結果の例である．この例文には「行く」と「する」の二つの動詞がある.「行く」という動詞は，ガ格に「太郎」，ヘ格に「学校」が相当する．一方,「する」という動詞は，ヲ格に「サッカー」が相当する．しかし，ガ格が取れていないため，これを図解にすると，主語のない意味のわからない図解を生成してしまう．そのため，ガ格に「太郎」という名詞を補完するために，述語項構造解析と同時に照応解析を行う必要がある．

表 2：述語項構造の例

	ガ格	ヲ格	ヘ格
行く	太郎		学校
する	φ	サッカー	

照応とはある表現が同一文章内の他の表現を指す機能をいい，指す側の表現を照応詞，指される側の表現を先行詞という．日本語の場合は述語の格要素の位置に出現している照応詞が頻繁に省略される．この省略された格要素をゼロ代名詞（記号 φ で表す）といい，ゼロ代名詞と照応関係となる場合をゼロ照応と呼ぶ[6]．このため，ニュース記事のような自然言語文をそのまま用いると，主体や客体が抜けたわかりづらい図解を生成してしまうという問題がある．

3.3 図の生成

文の内容を直感的に理解できる図解を生成するには，動詞の意味に合った図を生成する必要がある．しかし，図の種類が多くなると，膨大な数の動詞ひとつひとつに，意味に適した図を人手で登録することになり,管理コストが膨大になってしまう.逆に,図の種類が少なく，同じような図が繰り返し用いられていると，動詞の意味に適した図解にならず，理解の妨げになってしまうことや，直感的な理解に役

立たない問題がある．

3.4 研究目的

3.2 節と 3.3 節から，ニュース記事の内容を端的に表す図解を生成するためには二つの課題を解決する必要がある．

(1) ニュース記事をそのまま入力として与え格解析をするだけでは，ゼロ代名詞が頻繁に出現するため，主体や客体の抜け落ちた図解を生成してしまう

(2) 動詞の意味に合った図が少なすぎると直感的に理解できる図を生成できず，逆に多すぎると図と動詞を対応づける管理コストが膨大になる

本研究は，この二つの課題を解決し，生成物の要約としての有用性を確認することを目的とする．

そこで，本研究ではセンタリング理論を応用し図解出現要素に着目したゼロ代名詞補完手法と，動詞シソーラスを用いた動詞に対応する図の階層管理手法を提案する．

4 提案手法

4.1 ゼロ代名詞補完手法

センタリング理論は英語の代名詞の照応関係を決定する手法として Grosz[7]らによって提案され，大規模な知識を必要とせず，計算機上で実現容易であるなどの利点を持つ．文の中心になっているものをセンターと呼び，談話中でセンターが連続している場合，つまり話題が連続している場合には代名詞が使われているはずである，という基本規則を利用して照応解析を行っている．本研究では特に，図解に現れる格要素や人物に特化した．

4.1.1 センターの定義

談話単位中の各発話 U には，前向き中心(forward-looking-center) $C_f(U)$ と後向き中心(backward-looking-center)$C_b(U)$が結びついている[8]．C_fは発話U_iで実現される名詞リストを次発話U_{i+1}での参照されやすさで並べたもので，C_fのうち現在の話の中心になっている特別な要素がC_bである．C_fの要素は次のランキングで順序付けられる．

主題 ＞ ガ格 ＞ 二格 ＞ ヲ格 ＞ その他

4.1.2 センターの制約

発話列$U_1, \ldots, U_m$からなる談話単位中の各発話U_iについて，以下の制約が成り立つ[8]．

(a) ただ一つの$C_b(U_i)$が存在する．

(b) $C_f(U_i)$のあらゆる要素はU_iで実現されている

(c) $C_b(U_i)$は，$C_f(U_{i-1})$の要素のうちU_iで実現されているものの中で，$C_f(U_{i-1})$での序列が最も高かったものである．

4.1.3 ゼロ代名詞補完規則

センタリング理論に基づく，図解出現要素に着目したゼロ代名詞補完手法について説明する．KNP の出力する格をすべて必須格と考えてゼロ代名詞の補完を行うと，補完対象の誤りが頻繁に発生してしまう．そのため，ゼロ代名詞の補完の際に，図解生成に必要な格(図解出現格)を考慮し，ガ格，二格，ヲ格のみを補完対象とした．補完の規則として以下のとおり定めた．

(a) 図解出現格のみを補完対象とする

(b) 補完対象格の優先順序は以下の通りである

ガ格 ＞ 二格 ＞ ヲ格

(c) 述語項構造解析結果に，補完する語が含まれている場合はC_fの序列が次に高いものを補完する

4.1.4 提案手法の適用例

提案手法の適用例について説明する．

A) 彼は彼女に花を贈りました。

B) φ ガ とてもドキドキしました。

表 3：センターの遷移

	C_b	C_f
(A)	彼	彼<主題，人>，彼女<ヲ格，人>，~~花<二格，植物>~~
(B)	彼	彼<ガ格，人>

表 3(A)のように人の名詞にのみ着目すると，「花」が除外される．「彼」は「彼女」より序列が高いため，(B)のゼロ代名詞には「彼」が補完される．

4.2 階層管理手法

シソーラスを用いることで，階層構造で管理できるため，上位の動詞に登録されている図を下位の動

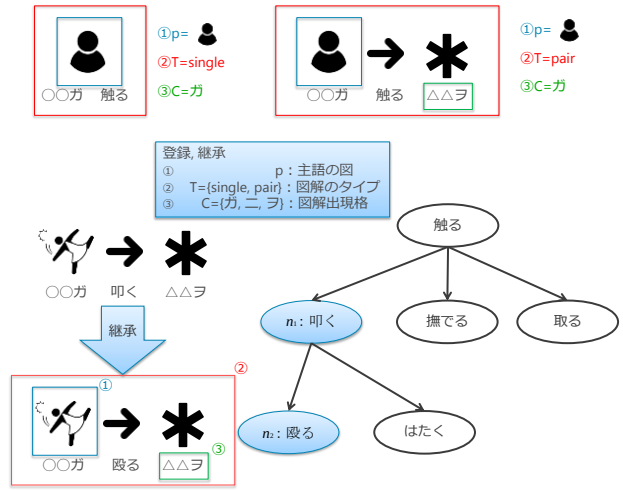


図 2: 階層管理手法

詞に継承することができる．階層管理手法の例を図 2 に示す．ある動詞に対して図が登録されている場合はその図を出力に用い，登録されていない場合は上位の動詞を参照し，上位の動詞に図が登録されている場合はその図を継承し，出力に用いる．

提案手法は以下の式で表すことができる．あるノードnの図をd(n)，タイプをtype(n)とし，図および図解のタイプの登録はそれぞれ以下の式で表現する．

$$\mathrm{reg\ d(n)} = \mathrm{p},\ \mathrm{reg\ type(n)} = (T, C)$$

上位ノードを$\mathrm{sup}(n_2) = n_1$と表すとき，type の継承は以下の式で表す．

$$\mathrm{type(n)} = \begin{cases} \mathrm{reg\ type}(n) & if\ \mathrm{type}(n) \neq \varphi \\ \mathrm{type(sup}(n)) & if\ \mathrm{type}(n) = \varphi \end{cases}$$

図の継承も同様に，以下の式で表現される．

$$\mathrm{d(n)} = \begin{cases} \mathrm{reg\ d}(n) & if\ \mathrm{d}(n) \neq \varphi \\ \mathrm{type(d}(n)) & if\ \mathrm{d}(n) = \varphi \end{cases}$$

次章より，提案手法を用いた要約生成処理について述べる．

5 要約生成処理

提案手法を用いた要約生成処理について説明する．簡単な処理の流れを図 **3** に示す．

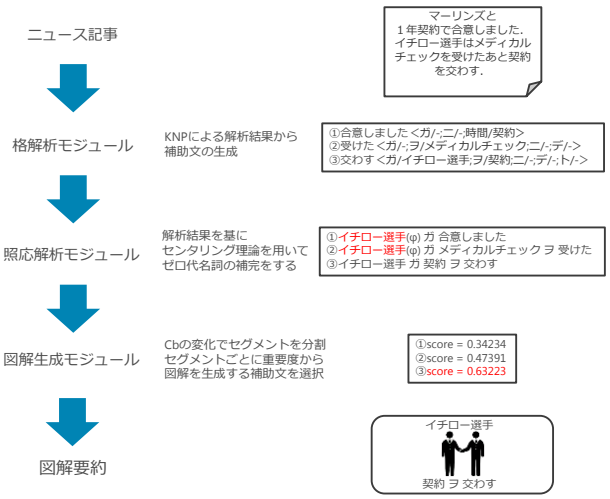


図 3: 処理の流れ

最初に入力テキストであるニュース記事を文単位に分割して，一文ずつ KNP で述語項構造解析を行う．次に，解析結果を用いて補助文を生成する．そして，生成した補助文のゼロ代名詞に人，組織・団体の名詞を補完する．続いて，補助文の重要度を計算し，セグメントごとの生成する図解を決定する．図解に用いる図は 4.2 節で述べた手法で決定する．生成する図解は図 4 のようなものになる．現在，補助文は簡易的なものを生成しているため，着目している動詞を含む一文を図解に並べて表示している．これをニュース記事全文に対して繰り返し行うことにより複数の図解を生成し，ニュース記事全体の要約を生成する．

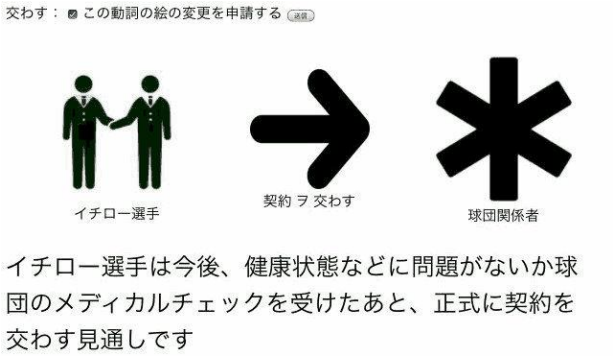


図 4：要約に用いる図解とニュース文

5.1 重要文選択

短時間で概要を把握するには，ニュースの中から要約に相応しい文を選び出す必要がある．重要文選択は以下の二つの仮定に基いて行っている．

1. C_bが変化するとセグメントを分割可能
2. TF-IDF の総和が大きい文が重要文

重要文選択の例を表 4 に示す．

5.1.1 C_bの変化によるセグメンテーション

本研究ではC_bは常に人や組織の名詞である．つまり，話の中心人物が変化すると，セグメントを分割する．セグメンテーションの例を表 4 に示す．各発話中，C_bとなっている要素を下線で示している． この例では三つのセグメントに分割される．

5.1.2 TF-IDF 重み付け

TF-IDF は，文書中の単語の重みとして広く使用される尺度である．単語の文書内での頻度を表す TF と，単語が現れる文書数に基づき語の珍しさを表す IDF を掛けあわせた尺度であり，その値が大きいほど各文書を特徴付ける語だと言える．ここでは TF-IDF の総和で文の重要さを表すという単純な手法をとる．すなわち，語wの TF-IDF 値をtfidf(w)で表す時，語$w_1, w_2, \cdots w_m$から成る文sの重要度は以下の式で求める．

$$score(s) = \sum_{i=1,\cdots,m} \mathrm{tfidf}(w_i)$$

表 4 の文(3)の場合，重要度は以下のように求める．

$$score(s_{(3)}) = \mathrm{tfidf("ジェリー")} + \mathrm{tfidf("頭")} + \mathrm{tfidf("ネズミ")}$$

表 4：重要文選択例

発話	score
(1)トムはいつもジェリーを追いかけています	0.123456
(2)ジェリーはトムと同じ家に住んでいます	0.345678
(3)ジェリーは頭のいいネズミです	0.678912
(4)飼い犬のスパイクとジェリーは仲良しです	0.456789
(5)スパイクはトムに仕返しをします	0.567891

このように(2), (3), (5)が各セグメント内で最も重要な文と考え，図解を生成し要約に用いる．

5.2 図と図解出現格の登録

シソーラスには日本語 WordNet[9]を利用した．日本語 WordNet は大規模な語彙データベースであり，語を類義関係のセット（synset）でグループ化している点に特徴があり，一つの synset が一つの概念に対応している．また，各 synset は上位下位関係などの多様な関係によって結ばれている．この synset に対して図と図解出現格を登録する．

ユーザが要約の図解を見て，わかりづらいと感じた場合，図 5 の中にあるチェックボックスにチェクを入れることで, 図の変更を申請することができる．図 5 は管理者の図と図解出現格登録画面の一部である．この画面では，ユーザがわかりづらいと感じた動詞の現在の図と，その一文が表示されている．管理者はその文を見て動詞に適した絵と図解出現格を選択して登録する．

図 5 : 図と図解出現格の登録

6 評価実験

本実験の目的は，提案手法を用いて生成する要約について，図解の有用性および既存サービスの生成する要約との比較による優位性の確認である．被験者として学生 6 名を対象にアンケート実験を行った．

比較する要約は，既存サービス，提案手法（自動解析），提案手法（手動解析）の三つである． KNP やセンタリング理論による解析は，生成する要約の品質に大きな影響を与える．つまり，解析の誤りにより図解生成に誤りが生じることがあり，ニュース記事の自動要約における図解の有用性が確認できない可能性がある．そのため，提案手法（自動解析）が出力した結果において誤りを含む解析結果を，人手で修正した解析結果で図解を生成したものが提案手法（手動解析）である．また，既存サービスには，自動で記事を 3 文に要約するニュースサービスである SLICE NEWS[10]を用いた．

6.1 実験の手順

実験は以下の手順で行った．

(1) テキストのみのニュース記事を提示
(2) 要約を被験者に提示
(3) 評価を用紙に記入
(4) 手順(2)と手順(3)を残りの要約に対して同様に行う
(5) 手順(1)〜(4)を提示する要約の順序を変え，3 つのニュース記事に対して行う

要約を読むたびにニュースの内容を把握していくため，順序が後になった要約の評価のスコアが高くなってしまうことを避けるため，手順(5)のように，ニュース記事ごとに提示する要約の順序を変えて実験を行った．また，本システムの想定する利用形態である，スマートフォン（iPhone 5S）を用いて要約の提示を行った．

評価項目は，二つの基本的な評価軸ごとに四つの項目を設定した合計八つであり,「とてもそう思う（5 点)」「そう思う（4 点)」「どちらともいえない 3 点」「あまりそう思わない（2 点)」「全く思わない（1 点)」の 5 段階で評価を行った．(▼は逆転項目の意．)

- 内容的品質: 現文書の内容を適切に反映した要約になっているか
  - ① 文章表現が適切である
  - ② 必要な情報が省略されている▼
  - ③ 同じ情報が繰り返されている▼
  - ④ 無関係な情報が含まれている▼
- 読解的品質: 読みやすい要約になっているか
  - ⑤ 読みやすい
  - ⑥ 登場人物の関係がイメージしやすい
  - ⑦ すぐに概要を把握できる
  - ⑧ ほしい情報がすぐに見つかる

これらの評価項目の合計点の平均スコアを求め評価を行う．

6.2 実験結果・考察

評価実験の結果を図 6 に示す．

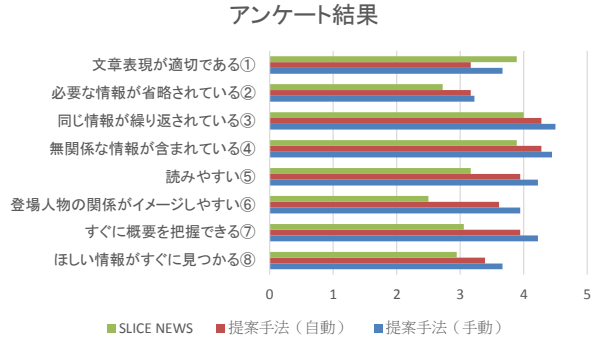


図 6: 評価実験結果

まず，特に手法間の差が顕著に見られた①，⑥，

⑦の項目について考察を述べたのち，全体の考察を述べる．

① 文章表現が適切である

八項目の中で唯一既存サービスが最も高いスコアを示した項目である．理由として，要約を生成する手法の違いが挙げられる．既存サービスはニュース記事を文に分解し，要約として相応しい文を選び，それらを繋げることで作る抽出的要約であり，作文は行っていない．それに対して，本システムでは述語項構造解析と照応解析の結果から，基本句単位に分解した後に作文や図解生成を行う生成的要約である．そのため，解析の誤りや作文の誤りによって誤りのある図解，すなわち要約が生成されたためであると考えられる．

⑥ 登場人物の関係がイメージしやすい

既存サービスは，元のニュース記事から重要な文を抽出し，それを繋ぎ合わせて要約を生成している．そのため，登場人物の関係を把握することに関しては工夫がなされておらず，元のニュース記事を読む場合となんら変化はない．それに対して，本システムでは二者間の関係を表す図解を複数生成しているため，既存サービスよりも高いスコアが得られたと考えられる．また，自動解析よりも手動解析のほうが高いスコアを得られた理由として，ゼロ代名詞補完の誤りによって自動解析では主体や客体の誤った人物関係を出力しており，イメージのしやすさの妨げとなっていたことが考えられる．

⑦ すぐに概要を把握できる

項目②の結果と合わせて本システムは，必要な情報を省略することなく，概要を把握しやすいという結果が得られたことから，図解が概要把握の手助けの一因になっているということが考えられ，自動要約における図解の有用性を確認することができた．

全体の結果では，項目①以外で，既存サービスよりも提案手法(自動)が，提案手法(自動)よりも提案手法(手動)が高いスコアを得られた．したがって，既存のテキストのみによる要約よりも本システムの図解を用いた要約は読み手にとってわかりやすい要約であったといえる．

また，提案手法(自動)よりも提案手法(手動)が高いスコアを得られたため，解析結果が要約品質に大きく影響することがわかった．述語項構造解析は KNP の結果に依存しているため，他手法によるゼロ代名詞補完は今後の検討課題である．

7 おわりに

本論文では，図解に出現する格要素および人や組織の名詞に着目したゼロ代名詞補完手法と，動詞シソーラスを用いた図解の階層管理手法を提案した．提案手法を用いてニュースの要約を自動生成し，評価実験を行った．

提案手法の生成する要約は，既存のニュース自動要約サービスの生成する要約と比較して，アンケート評価において内容的品質および読解的品質ともに高いスコアを得ることができた．特に，「登場人物の関係がイメージしやすい」，「すぐに概要を把握できる」の項目において顕著に見られた．また，既存のサービスと提案手法の生成する要約は同程度の文章量でありながら，提案手法のほうが概要を把握しやすいという結果が得られたということから，図解が概要把握に有用であることが確認できた．

今後の課題としては，ゼロ代名詞の補完精度を向上させるために，センタリング理論以外の機械学習による手法の検討を行う予定である．

謝辞

本研究の一部は，JSPS 科研費若手研究(B) (No.25870321) の助成を受けた．

参考文献

[1] Mani, I.: *Automatic Summarization*, John Benjamins Publishing (2001)

[2] 神代大輔，高村大也，奥村学: 物語テキストにおけるキャラクタ関係図自動構築，言語処理学会第 14 回年次大会発表論文集, Vol. 14, pp. 380-383 (2008)

[3] 二宮和弘，岡田信一郎，後藤寛幸，藤原祥隆: コミュニケーション支援のための会話内容の図式化ツールの開発，電子情報通信学会技術研究報告．TM，Vol. 104, No. 567, pp. 37-41 (2005)

[4] 岸邊賢太，横山晶一，井上雅史: 形容詞、複合動詞を扱う述語項構造解析システム，平成 22 年度第 6 回情報処理学会東北支部研究会，資料番号 10-6-B3-4, (2010)

[5] KNP, http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

[6] 飯田龍，小町守，井之上直也，乾健太郎，松本裕治: 述語項構造と照応関係のアノテーション: NAIST テキストコーパス構築の経験から，自然言語処理，Vol. 17, No. 2, pp. 25-50, (2010)

[7] Grosz., B. J., Joshi., A. K., and Weinstein, S.: Providing a Unified Account of Definite Noun Phrases in Discourse, In *Proc. of the 21st Annual Meeting of the Association for Computational Linguistics*, pp. 44-50 (1983)

[8] 佐竹正臣: 新聞記事の固有表現を対象とした参照関係の解析，JAIST 学術研究リポジトリ，http://hdl.handle.net/10119/1558 (2002)

[9] 日本語 WordNet, http://nlpwww.nict.go.jp/wn-ja/

[10] SLICE NEWS, http://slicenews.net

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-04

タグマッピングによるTwitter特性と話題の関係解析

Relationship Analysis between Twitter's Parameter and Topic using Tag Mapping

清政 貴文[1*]　六井 淳[1]
Takahumi Seimasa[1], Jun Rokui[1]

[1] 島根大学総合理工学研究科
[1] Graduate School of Synthesis Sience and Enginearing, Shimane University

Abstract: We propose the new analysis method of Twitter, which uses 3 types parameters in single Tweet. We show relations of Topics-User-Trend visually based on Hashtag. System contains 3 types Self-Organizing Map used for visualize. We operate maps and analyze interactively.

1　はじめに

近年，世間の思想や流行を解析するための情報源としてTwitter[1]やFacebook[2]などのソーシャルメディアサービスが注目されている．本研究ではTwitter上の話題について興味をもつユーザ層や盛り上がりへの寄与度を解析し，影響力の強い話題の特徴を抽出する手法を提案する．解析にはハッシュタグが付けられているツイートを利用し，1個のツイートからツイート主のフォロワー数などユーザ属性，流行を示す被リツイート数，ツイートに含まれる単語群の3種の情報を抽出する．ハッシュタグを含むツイートはタグが示す話題に関するものであると考え，3種の情報から生成したマップをタグで繋げて各話題を視覚的に解析する．

Twitterユーザの影響力を解析する手法は数多く研究されている．フォロー関係をリンク，発言内容の類似度を重みとしてPageRankを応用する手法 [1] では，より多くのユーザに影響を与えそうな人物の特定に成功している．また，特定の話題において影響力をもつユーザを特定する手法 [2] では，キーワード検索で集めたツイートのリツイートやお気に入り情報を解析している．

ユーザの影響力ではなくタイプを分類する研究も盛んに行われている．ユーザがＢｏｔなのか人間なのかを判定する研究 [3] では，ユーザのフォロー関係やツイートの内容・時刻のパターンを利用して分類している．その他に，ユーザの主なツイートがどのような範囲の集団に向けたものであるかを解析して分類する手法 [4][5] なども存在する．

*連絡先：島根大学総合理工学研究科
〒 690-8504　島根県松江市西川津町 1060
E-mail: s149505@matsu.shimane-u.ac.jp

[1]https://twitter.com/
[2]https://www.facebook.com/

SOM(Self-Organizing Map, 自己組織化マップ)[6] を利用してTwitter上の豪雪トピックを解析する先行研究 [7] では，1種のタグツイートについて単語頻度ベクトルのSOMを構成し，内在する話題同士の類似関係を視覚化している．

SOMをインタフェースに用いてユーザ適応的なWeb検索を行う研究 [8] では，SOMとして表示した検索結果を操作して主観を反映した情報統合を行っている．SOMを用いた対話的推薦システムの研究 [9] では，ユーザのコンテンツ利用履歴から構成したSOMの操作により推薦を行っている．

本稿の2章では収集したデータ対する予備解析の結果を解説し，3章で構築したシステムの機能と仕組みについて解説，検証と考察を行う．4章ではまとめと今後の展望を述べる．

2　対象データ

解析対象としてハッシュタグがつけられたツイートを収集した．対象のタグは2014/11/8〜11/15の予備期間にTwitterStreamingＡＰＩのパブリックストリームから収集した出現頻度上位10%(112233種) から，日本語を含むタグ5839種を抜き出し，そこからランダムに100種類を選択した．その後，Twitter REST APIのsearch/tweetsにより2014/12/1〜12/25の間に出現した各タグツイートを収集した．なお，データの取得にはtwitter4j[10] を利用した．本調査期間で一日平均10回以上利用された89種のタグを解析対象とする．89種で合計1,834,923個のタグツイートを収集し，各タグの平均では20,617個，最も多く集まったタグで168,846個，最少のタグで509個のツイートが集まった．

各タグツイートからは表1に示す情報を取得して解

析に利用している．ツイート主の情報は，タグツイートをしたユーザのプロファイルから取得する情報である．オリジナルツイート主の情報は，タグツイートがリツイートだった場合にツイート主の情報とは別に元のツイートをしたユーザから取得する情報である．リツイート以外の場合，オリジナル情報の各値は全て 0 として扱っている．被リツイート数は各タグが付与されたリツイートについて，収集時点での被リツイート数である．被リツイート数に関して，クロールは期間中 5 分おきに行ったため，各リツイートの生成時点ではなく，クローラが最初に各リツイートを収集した時点での大元の被リツイート数を参照している．なお，被お気に入り数は評価タイミングが難しいことに加え，タイミングを合わせる場合に調査期間初期のツイートの修正に手間が掛かることから除外している．単語の抽出には形態素解析器 kuromoji[11] を利用し，2 文字以上の数でない名詞，または動詞であると分類されたものを単語として扱っている．また，本文から全てのハッシュタグ部分を除いたテキストの単語頻度を求めている．解析対象として選ばれた 89 種以外のハッシュタグも含有タグとして収集しており，単語解析時には本文から除いている．

ツイート主の情報	フレンド数
	フォロワー数
	お気に入り数
オリジナルツイート主の情報	フレンド数
	フォロワー数
	お気に入り数
流行情報	被リツイート数
単語情報	本文単語ヒストグラム
タグ情報	含有ハッシュタグリスト

表 1: タグツイートからの取得データ

解析対象中でツイート数の多い上位 3 種のタグの統計を表 2 に示す．ツイート数は期間中のタグツイート数，利用者数はタグを利用したユーザ数である．ツイート数に比べユーザ数が少ないほど，個々のユーザの呟きが多いことを示す．リツイート数はタグツイートの内リツイートの数であり，R ユーザ数はリツイート元のユーザ数を表す．リツイート数に比べて R ユーザ数が少ないほど，少数のユーザに注目が集まっていることを示す．フレンド平均はタグ利用ユーザのフレンド数の平均を表す．同様にフォロワー平均はフォロワー数の平均を表し，お気に入り平均は利用者のお気に入り数の平均である．ツイート主の統計は，各タグツイートのユーザ数として数えられているユーザを対象に求めている．頭に R がついているパラメータは，それぞれリツイート元のユーザについての統計である．リツイート元の統計は，各タグツイートの R ユーザ数を構成するユーザを対象にしている．各パラメータの偏差は，それぞれ利用者の多様性，リツイート元の多様性を示している．被リツイート平均は全てのタグツイートについて被リツイート数の平均を求めたものである．単語数平均，分散は各タグツイートに含まれる単語数の平均と分散を表す．総単語種は期間中の各タグツイート全体で出現した単語の種類数を表す．

表 2 の結果からは，次のようなことが推測できる．ツイート数とユーザ数の比率から，「#2ch まとめ」は広報用のアカウントに多く利用されている．フレンド平均，フォロワー平均から，「#進撃の巨人」は他 2 種よりも閉じたコミュニティに属すユーザに利用されている．お気に入り平均と R お気に入り平均から，リツイートされる側よりも，する側の方がツイートの収集意欲が高い．ツイート数と総単語数の比から，「#進撃の巨人」は他 2 種よりもツイートの内容が偏っている．

ハッシュタグ名	89 種全体	# アニメ	#2ch まとめ	#進撃の巨人
ツイート数	1,834,923	168,864	142,890	122,592
ユーザ数	180,268	12,504	555	24,475
リツイート数	505,930	23,329	937	66,024
R ユーザ数	12,808	1,235	43	1,357
フレンド平均	599	1,011	1,465	790
フォロワー平均	696	1,263	1,708	886
お気に入り平均	2,898	4,248	4,820	2,441
フレンド偏差	2,550	4,217	8,064	2,894
フォロワー偏差	4,495	7,974	8,795	5,825
お気に入り偏差	12,775	18,814	20,164	9,323
R フレンド平均	1,019	2,020	1,759	1,939
R フォロワー平均	2,355	3,195	1,695	4,568
R お気に入り平均	2,296	2,212	1,290	1,789
R フレンド偏差	6,227	9,333	3,675	8,636
R フォロワー偏差	48,594	19,737	3,743	26,420
R お気に入り偏差	14,272	12,319	7,955	8,605
被リツイート平均	33	4.1	0.015	68
被リツイート偏差	181	28	0.43	274
単語数平均	7.2	7.2	5.0	7.0
単語数偏差	5.2	4.8	4.0	5.0
総単語種	47,740	21,630	21,568	8,377

表 2: 抜粋したタグツイートの統計

また，89 種のタグについて，標本が 89 個あるとして統計パラメータ同士の相関係数を求めた結果，表 3 に示す傾向が得られた．表 3 から，フレンド数が多いユーザに利用されるタグはフォロワー数が多いユーザにも利用されている．これは，先行研究におけるフレンド数とフォロワー数の正の相関 [1][3] がタグ毎の統計にまとめられても有効であることを示している．また，フレンド数，フォロワー数の高いユーザに利用される

タグではツイート数に対するユーザ数，リツイート数が少なくなる傾向が現れている．この傾向から，フレンド数，フォロワー数の高いユーザに使われるタグは個人の使用頻度が高く，新規のタグツイートが生まれやすいと推測できる．次に，個々のユーザが均等に少なくツイートしている話題について，被リツイート数が高くなる傾向が現れている．これは，一過性に近い話題ほどリツイートされやすいと考えられる．お気に入り数が多いユーザに利用されるタグでは，出現する単語の種類が多くなる傾向がある．このことから，収集意欲の高いユーザはタグが示す話題の多様性に貢献している，または，多様な話題を好むと考えられる．

パラメータ組	係数
(フレンド平均, フォロワー平均)	0.886
(フレンド平均, ユーザ数/ツイート数)	-0.358
(フレンド平均, リツイート数/ツイート数)	-0.376
(フォロワー平均, ユーザ数/ツイート数)	-0.349
(フォロワー平均, リツイート数/ツイート数)	-0.426
(被リツイート平均, ユーザ数/ツイート数)	0.417
(お気に入り平均, 総単語種/総単語数)	0.479

表 3: 統計パラメータ間の相関

本章の解析結果から，タグツイートの統計をもとに各話題の傾向を読み取ることが可能であると示唆された．

3 提案システム

3 章では収集したデータをタグ毎に分類して傾向を見たが，フレンド数別や単語別など，様々な観点から解析を行うことでより詳細な傾向を観察できると考えられる．そこで，SOM[6] を用いてデータの特徴を圧縮・抽出し，マップの操作により対話的に解析を行うシステムを提案する．システムの構成を図 1 に示す．提案システムは学習ユニットと解析ユニットからなり，下記の 3 ステップで解析を行う．

1. 対象データを学習する
2. 各 SOM のセルにツイートをマッピングする
3. マップを操作して解析を行う

3.1 SOM の仕様

SOM とは与えられた高次元入力の特徴を低次元のマップに写像する手法であり，提案システムでは各ツイートの特徴を 3 種類の 2 次元マップに写像する．各 SOM への入力を表 4 に示す．ユーザ SOM，リツイート SOM への入力は全てツイートから取得した値をそのまま与える．単語 SOM について, ツイート本文に各

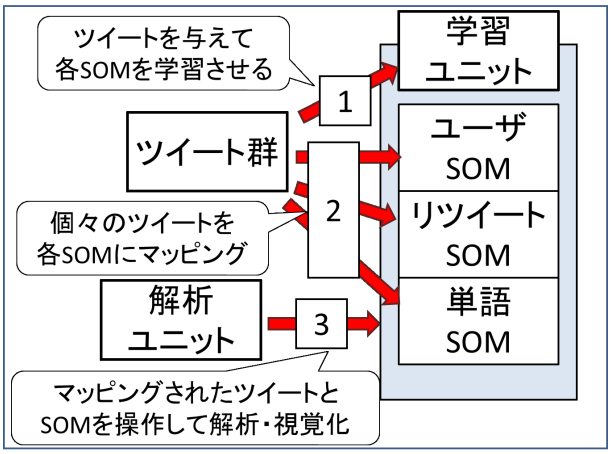


図 1: システムの概要図

タグの代表語が含まれる個数を与える．代表語は各タグツイートの頻出上位 5%の単語について，他のタグの上位 5%に含まれていないものを抽出する．タグ 89 種の内，6 種について代表語が存在しない結果となったが，単語 SOM の入力は 89 次元で行っている．各 SOM のセルはそれぞれに対する入力を表現できる重みをもつ．ユーザ SOM，リツイート SOM のセルの初期化には，表 2 に示す 89 種の全タグツイートについて求めた統計データを基に式 (1) で発生させた正規乱数を利用する．負の値で初期化される場合もそのまま変更していない．単語 SOM の初期化では，各セルに対して 0〜4 の一様な整数乱数を発生させ，その後 89 個の次元から 0〜4 個をランダムに選択して 1 とし，他の次元を 0 とする．学習ではバッチ学習でマップを均した後，逐次学習を行う．勝者セルの探索に用いる距離計算は式 (2) により行う．各パラメータ毎に値の大きさが異なるため，式 (2) ではパラメータ毎の分散を用いて各パラメータの距離の正規化を狙っている．重みの更新について，勝者セルには学習データのパラメータをコピーし，近傍セルの重みはパラメータの差分と勝者ユニットからの距離に応じて近づける．

ユーザ SOM	リツイート SOM	単語 SOM
フレンド数	R フレンド数	タグ 1 の代表語数
フォロワー数	R フォロワー数	タグ 2 の代表語数
お気に入り数	R お気に入り数	…
	被リツイート数	タグ 89 の代表語数

表 4: 各 SOM への入力

$$\begin{cases} p = p\text{の平均} + \text{標準正規乱数} \times\ p\text{の標準偏差} \\ p : \text{ユーザ } SOM,\ \text{リツイート } SOM \text{ の各入力} \end{cases} \tag{1}$$

$$\begin{cases} D_{ij} = \sum_{p \in\ P} d_{ijp} \\ d_{ijp} = (Tjp - Cip)^2 / VARp \\ P : \text{セルのパラメータ集合} \\ Tjp : \text{学習データ } j \text{ のパラメータ } p \\ Cip : \text{セル } i \text{ のパラメータ } p \\ VARp : \text{パラメータ } p \text{ の分散} \end{cases} \tag{2}$$

3.2 システムの機能

システムのメイン画面を図 2 に示す．メイン画面は上部のマップパネルと下部の情報パネルからなる．上部で選択したセル，セルにマッピングされているツイートの情報を下部に表示し，セルに集約された傾向を確認できる．下部は 3 領域に分かれており，それぞれ各 SOM で選択された 1 個のセルに対応した情報を表示する．各領域では対応パラメータに加え，マッピングされたハッシュタグの頻度情報を表示している．また，ある SOM で指定した 1 個以上のセルにマッピングされたツイートが他の SOM でどのように分布しているかを確認することができる．

より詳細な傾向を見る場合，1 個以上指定したセルのマッピングツイートを図 3 に示すサブ画面に表示する．サブ画面では SOM の操作で得た集合を平面に配置し，各領域ごとの傾向を掘り下げて見ることができる．平面に対応するパラメータを変更することで視点を変えた解析が可能である．

図 4 のタグ画面から 3 種までのタグを選択することで，メイン画面・サブ画面におけるタグの分布状況を比較することができる．比較の際は各セル・領域の内，マッピングの最大個数との比を RGB の強さに対応させ，マップの色として重ねて表示している．人気タグと似た傾向をもつ有望なタグの探索，特定のタグに注目した解析が可能である．

3.3 検証

タグ 89 種について集めた 1,834,923 個の全対象データを学習したシステムに対して検証を行った．3 種の SOM 全てに関してマップサイズは 10 × 10，即時的な分類性能を測るため，学習回数はバッチ学習 1 回，逐次学習 1 回と少なく設定している．

学習ユニットの結果において，ユーザ SOM のお気に入り数，リツイート SOM の R お気に入り数，被リツイート数の 3 パラメータは学習後の偏りが強かった．

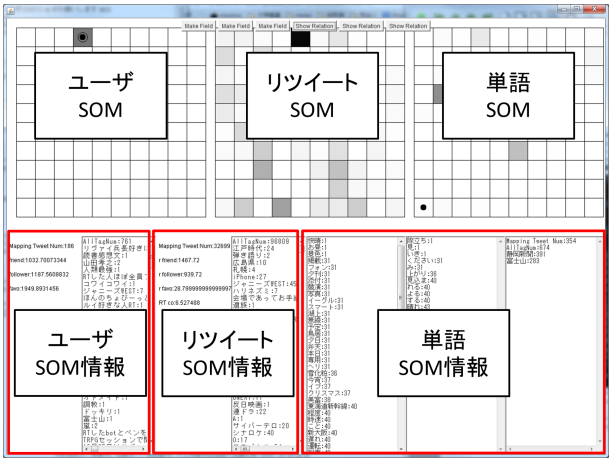


図 2: メイン画面

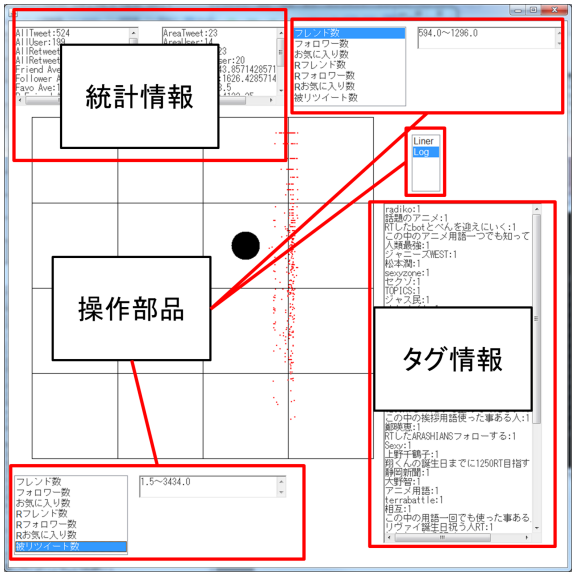


図 3: サブ画面

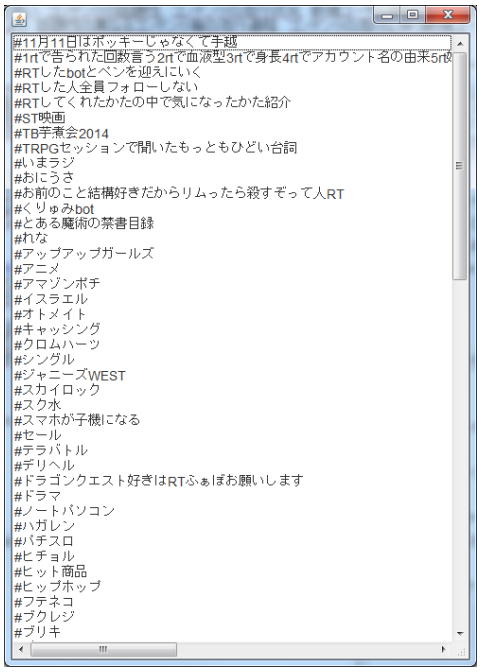


図 4: タグ操作画面

各試行で前述の 3 パラメータが 0 に近いセルが大多数を占め，SOM 上で形成される集団の多様性を狭めていた．そのため，SOM 上でお気に入り数，被リツイート数に着目した解析を行うことができなかった．多様性欠如の原因として，お気に入り数の分散が他に比べて大きく距離計算で軽視されやすいこと，元が同じリツイートを各リツイートで別々に学習させたことが考えられる．話題の盛り上がり要因を解析するためには被リツイート数の多様性が求められるため，リツイート SOM は特に改善が必要である．ユーザ SOM ではマッピング数の偏りや無駄なセル数が最も小さく，良好な学習がなされている．単語 SOM ではマッピング数 0 のセルが目立ち，入力構成や代表語抽出法の改善が必要である．

システムを利用した解析について，タグ操作画面から表 2 に記載した 3 種のタグを選択して分布を比較した画面を図 5，6 に示す．図 5 はユーザ SOM における 3 タグの分布を示しており，赤が「#進撃の巨人」，緑が「#2ch まとめ」，青が「#アニメ」に対応している．各セルの色の強さは全セル中の最大マッピング数との比なので，濃い色のセルが多いほど均等に分布しており，少ないほど集中している．

図 5 では一番右列の中心辺りの赤いセル 1 に「#進撃の巨人」が集中しており，その直上の緑のセル 2 に「#2ch まとめ」，中央よりやや左上の青いセル 3 には「#アニメ」のタグツイートが多くマッピングされている．また，赤と青は緑よりも広くマップ上に点在している．

図 6 は図 5 の画面のユーザ SOM から各タグが最も多くマッピングされている赤，緑，青の 3 セルについて，マッピングツイートを細かく表示したものである．

図 6 では赤点でタグツイートの分布を示しており，右ほどフレンド数が多く，下ほどフォロワー数の多いユーザが呟いたものである．マップ中の線は対数で引いており，一番左上のエリアはフレンド数・フォロワー数が 10 未満，その右隣はフレンド数が 10〜99 でフォロワー数が 10 未満のユーザによるツイートがマッピングされている．各エリアの色の対応は図 5 と同じである．図 6 中のエリア 1〜4 のタグ分布を表 5 に示す.「#進撃の巨人」は主にエリア 1,2，次いでエリア 4 に分布しており,「#2ch まとめ」は大部分がエリア 3,「#アニメ」はエリア 4，次いでエリア 1,2 に存在する．表 2，5 より，各タグツイートの約 1/4〜1/3 が 3 個のセルに集約されており，主利用者層の抽出に成功していると考えられる．

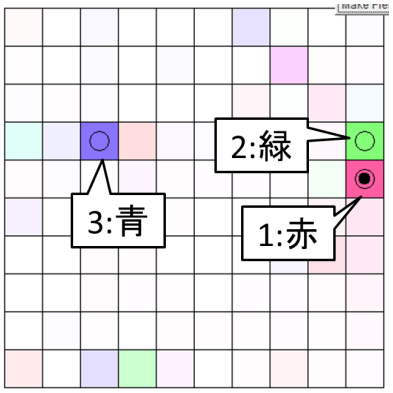


図 5: タグ 3 種のユーザ SOM 上の分布比較

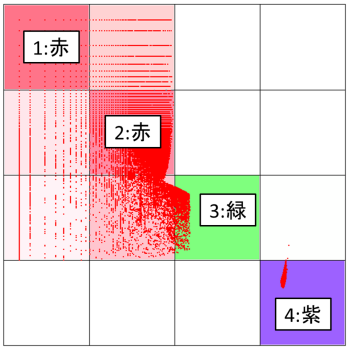


図 6: 注目セル 3 個中のタグ 3 種の分布比較

主利用者層の比較として見ると,「#進撃の巨人」は小規模コミュニティのユーザに多用され,「#2ch まとめ」はフォロワー数の方が多く 100〜1000 人にフォローされているユーザに利用されている.「#アニメ」は両方が 1000 人以上のユーザに多く利用されている.「#進撃の巨人」,「#アニメ」の分布は 2 章のタグ間比較に符合する.「#2ch まとめ」は表 2 に照らし合わせると，主利用者とは別層の人気アカウントにも利用されていると考えられる．セル 3 個分のツイートでは各タグの全

	ツイート数	ユーザ数	#アニメ	#2chまとめ	#進撃の巨人
3 セル合計	442,675	40,197	54,406	57,455	33,613
エリア 1	63,229	5,090	5,256	2	10,198
エリア 2	100,867	26,150	6,097	96	8.265
エリア 3	61,165	1,243	1,196	56,447	375
エリア 4	91,555	1,140	35,852	180	3,888

表 5: 図 6 のタグ分布

体傾向を見れていないが，各利用者層の視覚的解析は成功している．SOM の選択セル数を増やして各タグの解析範囲を広げることで表 2 を踏まえた結果に近付くと考えられる．

3.4 考察

検証結果より，提案システムにより対象データの傾向を発見できる可能性が示唆された．そのため，ユーザと話題の関係を解析する，という観点において提案システムは有効だと言える．

盛り上がりの要因解析についてはリツイート SOM の学習が上手く行えておらず，現状の提案システムでは不十分である．学習手法や入力構成の改善によりリツイート SOM の分類性能が向上すれば有用になると考えられる．

また，本研究では個々のツイートを主体としているため，ユーザ側のフォロー関係や前後の発言内容は考慮していない．そのため，各話題に参加している Bot や広告用アカウントなどを区別できておらず，話題の性質をより詳細に解析するためには個々のユーザを詳しく見る必要がある．また，各ユーザの Twitter 利用期間を考慮せずに各プロファイルを参照しており，各パラメータの高低が充分にユーザの利用傾向を表しているとは限らないため，改善の余地が大きい．

対象データ内において各タグの利用者特性に差異が認められたため，ユーザの影響力やタイプを解析する研究 [1][2][3][4][5] などを包含した，当該話題に対して各系統のユーザがどの程度参加しているか，という指標が Twitter 解析において有効であることが示唆された．

4 まとめ

本稿ではツイート情報を 3 種の SOM に学習させ，視点を変えたマップ表示により対話的に解析する手法を提案した．対象データの静的解析と提案システムを用いた解析の比較から，提案システムは一定の解析性能を有することが示された．今後は考察で述べた改善に取り組むと共に，細かい期間におけるタグツイートの統計情報とタグツイート数の増減の関係を解析して流行を予測する手法を研究する予定である．

参考文献

[1] Weng, Jianshu. , Lim, Ee-Peng. , et al.: Twitterrank: finding topic-sensitive influential twitterers, *Proceedings of the third ACM international conference on Web search and data mining. ACM*, (2010)

[2] Noro, Tomoya. , Ru, Fei. , et al.: Twitter user rank using keyword search, *Information Modelling and Knowledge Bases XXIV. Frontiers in Artificial Intelligence and Applications*, Vol.251, pp.31-48 (2013)

[3] Chu, Zi. Gianvecchio, Steven. , et al.: Who is tweeting on Twitter: human, bot, or cyborg?, *Proceedings of the 26th annual computer security applications conference. ACM*, (2010)

[4] 竹村 光, 田島 敬史: 情報発信の対象範囲に基づく Twitter ユーザの分類, *DEIM Forum B1-6*, (2013)

[5] Yan, Liang. , Ma, Qiang. , et al.: Classifying Twitter Users for Spatio-temporal Entity Retrieval, 電子情報通信学会技術研究報告. *DE,* データ工学, Vol.112, No.346, pp.93-98 (2012)

[6] Kohonen, Teuvo.: The self-organizing map, *Neurocomputing*, Vol.21, No.1, pp.1-6 (1998)

[7] 澤田 義人, 他: 2011 年山陰豪雪に関連する Twitter メッセージ解析法の開発, 生産研究, Vol.64, No.4, pp.467-473 (2012)

[8] 佐野 綾一, 波多野 賢治, 田中 克己: 自己組織化のマップを用いた Web 文書の対話的分類とその視覚化, 情報処理学会研究報告. データベース・システム研究会報告, Vol.98, No.57, pp.33-40 (1998)

[9] 藤森 洋昌, 土方 嘉徳, 西田 正吾: 協調フィルタリングにおける近傍グループの可視化, 情報処理学会研究報告. 情報学基礎研究会報告, pp.59-66 (2004)

[10] twitter4j-3.0.4:http://twitter4j.org/ja/index.html

[11] kuromoji-0.7.7:
http://www.atilika.com/ja/products/kuromoji.html

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-05

キーワードのシソーラス上の位置関係にもとづく文章の話題の推敲支援

Polishment of Document Topic by Keyword Relationship in a Thesaurus

大野 祐樹 [*1]　　砂山 渡 [*1]
Yuuki Ohno　　Wataru Sunayama

[1] 広島市立大学 情報科学部
Faculty of Information Sciences, Hiroshima City University*

Abstract: 文章の推敲支援の多くは，表層的な修正を促すものが多く，文章の主題や主題に関連する話題の吟味を促すものはあまり見られない．そこで本研究では，文章からキーワード（文章の主題および話題を表す単語）を抽出した上で，それらのシソーラス上の意味のつながりをもとに，文章内で述べられている内容を推敲するための指標を計算して利用者に提示し，利用者の推敲を促すシステムを提案する．

1　はじめに

近年，インターネットの普及により Twitter や Facebook 等の SNS サイトやブログを通じて誰でも手軽に情報発信ができるようになった．インターネット上では自分の発信した情報を必ずしも親しい人だけが見ているとは限らない．名前も顔も知らない相手に対して文章だけで誤解なく意図を伝えるためには，文章を見直し改める推敲が必要となる．

近年では，フリーソフトの推敲支援ツール [1] や [2] があり，手軽に文章の推敲を行う事ができるようになった．しかし，従来のツールでは文章の文法間違いや適切ではない単語を見つけるなどの表面的な推敲はできても，自分が主に述べたいことを見直す話題自体の見直しや推敲を行う事はできない．その要因として，自分の主張が文章中でどのようなキーワードとして出現しているかわからないことや，それらをどのように推敲すれば，読み手に意図が伝わりやすくなるのか分からないということが挙げられる．

そこで本研究では文章中に現れる筆者の主張を特徴づける単語をキーワードとして抜き出した上で，それら抽出したキーワード集合が，筆者の期待する話題としてふさわしいかを検証するための指標を提示し，話題の推敲を支援するシステムの構築を目指す．本稿では，具体的な話題の推敲支援システムの構築に向けて策定した指標と，その有効性について検証した結果について述べる．

*(連絡先) 砂山渡, 731-3194, 広島市安佐南区大塚東 3-4-1, 広島市立大学大学院情報科学研究科, sunayama@hiroshima-cu.ac.jp

2　関連研究

文章の推敲を支援するための研究は，PC があまり普及していない時代から行われてきており [3]．文法の正しさや誤字の検出など，表層的な指摘を行った上で文章の体裁を整える支援をする研究や，修正を促すシステムはこれまでに開発されてきている [4]．また表層的な表現に加え，談話レベルでの推敲を促す研究 [5] もある．これらの研究とは，文章の特徴を抽出し推敲支援を行う点で類似しているが，本研究では，表層的な表現，また言い回しなどの表現の推敲ではなく，表現のおおもととなる話題の推敲を扱う点で異なる．

表層的な表現の推敲に加え，キーワードを抽出し文章の特徴づけにより推敲の支援を行う研究 [6] もあり，キーワードを抽出することで，文章の特徴を捉えるという点で類似している．本研究では，抽出した各キーワードの位置づけに基づいて，それらをどのように扱うべきかの指標を与える点で異なる．

また，話題に対して指針を与える研究 [7] もあり，文章の中から話題を抽出し話題に対する推敲を支援する点で類似している．しかし，与える指針はふさわしくない話題の削除を促すものとなっており，本研究では，ふさわしくない話題に対してだけでなく，良い話題を広げる指針としての活用も期待できる．

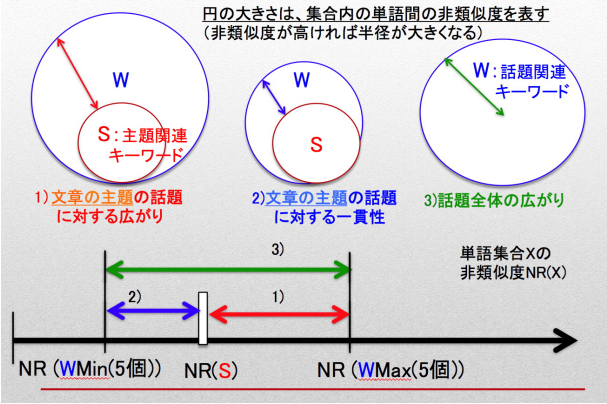


図 1: 話題推敲のための指標とキーワードとの関係

3 文章の話題の推敲支援のための指標

文章の話題の推敲に役立てられる指標として，以下の 3 つを用意する．

1 文章の主題の話題に対する広がり

2 文章の主題の話題に対する一貫性

3 話題全体の広がり

1 は，文章の主題に関連した話題が幅広く述べられているかを確認するため，2 は，文章の主題に関連した話題のみでまとめられているかを確認するため，3 は，文章が話題の全体が幅広い内容を取り扱っているかを確認するための指標となる．

これらの指標の計算のために，以下の 3 種類のキーワードを文章から抽出する．

a 主題キーワード：文章のテーマとなるキーワード

b 主題関連キーワード：文章に現れる筆者の主張を表すキーワード

c 話題関連キーワード:文章の特徴を表すキーワード

この 3 種類のキーワードは，文章内の名詞の出現頻度を用いて抽出する．これは，文章の主題、話題に関わる単語ほど多く出現し，また読み手にそう解釈される可能性が高いと考えたことによる．すなわち,「主題キーワード」は再頻出語,「主題関連キーワード」は頻度上位 5 単語,「話題関連キーワード」は頻度上位 10 単語とする．

1 に，文章の話題の推敲に役立てられる指標とキーワードとの関係を表した図を示す．すなわち，単語集合が与えられたときに，その単語集合内の単語がどれだけ似ているか似ていないかを表す非類似度を定義し，非類似度が高いほど，その単語集合が表す話題の広がりが大きいと考える．その上で，話題の広がりや一貫性を表す指標を計算する．以下で，この各指標の計算方法について述べる．

3.1 単語集合の非類似度

単語集合 X の非類似度 $NR(X)$ は，シソーラス（ある概念に沿って上位-下位のリンクでつながれた木構造のデータベース）内の単語集合の位置関係に基づいて計算する．本研究ではシソーラスに日本語 WordNet[8] を用い，その中の上位語と下位語のリンクでつながれた，全ての名詞のシソーラス内の位置情報を利用することとした．

まず，単語 w の深さ Depth(w) を，ルートノードから単語 w のノードまでの階層（リンク）の数，単語 w の高さ高さ Height(w) を，単語 w のボトムノード (w_b) からの階層の数として，式 (1) で表す．

$$Height(w) = Depth(w_b) - Depth(w) \qquad (1)$$

ただし，ノードはシソーラス上の一つの単語，ルートノードはシソーラス内の最上位語，ボトムノード w_b はシソーラス内で一番深い（ルートノードから最も遠い）ノードとする．

次に単語の非類似度について，単語 $W = \{w_1, w_2, ..., w_n\}$ の非類似度 $NR(W)$ を，式 (2) で与える．すなわち，式 (3) で表される単語間の相対的な距離の遠さと，式 (3) で表されるシソーラスの構造によるお互いの類似性の積によって，単語集合の非類似度を表す．式 (3) は，各単語がシソーラス内で深い位置にあるほどお互いの相対的な距離が遠くなることにより定めた．また式 (3) の sh_k は，シソーラス内で各単語をリンクに沿って上位にたどったときに，各単語がシソーラス内で交わるノードの高さを表す（n 個の単語があったとき，それらはシソーラス上で最大 $n-1$ 箇所で交わる）．これにより，シソーラス内で各単語が上の方で交わるほど，お互いの類似性が低いと考えたことにより定めた．

$$NR(W) = RD(W) \times RH(W) \qquad (2)$$

$$RD(W) = \prod_{i=1}^{n} Depth(w_i) \qquad (3)$$

$$RH(W) = \prod_{k=1}^{n-1} sh_k \qquad (4)$$

3.2 文章推敲のための指標

1 に示す 3 つの指標を非類似度を用いて計算する方法について述べる．

3.2.1 文章の主題の話題に対する広がり

文章の主題の話題に対する広がりの指標 C_1 は，主題関連キーワード集合 S の非類似度 $NR(S)$ と，話題関連キーワード集合 W の部分集合として作られる 5 個の単語による非類似度のうち，非類似度が最大になる単語の部分集合 W_{max} の非類似度 $NR(W_{max})$ の差として式 (5) で与える．これにより，主題を表す単語に対して，どの程度広い話題が取り扱われているかを測ることで，主題に関連してどの程度話題が広げられているかを確認できる．

$$C_1 = NR(W_{max}) - NR(S) \tag{5}$$

3.2.2 文章の主題の話題に対する広がり

文章の主題の話題に対する一貫性の指標 C_2 は，主題関連キーワード集合 S の非類似度 $NR(S)$ と，話題関連キーワード集合 W の部分集合として作られる 5 個の単語による非類似度のうち，非類似度が最小になる単語の部分集合 W_{min} の非類似度 $NR(W_{min})$ の差として式 (6) 計算する．これにより，主題を表す単語に対して，取り扱われている話題の狭さ，すなわち一貫性の程度を確認できる．

$$C_2 = NR(S) - NR(W_{min}) \tag{6}$$

3.2.3 話題全体の広がり

話題全体の広がりの指標 C_3 は，先の述べた非類似度 $NR(W_{max})$ と非類似度 $NR(W_{min})$ との差として式 (7) で与える．すなわち，やみくもに単語の類似性がないことを話題の広がりと呼ぶのではなく，話題集合の中でも，核となる類似性が高い単語集合の非類似度 $NR(W_{min})$ に対して，どの程度話題が広げられているかを図る．

$$C_3 = NR(W_{max}) - NR(W_{min}) \tag{7}$$

4 文章の話題の推敲支援に用いる指標の有意性検証実験

4.1 単語集合内の単語の非類似度とストーリーとの関係の調査

5 個の主題関連キーワード集合 S 内の単語間の非類似度と，文章のストーリーの想像のしやすさとの関係を調査する実験を行った．実験は，40 人の男女に対して，類似性のパターンが異なる 7 種類のキーワード集合 3 セットに対して，文章のストーリーを想像しやすい順に並べてもらった．用意したパターンの単語の類似性について，シソーラスとして用いている WordNet 内において，共通の親ノードをもつ単語間には類似性がある，またそうでない単語には類似性がないとして，単語のパターンを生成した．実験に用いた単語の例を 1 に示す．

表 1: 用意したキーワード集合と類似性パターンの例 (括弧でまとられた単語間には類似性がある)

パターン	使用したキーワード
5	(戦争，敵，反逆，知恵比べ，縄張り争い)
1,1,1,1,1	(戦争)(選挙)(出席)(仲裁)(同盟)
2,2,1	(戦争，敵)(選挙，市長)(出席)
3,1,1	(戦争，敵，反逆)(選挙)(出席)
4,1	(戦争，敵，反逆，知恵比べ)(選挙)
2,1,1,1	(戦争，敵)(選挙)(出席)(仲裁)
3,2	(戦争，敵，反逆)(選挙，市長)

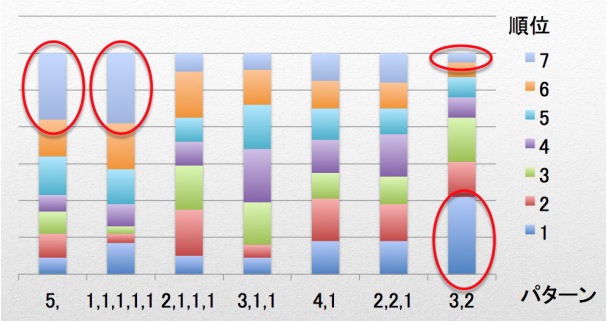


図 2: キーワードの類似度とストーリーの想像のしやすさ

図 2 に，キーワードの類似度とストーリーの想像のしやすさの関係の結果を示す．この結果から，類義語がパターン (3,2)，(2,2,1) や (4,1) のように，バランスよく含まれているほどストーリーを想像しやすかったことがわかる．特にパターン (5) のように，すべての単語が類似している場合，単語が表す範囲が狭すぎて，ストーリーが想像しにくい，またパターン (1,1,1,1,1) のように，すべての単語が類似していない場合は，単語間の関連によるストーリーが想像しにくくなったと考えられる．そのため，単語集合には一定の類似性と非類似性を併せ持つことが，ストーリーの想像には有効となることがわかった．このことから，本研究で提案した主題関連キーワード集合 S の非類似度 $NR(S)$ が，$NR(W_{max})$ と $NR(W_{min})$ の中間の値に近い，すなわち C_1 と C_2 の値が近いほど，文章のストーリーがわかりやすいと考えられ，これらをそのための指標として用いられる可能性を確認した．

表 2: 実験に用いた文章の主題キーワードと W_{max} と W_{min} との関係

文章	W_{max} が含む	W_{min} が含む
1	○	×
2	×	○
3	○	○
4	×	×
5	×	○
6	○	○
7	×	×
8	○	×

表 3: 話題の広がりの評価結果

文章	W_{max} が含む	評価の平均
6	○	7.1
3	○	7.0
1	○	6.9
8	○	6.4
5	×	5.5
2	×	5.3
4	×	5.1
7	×	4.3

4.2 話題の幅広さと一貫性の指標の検証

主題キーワードが単語集合 W_{max} に含まれていると，主題に関連した話題が幅広く述べられていると言えるか，また主題キーワードが W_{min} に含まれていると，主題に関連した話題のみでまとめられていると言えるかを検証する実験を行った．実験は 40 人の男女に 800 字程度の 8 つの文章を被験者に読んでもらい，各文章について,「文章の話題の広がり」と「文章の一貫性」を 10 段階で評価してもらうことで行った．用意した各文章の W_{max} と W_{min} が，主題キーワードを含むか否かについてまとめたものを表 2 に示す．

表 3 に話題の広がりの結果を示す．主題キーワードが W_{max} に含まれていると文章の話題の広がりの評価結果は高くなった．このことから，話題の広がりを大きくしたい時は、W_{max} に主題キーワードが含まれるように修正を促すことができると考えられる．

表 4 に話題の広がりの結果を示す．主題キーワードが W_{min} に含まれていると文章の一貫性の評価結果は高くなった．このことから，文章に一貫性をもたせたい時は，W_{min} に主題キーワードが含まれるように修正を促すことができると考えられる．

5 おわりに

本稿では，キーワードのシソーラス上の位置関係にもとづいて，文章の話題の推敲に用いられる指標を提案した．評価実験により，指標を推敲に有効に役立てられる可能性を検証した．

今後は，話題の推敲支援システムとして具体的に実装と評価を行っていきたい．

表 4: 話題の一貫性の評価結果

文章	W_{min} が含む	評価の平均
6	○	7.9
5	○	7.1
2	○	7.1
3	○	6.4
7	×	5.4
1	×	5.4
8	×	5.3
4	×	4.6

参考文献

[1] 日本語小論文評価採点システム Jess：(URL) http://coca.rd.dnc.ac.jp/jess (2015/3/6 access)

[2] 森リン，日本語の文章解析ソフト：(URL) http://www.mori7.info/moririn/ (2015/3/6 access)

[3] 倉田昌典，菅沼明，牛島和夫：日本語文章推敲支援ツール『推敲』のパソコン上での実用化，コンピュータソフトウェア，Vol.6, No.4, pp.373-385 (1989)

[4] 奥村有希，大野博之，稲積宏誠：技術文章作成支援ツールの推敲支援機能の拡張?長い修飾節に起因する悪文の検出手法の提案，教育システム情報学会研究報告，Vol.22, No.6, pp.186–191 (2008)

[5] 飯田龍，徳永健伸：談話レベルの推敲支援のための人手修正基準，言語処理学会第 19 回年次大会発表論文集，pp.830–833 (2013)

[6] 石岡恒憲．亀田 雅之：コンピュータによる小論文の自動採点システム Jess の試作，計算機統計学，Vol.16, No.1, pp.3–19 (2003)

[7] 菅沼明，小野貴博：文章推敲支援における読み手に誤解される文の抽出，情報処理学会研究報告，DD, Vol.2007, No.50, pp.31–38 (2007)

[8] 日本語 WordNet：(URL) http://nlpwww.nict.go.jp/wn-ja/ (2015/3/6 access)

[9] 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)

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-06

TETDM を用いた文章推敲スキル育成のための チュートリアルシステムの開発

Development of Tutorial System for Acquiring Document Polishing Skills by Using TETDM

中垣内　李菜[1]　川本　佳代[1*]　砂山　渡[1]

Rina Nakagochi[1], Kayo Kawamoto[1], and Wataru Sunayama[1]

[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

Abstract: Document polishing is indispensable to create easy-to-read text. Hence, in this paper, we defined document polishing skills as to find the problems in the document and to consider modified policy. Then we proposed a tutorial system which can acquire document polishing skills by using TETDM and showed the effect of the system by experiment. According to experiments, users of the system was able to polish document not only from the local perspective but also from global perspective.

1　はじめに

あらゆる職種・分野において，文章を書く機会はある．学生ならばレポートや論文，報道機関で働く人ならばニュース原稿や新聞記事などさまざまな文章がある．文章は文学的文章と説明的文章に大別される．文学的文章が人物の心情や出来事など，主観を含む文で構成されるのに対し，説明的文章は，ある物事について，それがどんな性質，特質であるかなど，物事に関する具体的な説明を明確に，さらに客観的な視点から示さなければならない．読者に伝わりやすい説明的文章を書くためには，筆者は誤字脱字をしないことはもちろん, 文章の構造を意識し, 物事について論理的に順序良く説明する必要がある. さらに，よりよい文章を書くためには，執筆が終わった後に推敲を行うことが重要である．そこで本研究では，説明的文章を対象として，テキストマイニングを用いて文章中の問題点を見つけ出して，修正方針を検討する練習ができるようなチュートリアルシステムを開発し，提供することによって文章推敲のスキルを育成することを目的とする．

なお, 本研究では, テキストマイニングを用いて, 文章中の悪い点を見つけ出して修正方針を検討することを文章推敲スキルと定義する.

*連絡先：広島市立大学大学院情報科学研究科システム工学専攻
〒731-3194 広島市安佐南区大塚東三丁目 4 番 1 号
E-mail: kayo@sys.info.hiroshima-cu.ac.jp

2　関連研究

2.1　既存の推敲システム

推敲支援システムの研究はこれまでも行われてきた．例えば，文章中の問題となりそうな箇所を，可能な限り高速で探し出して指摘するという目的で開発された「推敲」というシステム[1]がある．このシステムは，推敲作業における有用な情報として，指示語，受身形，二重否定の指摘など，文章中の細部に着目して推敲を支援する．一方で，一文または一段落に着目し，文や段落内の構成をグラフで表す可視化システム[2]がある．このシステムでは，ある文(段落)の中に含まれる修飾語と被修飾語の間に多数の語が挟まれている場合，文(段落)単位のグラフ(図1)を作成してユーザに表示する. ユーザは, グラフをみて意図どおりの構文解析が行われていない箇所について修正を行う．いずれのシステムも，文章推敲の参考となりそうな項目すべてを提示するのではなく，ある一部の項目に特化して推敲支援を行っている．本研究では，細部や文構成のみに焦点を当てるのではなく，一つのシステムで文章中の局所的な部分から大局的な部分までに焦点を当てて推敲を支援できるようなシステムを作成した．

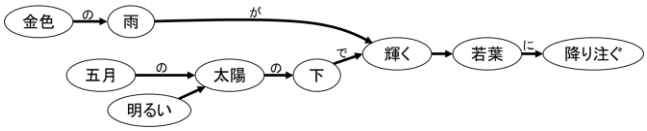


図 1：「金色の雨が五月の明るい太陽の下で輝く若葉に降り注ぐ」という文を可視化した図

2.2 小論文の自動採点システム

小論文をコンピュータで自動評価する試みは数多く行われている[3][4]．そのなかでも，最も有名であり，他の研究でも参考にされるシステムとして E-rater がある．E-rater は，アメリカのテスト機関 Educational Testing Service, ETS の Burstein らの研究グループにより開発された自動採点システムである．E-rater は，アメリカの経営大学院の入学試験における小論文の採点に用いられている．E-rater は，構造・組織化・内容の 3 つの観点から小論文を評価し，評価結果を 6 点満点で示す．なお，E-rater と専門家による採点の一致率，すなわち，両者の採点結果が 1 点差以内となったのは，97%と検証されている[5]．日本語を処理する小論文の自動採点システムに石岡らが開発した Jess[6]がある．このシステムは，質問文と解答文が入力されると，入力した情報をもとに小論文を自動的に 10 点満点で採点する．Jess の採点の観点は，E-rater の採点の観点を踏襲している．Jess と E-rater の採点結果を比較すると，得点がかなり一致したことがわかっている．このことから，Jess も信頼できるものと考えられる．本研究では，Jess の s 採点の観点を参考に課題を作成した．ただし，本研究では，対象が説明的文章であること・分析方法が異なること・採点が目的ではないことなどの相違点があるため，採点基準をそのまま使用せず，目的に適すよう採点基準を検討し，適宜利用した．検討した観点については，第 3.3.1 節で詳しく述べる．

2.3 TETDM

TETDM[7]は，豊富なツールが提供されているテキストマイニングの統合環境である．TETDM は，幅広い利用者と開発者の参入を目的に，多くのツールを比較的容易に実装および利用できる環境を構築し，無償で一般公開している[8]．TETDM に関するこれまでの研究には，処理ツールや可視化ツールの開発に関する様々な研究[9][10]があるが，ツールの見方や使い方の説明が十分には備わっていない．本研究では，チュートリアルシステムを TETDM の機能の一つとして提供する．なお，本研究では，公開中の TETDM(Ver.0.62)を改良してチュートリアルを作成した．

2.4 既存のテキストマイニングツールのチュートリアルやマニュアル

テキストマイニングツールは数多く存在する．現在，日本で公開されているテキストマイニングツールには DIAMining[11]，Text Mining Studio[12]などがある．いずれのツールも有償であり，ツールとは別に独立した大まかな分析手順をインターネット上で Flash コンテンツやプレゼンテーションのスライド形式で公開している．しかし，これらのコンテンツはあくまでツールを使ってどのように分析作業を行えるかを一通り知りたいユーザ，例えば購入を検討しているユーザを対象にしており，実際にツールを使用するときに参考にするには内容が十分ではない．製品を購入すると詳しい説明が書かれたテキストを取得できるが，ごく一般的な目的のために汎用的なツールを紹介するにとどまり，ユーザには，高いモチベーションを維持し各自で模索してテキストマイニングの知識を得ながら，経験を積むことが求められる．本研究では，ユーザにテキストマイニングを実際に行わせながらその都度最小限の情報を提示して，自然と知識と技術を得ることができるようなチュートリアルを作成した．

3 チュートリアルシステム

本チュートリアルシステムの使用時のフローチャートを図 2 に示す．まず，システムに事前に作成した必要な情報を記述した複数のテキストファイルを入れたフォルダを与える．これに対し，システムは入力された情報をもとに課題データを作成して，各課題を一覧できる「チュートリアルウィンドウ」(図 3)を提示する．次にユーザがこの中から任意の課題を選択すると，システムは，選択した課題について，「課題の詳細ウィンドウ」(図 5)上で詳細を提示する．そして，ユーザは課題の詳細を読んで分析の目的を理解し，ツールをセットし，ツールの使用法を学習し，結果と解釈を考案する．最後に，ユーザが TETDM の機能を用いて結果と解釈を登録すると，チュートリアルシステムは入力結果が課題クリアの条件を満たしているかを判定する．

以降，本研究では，「ツール」を TETDM で用意されているテキストマイニングツールのこと，「セット」をユーザがツールを選択して使用可能な状態にすること，「課題」をチュートリアルで提供する 1 つの話題についてまとめた学習単元のことと定義する．

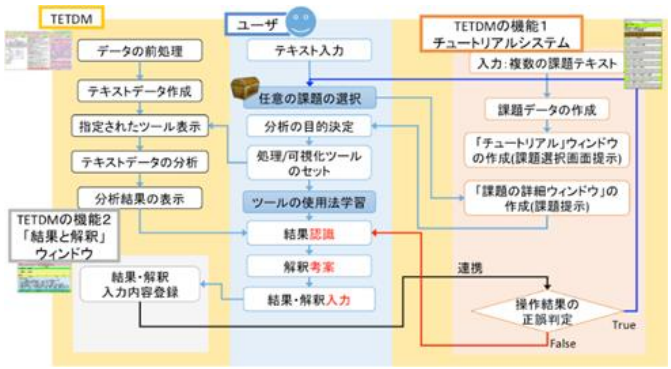


図 2：システム使用時のフローチャート

3.1　入力：複数の課題テキスト

本チュートリアルシステムには，あらかじめ作成した複数の課題に関するテキストを含むフォルダを入力として与える．テキストは，チュートリアルの各課題の具体的な内容を記したものや，各課題に関する情報と，課題を内容に基づき大まかに分類するためのカテゴリに関する情報を記したものがある．カテゴリに関する情報には，カテゴリの通し番号や名前があり，各課題に関する情報には通し番号や名前，取得できる経験値，所属するカテゴリの通し番号，クリア状況などがある．

3.2　課題データの作成

入力された情報をもとに，各カテゴリおよび各課題に対する課題データを作成する．

3.3　出力:「チュートリアルウィンドウ」の作成

課題データを作成して，各課題を一覧できる「チュートリアルウィンドウ」を作成する．図 3 にチュートリアルウィンドウの表示例を示す．チュートリアルウィンドウの画面の上部には，学習者のレベルや経験値，使用者がマウスのカーソルを合わせている課題のタイトルを表示する．ウィンドウ下部には，各課題に対応した複数の宝箱を表示する．宝箱は内容によって複数のカテゴリに分けられており，各カテゴリはカテゴリ名と，カテゴリに属する課題群で構成される．

図 3：チュートリアルウィンドウの表示例

3.3.1　提示する課題の内容

推敲スキルを獲得するためには，まず大局的な視点に立って文章の段落レベルに着目し，後に文レベル，単語レベルと徐々に局所的な視点に視野を狭めながら徹底的に文章をチェックできるような課題が必要である．各段階の視野の広さのことを視点レベルと定義する. 本研究では, 具体的な視点レベルを，広いほうから全体・全体的な段落間・局所的な段落間・1 つの段落内・全体的な文間・局所的な文間・1 つの文内・単語表現の 8 つとした．

以上の各視点レベルを踏まえて，TETDM で提供されているツールから，文章を書くときに利用できそうなツール 29 種類を選び，課題を作成した．課題は，9 種類(MISSION 0 から MISSION 8)の 36 個を用意した．ユーザが最初に取り組む MISSION0 では，チュートリアルの使い方および進め方を学ぶことができる. MISSION 1～8 の内容は各視点レベルに対応しており，MISSION1 から 4 では文章全体の大まかな構成に着目した課題を，MISSION5 から 8 では細かいポイントに着目した課題を学ぶことができる．具体的な課題の内容を表 1 に示す．さらに，各 MISSION 内の課題は, 1 つ目の課題に各 MISSION の目的を学ぶものを設定し，以降の課題で実際にユーザが練習を行って学習するものを設定した．

ユーザが実際に練習を行う課題は，すべてが同一の重要度ではなく，こなせばある程度の文章推敲スキルが付くように最低限に絞った重要度が高い課題, その他の重要度が低い課題の 2 種類に設定した．重要度は，第 2.2 節で述べた Jess の採点基準をもとに設定した．ただし，本研究では，対象が説明的文章

であること，分析方法が異なること，目的が採点ではないことなどの相違点があることから，以上の Jess の採点基準をそのまま採用することはできない．そのため，本研究の目的である文章推敲スキルを育成することに適するよう，修辞・論理構成・内容の各観点で着目する項目を以下のように設定した．

- 修辞
  文の長さ，文中の主語の有無，曖昧表現の有無，難読漢字の有無
- 論理構成
  段落の構造，主題一貫性
- 内容
  文章の採点結果，単語の頻度，主題語，類似度，重要文

実装した課題のうち，以上の項目を含む課題は重要度が高い課題，含まない課題は重要度が低い課題と設定した．

表 1：提示する課題の内容

カテゴリ	課題内容
MISSION 0	このチュートリアルの使い方：チュートリアルの進め方について
MISSION 1	視点「文章の構成に関する内容」：文章の大まかな雰囲気・構成を掴む練習について
MISSION 2	視点「全体的な段落間の関係」：文章中の段落の構成について，段落が適切な箇所，バランスで分けられているか，スムーズに読める文章を構成できているかを調べる手段について
MISSION 3	視点「局所的な段落間の関係」：文章中の一部の段落の構成について，段落が適切な箇所，バランスで分けられているか，段落間のつながりが明確かを調べる手段について
MISSION 4	視点「段落の表現に関する内容」：文章中の段落中の表現について，段落の主張点が明確か，主張点に対して一貫性があるかを調べる手段について
MISSION 5	視点「全体的な文間の関係」：文章中の文の構成について，文章の主張点に対して一貫性のある文で，無駄なく構成できているかを調べる手段について
MISSION 6	視点「局所的な文間の関係」：文章中の一部の文に着目して，文章の内容を把握する手段について
MISSION 7	視点「文の表現に関する内容」：文章中の文に関して，読者が読みにくかったり，誤解を招く可能性のあるポイントを掴む手段について
MISSION 8	視点「単語の表現に関する内容」：文章中の単語に関して，表現に問題がないかを確認する手段について

3.3.2 モチベーション維持のためのゲーム的要素

チュートリアルウィンドウ内で表示されるアイコンは以下に示す 3 種類があり，内容に応じて使い分けてある．

- 本＋鍵型　：チュートリアルやカテゴリの説明
- 上級宝箱型：重要度が高い課題
- 一般宝箱型：重要度が低い課題

また，各アイコンは，状況により画像が変化する．その変化を図 4 に示す．本ウィンドウは，初期状態では最初のカテゴリのすべての課題と，2 つ目のカテゴリの最初の課題が挑戦可，それ以外のすべての課題が挑戦不可のアイコンで描画される．常にすべての宝箱に挑戦可だった場合，情報量が多すぎて，ユーザがどの課題から挑戦すれば良いか迷ってしまう恐れがあるためである．挑戦可のアイコンの数を限定して描画することにより，情報量の削減を行なった．なお，挑戦中のカテゴリの最後の必修課題をクリアすることによって，次のカテゴリの課題群は挑戦可の宝箱となる．また，各課題をクリアするごとに経験値が加算され，獲得した経験値による学習者のレベルを SKILL LEVEL として表示した．学習者がモチベーションを維持しながら多くの知識やスキルを獲得できるように，ゲーム的要素としてこれらを実装した．

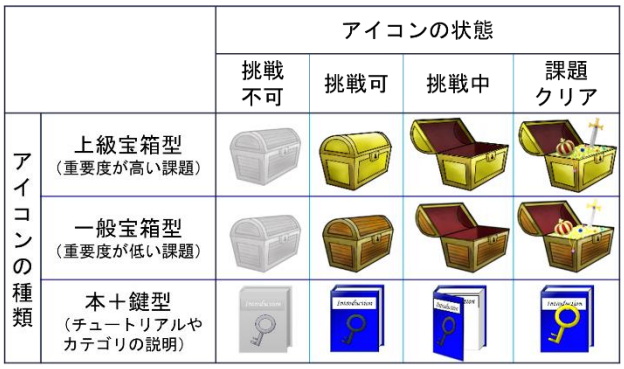


図 4：宝箱の状態

3.4 ユーザの入力：課題の選択

ユーザが「チュートリアルウィンドウ」上で任意の課題を選択すると，選択された課題に応じてシステムが「課題の詳細ウィンドウ」を作成する．図 5 に課題の詳細ウィンドウの表示例を示す．課題の詳細ウィンドウでは最上部に課題のタイトルを，中央部に課題の詳細を表示する．また，下部にはページ送りボタン(back <，next >)，画像表示ボタン(figure)，ウィンドウを閉じるボタン(close)を配置した．1 つの課題は，主に以下のステップで構成され，1 つのステップを 1 ページに表示する．

[STEP1]学習する推敲スキルを把握する
[STEP2]指定されたツールをセットし, 概要を学ぶ
[STEP3]ツールの見方・使い方を学ぶ
[STEP4]課題に挑戦する

ユーザは, [STEP1]から[STEP3]までの説明を読み, [STEP4]の指示に従って課題に挑戦する．課題では, もともと TETDM に備わっている機能を用いて,「結果」と「解釈」を入力し,「登録」ボタンを押して登録を確定する．なお，結果には，表示されたテキストマイニングの分析結果から探し出した悪い/良い点/気になったところの具体的な内容を，解釈には結果をどのように修正するか/どのような点が良いのかなどを各自で考案して登録する．以降，この機能を用いて「結果」と「解釈」を登録することを「登録」と呼ぶ.

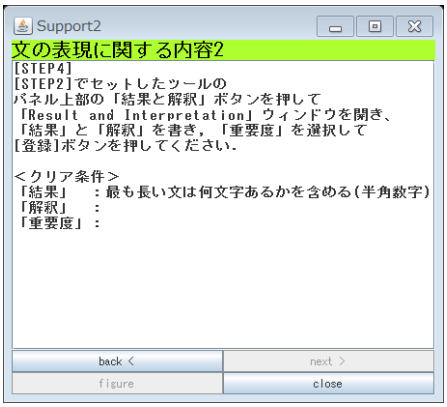


図 5：課題の詳細ウィンドウ

3.5 出力：ユーザの操作結果の正誤判定

ユーザは，[STEP4]で，結果と解釈を検討しながらミッションごとの課題をクリアすることで，チュートリアルを進行する. ユーザが[STEP4]に記述してある<クリア条件>に従って，すべての条件を満たして登録するとクリアとなる．ユーザが満たす必要のある条件を以下に示す．ただし，[STEP4]中に条件が記載されていない場合は, 登録内容に制限を設けない.

- [STEP2]で指定したツールを用いて登録を行っている
- 登録した結果の中に，<クリア条件>の解答が含まれており，指定文字数を上回っている
- 登録した解釈の中に，<クリア条件>の解答が含まれており，指定文字数を上回っている

4 チュートリアルシステムの評価実験

4.1 実験目的

提案したチュートリアルシステムが文章推敲スキル育成に有効であるか，また，使いやすいものであるかを検証することを目的とした.

4.2 実験方法

被験者は，予備研究で作成したチュートリアルシステムを使用し，TETDM そのものの使い方を学習済みである成人男女 17 人である．被験者には，実験実施前に TETDM の経験年数を問う事前アンケートに回答してもらい，経験がほぼ等しくなるように実験群の 8 人と統制群の 9 人に分けた.

まず，手順 1 では，所要時間の目安を 15 分として，実験群に本チュートリアルシステムの操作方法と進行方法を一通り学んでもらった．次に，手順 2 では, 練習フェーズとして, テキスト A について「結果」と「解釈」を登録してもらった．具体的には，実験群には，想定所要時間を 80 分として，提案システムの MISSION 1〜8 までの合計 33 課題を遂行することにより，ツールを使用した分析結果から読み取れる問題点や，ツールによる表示から読み取れる結果について一通り学習してもらいながら,「結果」と「解釈」を登録してもらった．ただし，80 分を超えても課題を達成していない場合は，時間を延長して解答することを許可した．統制群には，想定所要時間を 40〜80 分として，ヒントテキストを用いて，自由に「結果」と「解釈」を登録してもらった．想定所要時間が異なるのは，提案システムが課題の内容を閲覧してその後ツールをセットして「結果」と「解釈」を考案して登録する時間が必要であるのに対し，比較システムが用いるヒントテキストは提案システムと比較すると情報が少ないためである．なお，ヒントテキストは, チュートリアルの各課題の[STEP1]と[STEP2]の情報を箇条書きで記載した Word 文書であり, 被験者にディスプレイ上で提示した. そして，手順 3 では，実践フェーズとして，テキスト B について「結果」と「解釈」を登録してもらった．具体的には，制限時間を 60 分として，両群に「結果」と「解釈」を登録してもらった．両群とも，提案システムあるいはヒントテキストを用いることを禁止した．最後に，手順 4 では，提案システムの「使いやすさ」について，主観的評価を得るために，各群の被験者に，事後アンケートを行った.

4.3 評価方法

本実験の結果は，提案システムの妥当性と有効性の観点から評価する．具体的な評価方法を以下に示す.

● チュートリアルシステムの妥当性の評価：
チュートリアルシステムから取得するログデータから，「チュートリアルの課題クリア数と所要時間」を取得し，あらかじめ想定した時間と，実際にかかった時間を比較する. また, 事後アンケートのうち，「チュートリアルが使いやすかったかを問う項目」から，使いやすいシステムが作成できたかを確認する.

● チュートリアルシステムの有効性の量的評価：
TETDM から取得する両群の被験者のログデータから，実践フェーズの「結果および解釈の登録回数と所要時間」,「セットしたツールの種類数」,「登録に用いたツールの種類数」を取得し，両群のデータを比較する．また，事後アンケートのうち,「結果および解釈の考案が簡単だったかを問う項目」から，提案システムの結果と解釈の考案の学習の効果を評価する.

● チュートリアルシステムの有効性の質的評価：
TETDM から取得する両群の被験者のログデータから，実践フェーズの「被験者が登録した解釈」を取得する．両群の解釈の内容を比較する.

4.4 実験結果と考察

4.4.1 チュートリアルシステムの妥当性の評価

実験群の被験者が手順 2 を実施したときの平均所要時間，クリアした課題の数の平均，課題 1 つあたりにかかった平均時間を表 2 に示す．表 2 より，実験群の 8 人の被験者は, 全部で 36 個の課題から構成されるチュートリアルから, 平均 30 個の課題を 1 課題あたり平均 238.95 秒（約 4 分，標準偏差は約 1 分 30 秒）で達成した．あらかじめ設定されていた 1 課題あたり 5 分の所要時間を下回る結果となった．また，表 3 の「チュートリアルは使いやすかったですか？」のアンケート結果より，すべての被験者がチュートリアルを「使いやすかった」と回答した(2 項検定 : p=0.0078　p<.01). よって，チュートリアルの課題は利用者のレベルにあった適切なものであったと評価できる.

表 2：チュートリアルの課題クリア数と所要時間

	実験群
手順 2 を実施したときの平均所要時間（秒）	7144.88
標準偏差	2078.26
クリアした課題の平均数（個）	30.0
標準偏差	2.40
課題 1 つあたりにかかった平均時間（秒）	238.95
標準偏差	72.43

表 3：実験群の「チュートリアルは使いやすかったですか？」に対する回答

	「チュートリアルは使いやすかったですか？」に対する回答
使いやすかった	8
使いにくかった	0
両側検定(2 項検定)	p=0.0078　p<.01 **

4.4.2 チュートリアルシステムの有効性の量的評価

両群が練習フェーズと実践フェーズでそれぞれ挙げた「結果」および「解釈」の平均登録個数を表 4 に，実践フェーズにおける「結果」および「解釈」の登録所要時間を表 5 に示す．表 4，表 5 より，実践フェーズの実験群と統制群において，「結果」と「解釈」の平均登録回数と登録所要時間，平均文字数に有意な差はなかった．両群の被験者には，実験中にすばやく登録するようには指示しておらず，各々が練習フェーズの内容を踏まえてじっくりと「結果」と「解釈」を検討したため，これらの値に差が生じなかったと考えられる．一方で，両群とも被験者によって登録する「結果」や「解釈」の内容や量にばらつきが見られた．実験後にインタビューを行ったところ, しっかりとした解釈ができた被験者からは，「チュートリアルで詳しく練習したから」「自由に挙げられたから」という意見が得られ，あまり解釈ができなかった被験者からは，「何を書けば良いか迷った」という意見が得られた．実際，表 6 より，「実践フェーズで「解釈」を考案するのは簡単でしたか？」の質問で，実験群で「簡単だった」と回答した人と「難しかった」と回答した人は同数であった．本チュートリアルシステムでは，ユーザに固定観念を持たせないために，具体的な解釈例を明確に示さず，テキストマイニングツールによる分析結果から考えられる問題点について，詳しく説明するにとどまっていた．そのため，ユーザによっては，問題点に対する解釈が十分できなかったと考えられ，支援にさらなる工夫が必要であることがわかった．ただし，表 7 より「結果」について各群の回答結果を比較すると，実験群が「簡単だった」と回答した人が 3 人で，「難しかった」と回答した人が 5 人で有意差がなかった（2 項検定：p=0.7266　.10<p）のに対し，統制群では「簡単だった」と回答した人が 1 人で，「難しかった」と回答した人は 7 人で有意傾向があった（2 項検定 : p=0.0703　.05<p<.10）．すなわち，チュートリアルを使用した場合は，差がなかったが，使用しなかった場合は「難しかった」と回答する人が多い傾向があることがわかった.

表 4：各フェーズで挙げた「結果」および「解釈」の平均登録個数

		練習フェーズ	実践フェーズ
実験群	登録数	31.63	12.25
	標準偏差	12.08	2.68
統制群	登録数	10.00	11.13
	標準偏差	4.78	4.31
検定結果(t 検定)		t(15)=4.13 p=0.001 p<.01 **	t(15)=0.59 p=0.567 .10<p n.s.

表 5：実践フェーズにおける「結果」および「解釈」の登録所要時間

		実践フェーズ
実験群	登録所要時間（秒）	3018.75
	標準偏差	691.92
統制群	登録所要時間（秒）	2967.67
	標準偏差	875.81
検定結果(t 検定)		t(15)=0.12 p=0.903　.10<p n.s.

表 6：両群の「実践フェーズで「解釈」を考案するのは簡単でしたか？」に対する回答

	実践フェーズで「解釈」を考案するのは簡単でしたか？	
	実験群	統制群
簡単だった	4	2
難しかった	4	6
両側検定 (直接確率計算)	$\chi^2(1)=1.07$ p=0.6084　.10<p n.s.	

表 7：両群の「実践フェーズで「結果」を考案するのは簡単でしたか？」に対する回答

	実践フェーズで「結果」を考案するのは簡単でしたか？	
	実験群	統制群
簡単だった	3	1
難しかった	5	7
両側検定 (直接確率計算)	$\chi^2(1)=1.33$ p=0.5692　.10<p n.s.	

両群が各フェーズでセットしたツールの種類数を表 8 に，各フェーズでセットしたツールのうち，「結果」と「解釈」の登録時に用いたツールの種類数を表 9 に示す．表 8 より，両群が各フェーズそれぞれでセットしたツールの種類数には有意差がなかった．しかし，表 9 より，両群が実践フェーズで「結果」と「解釈」の登録に使用したツールの種類を比較すると，実験群が平均 14.25 種類，統制群が平均 10 種類のツールを用いており，実際に結果や解釈をできたツールの種類の数に差がある傾向があった（t 検定：p=0.071　.05<p<.10）．また，表 10 の「チュートリアルを使うことによって，ツールの使い方を理解できましたか？」のアンケート結果より，すべての被験者がチュートリアルによってツールの使い方を理解できたと回答した(2 項検定:p=0.0078　p<.01). よって, 本チュートリアルシステムは, 文章に対し，さまざまなツールを用いて，広い視点から結果や解釈を検討するうえで有効であったと考えられる.

表 8：各フェーズでセットしたツールの種類数

		練習フェーズ	実践フェーズ
実験群	セットツール総数（個）	21.50	22.00
	標準偏差	2.18	3.54
統制群	セットツール総数（個）	18.22	22.44
	標準偏差	5.41	4.00
検定結果(t 検定)		t(15)=1.50 p=0.153 .10<p n.s.	t(15)=0.23 p=0.824 .10<p n.s.

表 9：各フェーズでセットしたツールのうち，登録時に用いたツールの種類数

		練習フェーズ	実践フェーズ
実験群	登録に用いたツール総数（個）	19.13	14.25
	標準偏差	2.20	3.49
統制群	登録に用いたツール総数（個）	7.17	10.00
	標準偏差	2.54	4.78
検定結果(t 検定)		t(15)=8.70 p=0.000002 p<.01 **	t(15)=1.94 p=0.071 .05<p<.10 †

表 10：実験群の「チュートリアルを使うことによって，ツールの使い方を理解できましたか？」に対する回答

	チュートリアルを使うことによって，ツールの使い方を理解できましたか？
使いやすかった	8
使いにくかった	0
両側検定(2項検定)	p=0.0078　p<.01 **

4.4.3　チュートリアルシステムの有効性の質的評価

実践フェーズで，両群の「解釈」を比較して共起・非共起単語を調べ，次に，得られた単語が出現する解釈文が，文章中のどの視点レベルで検討されたものかを調べた．その結果，共起単語には「主題」「主語」「長文」「不適切」など，16 の単語が得られ，両群とも局所的な段落間や 1 つの段落内，1 つの文内，単語表現を視点に入れて，つまり文章中の一部分に注目した解釈を挙げたことがわかった．非共起単語には，実験群では「説明」「段落」「前半」「結論」など，52 個の単語が得られ，文章全体，局所的な段落間，1 つの段落内，全体的な文間，単語表現を視点に入れて，つまり文章をあらゆる視点から着目して解釈を挙げたことがわかった．統制群では「失礼」「客観」「分割」「曖昧」など，10 個の単語が得られ，1 つの段落内，1 つの文内，単語表現を視点に入れて，つまり文章中の細かい部分に着目して解釈を挙げたことがわかった．つまり，両群とも文章中の細かい部分に注目した解釈を挙げたが，文章全体に注目して解釈を挙げたのは実験群のみであった．これは，統制群が自由に自分なりの解釈を挙げたのに対し，実験群では，チュートリアルシステムの各カテゴリの目的を意識しながら，文章の全体構造から文中の細かい表現まで一通り推敲スキルを獲得できたためと考えられる．

5　結論

説明的文章を対象に，文章推敲スキル，すなわち文章中の問題点を見つけ出して，修正方針を検討するスキルを育成することを目的として，文章中の問題点を見つけ出し修正方針を検討する練習ができるようなチュートリアルシステムを開発した．実験の結果，システムの使用不使用にかかわらず，使用したツールの種類数は変わらなかったが，システムにより，より多くの種類のツールで解釈を登録でき，また，解釈を検討するときに文章をあらゆる視点から着目できたことがわかった．よって，本システムは，解釈の内容に広がりを持たせる点において，有効であることがわかった．しかし，被験者によっては解釈の量にばらつきがみられ，システムの効果が限定されることがわかった．今後の課題として，被験者に応じたチュートリアル課題を工夫することが挙げられる．

また，本研究で作成したチュートリアルシステムは，さまざまな職業・分野を対象にした新たな課題を簡単に実装できるように設計しているが，現在文章推敲スキルに対応した課題のみを実装しているため，今後の展開としては，さらに他の職業・分野に特化したチュートリアル課題を実装することが考えられる．

参考文献

[1] 菅沼明, 牛島和夫:テキスト処理による推敲支援情報の抽出，人工知能学会誌，Vol.23，No.1，pp.25-32，(2008)

[2] 松本章代，山田未央佳，山田翔[他]，鈴木雅人：理工系学生を対象とした技術文書作成支援システム，情報処理学会研究報告．コンピュータと教育研究会報告，Vol.2009，No.15，pp.91-96，(2009)

[3] Arijit De，Sunil Kumar Kopparapu：An Unsupervised Approach to Automated Selection of Good Essays，Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE，Trivandrum，pp. 662 – 666，(2011)

[4] Md. Monjurul Islam，A. S. M. Latiful Hoque：Automated Essay Scoring Using Generalized Latent Semantic Analysis，Computer and Information Technology (ICCIT), 2010 13th International Conference on，Dhaka，pp. 358 – 363，(2010)

[5] Toward Evaluation of Writing Style:Finding Overly Repetitive Word Use in Student Essays Proc.11th Conference of the European Chapter of the Association for Computational Linguistics，Budapest，Hungary，(2003)

[6] 石岡恒憲：記述式テストにおける自動採点システムの最新動向，日本行動計量学会(行動計量学) Vol.31，No.2，pp.67-87，(2004)

[7] 砂山渡，高間康史，西原陽子，徳永秀和，串間宗夫，阿部秀尚，梶並知記: テキストデータマイニングのための統合環境 TETDM の開発，人工知能学会論文誌，Vol.28，No.1，pp.1-12，(2013)

[8] TETDM サイト，”TETDM トップ”，http://tetdm.jp/ (2015/01/25 アクセス)

[9] 砂山渡，谷内田正彦: 展望台システムによる複数文書の要約と Web ページ集合への適用，一般社団法人情報処理学会，Vol.2001，No.86，pp.57-62，(2001)

[10] 山手砂都美，砂山渡：文章の話の組み立てと展開速度による段落間関係の評価，第 27 回人工知能学会全国大会，3K2-NFC-3-4 ，(2013)

[11] 三菱電機インフォメーションシステムズ株式会社，“DIAMining”, http://www.mdis.co.jp/products/diamining/, (2015/01/25 アクセス)

[12] NTT DATA Mathematical Systems, Inc.，“Text Mining Studio”，http://www.msi.co.jp/tmstudio/, (2015/01/25 アクセス)

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-07

価値観アイテムモデリング手法を利用した推薦理由提示手法についての考察

Consideration of Explanations for Recommender Systems with Personal-value-based Item Modeling

山口 貴之　　服部俊一　　高間康史 *
Takayuki Yamaguchi　Shunichi Hattori　Yasufumi Takama

首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

Abstract: 本稿では，価値観に基づくアイテムモデリング手法を提案し，これを用いた推薦理由の提示手法について考察する．近年，情報推薦では精度だけでなくその推薦過程をユーザに提示することで，システム全体に対する満足度の向上を意図した研究が行われている．提案手法ではユーザの価値観に着目し，相関ルールを用いて作成したアイテムモデルを推薦理由の説明に利用する．本稿ではアイテムモデルに関する予備実験結果について報告すると共に，推薦システムへの適用について考察する．

1　はじめに

近年，情報化技術の発展により，ユーザが膨大な情報の中から自分のニーズに合ったものを探すのが困難になるという問題が生じている．これに対する解決策として，ユーザの行動履歴から有用性の高い情報を推薦する情報推薦システムが注目されている．その中の一つに価値観に基づいたユーザモデルに関する研究があり，cold-start 問題 [2] や sparsity 問題 [3] に有用であることが示されている [1]．また，文献 [1] の研究では，ユーザモデリングを用いた手法を提案しているが，アイテムモデリングへの適用の可能性にも言及している．ここでのユーザモデリングとは特定ユーザのレビュー履歴からユーザのこだわりを求めるものであり，アイテムモデリングは特定アイテムに投稿されたレビューを収集し，そのアイテムがどの属性に着目して評価されているのかを求めることである．このことから一つのモデル構築に必要なレビュー数ではアイテムモデリングの方が集まりやすいという利点がある．さらにアイテムモデルを算出することで情報推薦システムの重要な要素技術の一つである，推薦理由の提示 [4] への応用が期待できる．

情報推薦システムにおいて，アイテムの推薦時にその推薦された理由を提示することは推薦アイテムに対して説得性を持たせることに繋がり，システム全体に対する信用度を向上させる期待ができることから，近年，重要視されるようになってきている [4]．

そこで本稿では価値観に基づいたアイテムモデリングを元に，推薦理由の提示を行う手法を提案する．価値観に基づくアイテムモデリングでは，一般的に低評価でも，ある属性にこだわりの強い人は好む傾向にあることなどを推定可能であるため，低評価のアイテムに対しても推薦を行うことが可能になることが期待される．本稿ではモデリング結果を活かした推薦理由文について検討することを目的として，モデリング結果に対する解釈やその表現方法等をアンケートにより収集し，分析を行う．その結果に基づきアイテムモデリングに基づく推薦理由提示方法について考察する．

2　関連研究

2.1　協調フィルタリング

協調フィルタリングとは口コミによる推薦の過程を自動化したものであり，Amazon.com[1] などのショッピングサイトで幅広く利用されている推薦手法である．協調フィルタリングの代表的なものには，蓄積されたユーザの嗜好データから類似するユーザを予測し推薦を行うメモリベースと呼ばれる手法が存在する．メモリベースはユーザベース [6] とアイテムベース [7] の 2 つに大きく分類することが可能である．ユーザベース

*連絡先：首都大学東京大学院システムデザイン研究科
〒 191-0065 東京都日野市旭が丘 6 − 6
E-mail: ytakama@tmu.ac.jp

[1] http://www.amazon.com/

では，類似した嗜好パターンを持つユーザを探し，そのユーザが好むアイテムを推薦するのに対し，アイテムベースは，ユーザの好むアイテムと類似したアイテムを推薦する．アイテムベースには，アイテムの類似度を事前に計算しておくことでユーザベースに比べ計算量を減らせる利点が存在する．

2.2 価値観に基づいたユーザモデル

価値観とはユーザがアイテムのどの要素を重視するかというものであり，消費行動に影響を与える要素であることからマーケティング等に利用されている．この価値観を用いることで，より少ない情報からユーザの嗜好や特性を推論できると考えられている．

ユーザベース協調フィルタリング [6] では他のユーザと共通に評価を行っているアイテムの評価値を元にPearson 相関を用いて類似度を計算するのに対し，価値観に基づくユーザモデルを用いて協調フィルタリングを拡張した手法が提案されている [1]．拡張手法 [1] ではユーザの価値観（Personal Values）に着目し，評価一致率と呼ばれる指標を用いて類似度を計算する．評価一致率の P_{uj} は，u のアイテム i に対する評価極性 p_{ui} と，i の属性 j に対する評価極性 p_{uij} から求められる．ユーザ u が評価した全アイテムについて，属性 j に対する評価とアイテムに対する評価の極性が一致した回数を $O(u,j)$，不一致であった回数を $Q(u,j)$ とした場合，評価一致率 P_{uj} は式 (1) のように計算される．

$$P_{uj} = \frac{O(u,j)}{O(u,j) + Q(u,j)} \tag{1}$$

これによりユーザモデルは属性数を n とした n 次元ベクトルで表され，Peaosn 相関を用いて他のユーザとの類似度が計算される．レビューサイトのデータセットを用いた評価実験により，低評価のアイテムにおける MAE が低減する結果などが得られている [9]．

2.3 推薦理由に関する研究

情報推薦において，推薦アイテムと共にその推薦理由を提示することで，以下の様な効果が期待されており，近年その必要性が高まってきている [4]．

- ユーザのシステムに対する親密性や信頼性を高める．
- ユーザの満足度を上げる．
- ユーザが目的のアイテムを早く探せるようになる．
- 説得性を持たせユーザにアイテムを購入させやすくする．

特に協調フィルタリングにおいては，推薦の仕組みや過程がユーザに伝わらない状態でアイテムが推薦されるため，推薦理由を提示することが効果的であると指摘されている [5]．協調フィルタリングに関する推薦理由提示手法として，Amazon.com では推薦時に「この商品を買った人はこんな商品も買っています」という説明文を提示し，推薦の仕組みをユーザに示している．また文献 [5] では，対象ユーザと類似度の高いユーザに付与した評価値を，ヒストグラムや表，説明文などの形式で提示することを提案し，有用性を示している．

類似ユーザが対象アイテムに下した評価を，アイテムの推薦理由として用いる手法に関しては欠点も指摘されている．文献 [8] では，推薦アイテムに対するユーザの評価を推薦時とアイテム使用後で比較した結果，評価が大きく変動することを報告している ．この問題の解決策として，ユーザの過去の評価アイテムが推薦に与えている影響力を数値化し提示する手法を提案し，類似ユーザの評価値を利用した手法よりも正確に評価を推定できることを示している．

3 価値観アイテムモデリング手法

3.1 アイテムモデリングへの適用

価値観に基づいたユーザモデリング手法 [1] では，あるユーザが評価した全アイテムのレビューよりユーザモデルを構築したのに対し，本稿で提案する価値観アイテムモデリング手法では，あるアイテムに評価されている全ユーザのレビューを元にアイテムモデルを作成する．提案手法ではどの属性に対する評価がアイテムの評価に影響を与えるのかを推論し，アイテムを推薦する．

3.2 評価一致率の拡張

2.2 節で述べたとおり，価値観ユーザモデリング手法 [1] では，アイテムに対する評価と属性に対する評価の極性が一致した回数に着目している．すなわち，好評で一致した場合と不評で一致した場合を区別せずに扱っている．提案手法では，アイテムの特性をより詳しく分析することを試みる．具体的には，アイテムに対する好評・不評，属性に対する好評・不評の組み合わせ 4 種類に分類してモデル化を行う．また，評価一致率の代わりにリフト値 [10] によって検出を行う．リフト値は，相関ルールの評価指標の一つであり，式 (2) で定義される．ここで，$P(X)$ は事象 X の生起確率である．式よりリフト値は，Y の生起確率が条件 X により何倍

に増加するかを表している．

$$lift(X \Rightarrow Y) = \frac{P(X \cap Y)}{P(X)P(Y)} \quad (2)$$

相関ルールでは，X，Y はあるアイテム集合がトランザクションに含まれる事象を表すが，提案手法では属性に対する評価，アイテムに対する評価が X, Y にそれぞれ対応する．アイテム i に対するレビューの評価極性が $p_t(\in \{\text{好評}, \text{不評}\})$ となる事象 X_i^t と，アイテム i に対するレビューにおいて属性 j の評価極性が $p_a(\in \{\text{好評}, \text{不評}\})$ となる事象 X_{ij}^a の間のリフト値 $lift_{ij}(p_a \Rightarrow p_t)$ は式 (3) で定義される．

$$lift_{ij}(p_a \Rightarrow p_t) = \frac{P(X_{ij}^a \cap X_i^t)}{P(X_{ij}^a)P(X_i^t)} \quad (3)$$

例えば $lift_{ij}$(好評 $\Rightarrow$ 好評) の値は，i が好評となる確率が，属性 j が好評の場合にどのくらい高くなるかを意味している．

式 (3) を用いて，p_a, p_t の組み合わせにより 4 通りの値が求められる．例として，ある映画に対して 4 人のユーザが評価を行った結果を表 1 に示す．また，表 1 の評価に基づいて属性ごとに 4 通りのリフト値を計算した例を表 2 に示す．ここで,「好」,「不」はそれぞれ好評，不評を表し，例えば 2 列目は属性に対する評価が好評の時に総合評価も好評となる場合のリフト値を意味する．この例では，属性「物語」が好評だとアイテムも好評になる傾向にあるといえる．また，属性「映像」に関しては不評の場合にアイテムは好評になると言える．このようにリフト値を計算することで，アイテムの評価に影響を与える「推薦時に重要度の高い属性の極性」を推論することができる．

表 1: アイテムにおける評価例

属性	ユーザ1	ユーザ2	ユーザ3	ユーザ4
総合評価	不評	不評	好評	不評
物語	不評	不評	好評	好評
映像	好評	好評	不評	不評

表 2: リフト値の計算例

属性	好⇒好	好⇒不	不⇒好	不⇒不
物語	2.00	0.67	0	1.33
映像	0	1.33	2.00	1.33

3.3 推薦理由のための説明文作成

推薦理由の説明文を，アイテムモデルを元に作成する．前述のとおり，表 2 の場合では属性「物語」を好む人はアイテムに対して高評価を付ける傾向にあることが推論できる．これより「物語を気に入った人は，この映画を好む傾向があります.」のような説明文の提示が考えられる．また属性「映像」については「映像を気に入らなかった人でも，映画自体には満足する傾向があります.」のような説明文を提示可能である．

4 予備実験

アイテムモデリングに基づく推薦理由提示方法について考察を行うため，予備実験を実施した．4.1 節に実験の概要を示し，結果を 4.2 節に示す．

4.1 実験概要

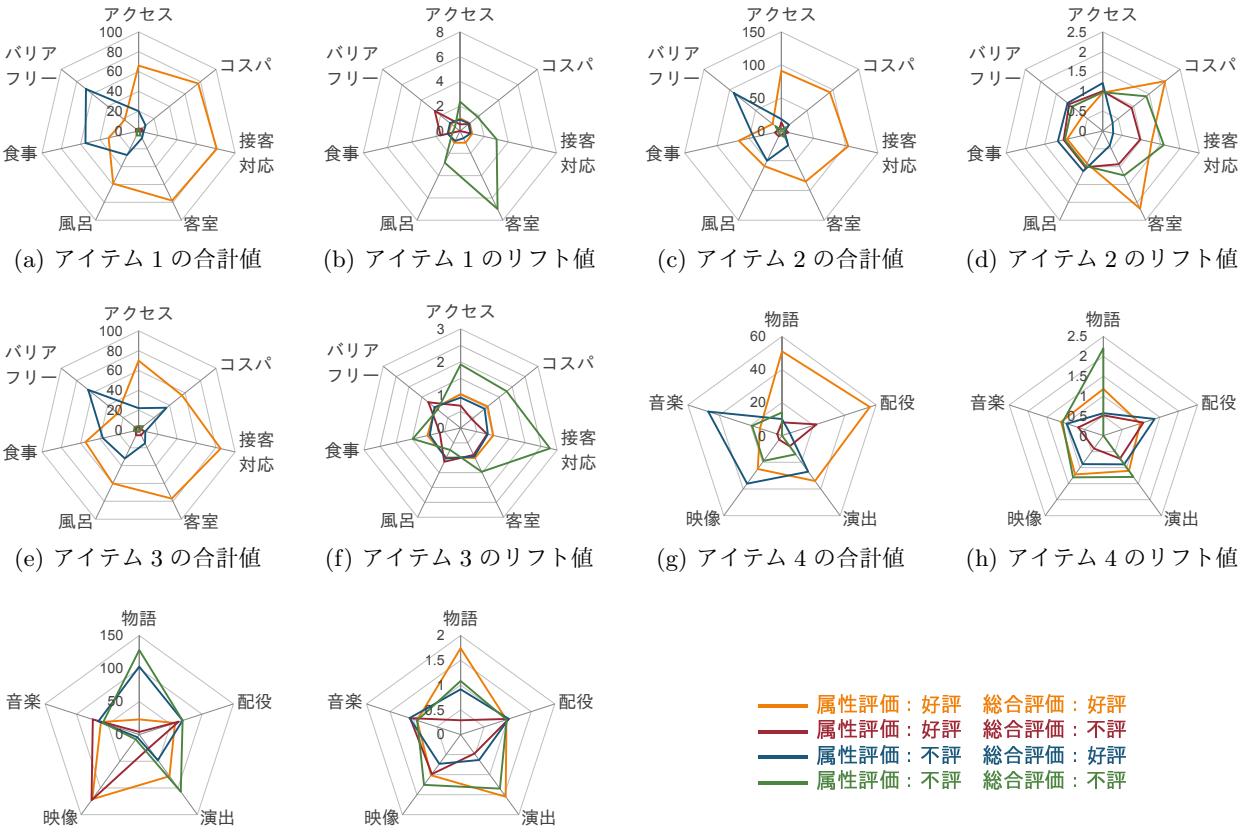
本実験では，同じ研究室に所属する 20 代の工学系大学院生 14 名に，提案するアイテムモデリングの結果を提示し，結果に対する解釈やこれを利用したアイテムの説明文等をアンケート形式で収集した．アンケートでは，アイテムの基本情報 (属性値やサムネイル画像等) とモデリング結果を提示し，実験協力者に「売り手として，アイテムの魅力を買い手に宣伝する」ことを想定し，提示された情報を元にそのアイテムの魅力や特徴を説明する文を記述してもらった．アイテムモデリングの結果として，各属性における $(X_{ij}^a \cap X_i^t)$ を満たすレビューの数と，4 種類のリフト値の値を表及びレーダグラフにしたものを提示した．

旅行サイト 4travel[2] より宿泊施設 3 件（アイテム 1, 2, 3），映画情報サイト Yahoo!映画 [3] より映画 2 件（アイテム 4, 5）の合計 5 件のアイテムを提示した．図 1 に，各アイテムについて提示したレーダーチャートを示す．

回答項目には文章の記述以外に，提示されたアイテムに対する認知度を 5 段階で評価してもらった．また説明文の記述に関しては，以下の 4 点に従ってもらった．

- 売り手として，買い手にアイテムの魅力を伝える表現を考えること．
- 売り手側としての評価や信頼を失わない表現にすること．
- 説明文に加え，提示情報内でその根拠としたものについても記述すること (理由文は買い手を想定した表現である必要はない)．
- リフト値を参考にした説明文を必ず 3 つ以上記述すること．

[2] http://4travel.jp/
[3] http://movies.yahoo.co.jp/



(a) アイテム 1 の合計値　(b) アイテム 1 のリフト値　(c) アイテム 2 の合計値　(d) アイテム 2 のリフト値

(e) アイテム 3 の合計値　(f) アイテム 3 のリフト値　(g) アイテム 4 の合計値　(h) アイテム 4 のリフト値

(i) アイテム 5 の合計値　(j) アイテム 5 のリフト値　(k) 各グラフの凡例

図 1: 実験にて提示したアイテム情報 (レーダチャート)

協力者にはアンケート前に，アイテムモデリングの作成方法やリフト値について説明を行った上で，2 週間の期限を設けアンケートに回答してもらった．

4.2 実験結果

アンケート結果より，アイテムへ対する認知度をまとめたものを図 2 に示す．図 2 より，対象のアイテムの利用経験による差は小さく，協力者はほぼアンケート内の提示内容のみからアイテムの特徴を捉え，説明文を作成していると考えられる．

アイテムモデリングに対する説明文は，協力者 14 名より 221 個の文章を得られた．得られた文章について，「リフト値が大きいこと」を根拠として作成された説明文は 190 個存在し，その中には「『属性』がおすすめです」,「『属性』が気に入られています」等の共通した説明文が見られた．そこで，回答された説明文を以下の 13 種類に分類し，それぞれの頻度を集計した結果を表 3 に示す．

1. 「属性」が好評 (おすすめ) である．
2. 「属性」を好評とした人は，アイテム自体には満足する傾向がある．
3. 「属性」に興味やこだわりを持つ人には，アイテムをおすすめできる．
4. 「属性」が不評である．
5. 「属性」を不評とした人は，アイテム自体には満足しない傾向にある．
6. 「属性」に興味やこだわりを持つ人には，アイテムをおすすめできない．
7. 「属性」がアイテムの評価に影響を与えている．
8. 「属性」を好評とした人は，アイテム自体には満足しない傾向にある．
9. 「属性」を不評とした人は，アイテム自体には満足する傾向がある．
10. 「属性」に興味やこだわりを持たない人でも，アイテムをおすすめできる．
11. 「属性」に興味やこだわりを持たない人には，アイテムをおすすめできない．
12. 「属性」がアイテムの評価に影響を与えていない．
13. その他

ここで 1 行目の文章の種類 (Index) は，分類した 13 種類の文章の番号に対応し，例えば一行一列目は「$lift_{ij}$(好評 $\Rightarrow$ 好評) の値が大きい」という理由で，「『属性』はおすすめである」と表現した説明文が 35 件あったことを示している．各リフト値の中で高頻度の 3 種類を赤字で示している．

表 3: リフト値が大きいことを根拠とした，説明文の分類結果

説明文を記述した理由	文章の種類 (Index)												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
$lift_{ij}$(好評⇒好評)が大きい	35	7	19	0	0	2	1	0	0	2	0	0	9
$lift_{ij}$(好評⇒不評)が大きい	3	1	2	0	1	4	0	1	0	11	0	7	2
$lift_{ij}$(不評⇒好評)が大きい	2	1	0	1	0	3	0	0	0	15	0	5	2
$lift_{ij}$(不評⇒不評)が大きい	6	0	1	18	1	21	2	1	0	2	0	0	8

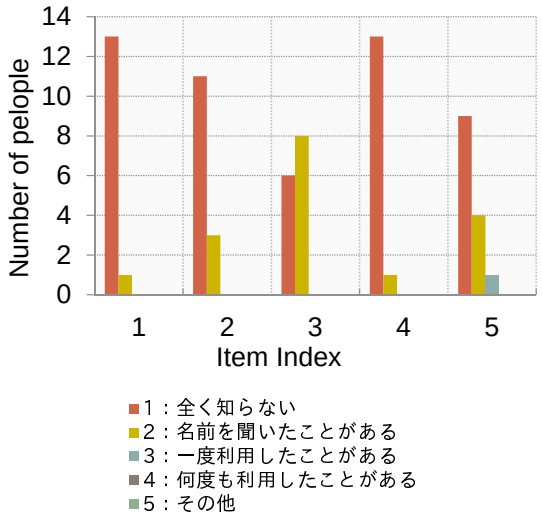


図 2: アイテムへ対する認知度のグラフ

表 3 より，$lift_{ij}$(好評 ⇒ 好評) の値が大きいという理由により，(1)「『属性』が好評である」，(3)「『属性』に対してこだわりを持つ人にはおすすめできる」という表現が多く作成されたことがわかる．また $lift_{ij}$(不評 ⇒ 不評) の値が大きいとき，(4)「『属性』が不評である」，(6)「『属性』に対してこだわりを持つ人にはおすすめできない」という解釈のされ方が多いことがわかる．

複数のリフト値を考慮した説明文も存在した．例えば，$lift_{ij}$(好評 ⇒ 好評) と $lift_{ij}$(不評 ⇒ 不評) の値が両方とも大きい時,「このアイテムは『属性』によって評価が分かれる，好き嫌いの大きいアイテムである」等の説明文が多く見られた．これらについては，その他に分類した．$lift_{ij}$(好評 ⇒ 好評) と $lift_{ij}$(不評 ⇒ 不評) において (13)「その他」に分類されている説明文が多いのはそのためである．

$lift_{ij}$(好評 ⇒ 不評) や $lift_{ij}$(不評 ⇒ 好評) の値が大きいという理由により，(10)「『属性』に対してこだわりを持っていなくても，アイテムをおすすめできる」という表現が多くされている．これらは，属性の評価がアイテムの評価に与えている影響は小さいという解釈に基づくものと言える．

回答された説明文の文中には,「リフト値」や「確率」といった用語を使用した表現は一切見られなかった．また $lift_{ij}$(好評 ⇒ 好評) に関する説明文以外では「『属性』を好評 (不評) とした人はアイテム自体を満足とする (しない) 傾向にある」といった表現はあまり見られなかった．これらの結果は協力者が売り手側としてアイテムの宣伝を考慮し，直接的な表現を控えたためと考える．

回答された説明文の中には，推薦システムで利用可能な有用な表現も存在した．例えば $lift_{ij}$(不評 ⇒ 不評) の値が大きいとき,「この『属性』にこだわりがある方は，レビューなどを確認の上ご利用下さい」,「この『属性』の評価はアイテムの評価に大きな影響を与えています」等の表現が見られた．また，$lift_{ij}$(好評 ⇒ 不評) のリフト値が大きいとき,「この『属性』を重視する方は，他の要素も気にして判断していただくと良いです」という表現が見られた．このような表現は「不評，おすすめできない」といった消極的な表現を使わないため，アイテムに対するイメージを下げずにユーザに注意を促すことが期待できる．

また各リフト値がとる値にあまり差が見られない場合,「バランスのとれたアイテムです」,「各要素よりもアイテム全体を見て評価されている」等の表現がされていた．一方，リフト値に大きな特徴が見られなくても $(X_{ij}^a \cap X_i^t)$ を満たすレビューの数から判断している回答も存在した．

4.3 推薦システムでの利用に関する考察

実験結果を踏まえ，4 種類のリフト値それぞれを根拠とした説明文として，表 4 に挙げるものが利用可能と考える．推薦システムでの利用方法としては，各リフト値が設定した閾値を超えた場合に該当する説明文をテンプレートとして生成することが考えられる．

またリフト値に大きな特徴が見られない場合は，$(X_{ij}^a \cap X_i^t)$ を満たすレビューの数による判断を行うか,「バランスのとれた (平均的に見られている) アイテムです」等の説明を行うことが考えられる．

表 4: 閾値を超えたリフト値に応じて提示する説明文の例

リフト値の種類	提示する文章の例
$lift_{ij}$ (好評⇒好評)	• この「属性」に興味(こだわり)を持つ方にはおすすめできます. • この「属性」の評価は，アイテム全体の評価に大きな影響を与えています. • この「属性」の評判は良いです.
$lift_{ij}$ (好評⇒不評)	• この「属性」にこだわりを持たない方は，他の要素で判断するほうが良いかもしれません • この「属性」の評価が，アイテムの評判に繋がることはあまりありません. • この「属性」に興味(こだわり)を持たない方にもおすすめできます.
$lift_{ij}$ (不評⇒好評)	• この「属性」の評価が，アイテムの評判に繋がることはあまりありません. • この「属性」に興味(こだわり)を持たない方にもおすすめできます.
$lift_{ij}$ (不評⇒不評)	• この「属性」に興味(こだわり)を持つ方は，事前にそれに関するレビューに目を通すことをおすすめします. • この「属性」の評価は，アイテム全体の評価に大きな影響を与えています

5 おわりに

本稿では推薦システムにおいて，価値観アイテムモデリング手法を提案した．提案手法では，価値観に基づいたユーザモデルに関する先行研究を拡張し，アイテムモデルに適用した．評価一致率に代わる指標として，リフト値を用いてモデリングを行った．

モデリング結果に対する協力者の意見を予備実験にて収集し分析することで，アイテムモデリングに基づく推薦理由の説明文生成の可能性について考察し，推薦システムでの利用について検討した．

今後は，提案手法を用いた推薦システムを構築する予定である．そのために，各リフト値を根拠とした説明文を生成する際の閾値の設定，およびテンプレート文章の増加について検討を行う．

参考文献

[1] S. Hattori and Y. Takama, “Consideration about Applicability of Recommender System Employing Personal-Value-Based User Model,” TAAI2013, pp.282-287, 2013.

[2] A. I. Schein et al, “Methods and metrics for cold-start recommendations,” Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 253-260, 2002.

[3] S. Lee, J. Yang and S. Park, “Discovery of hidden similarity on collaborative filtering to overcome sparsity problem,” Discovery Science. Springer Berlin Heidelberg, pp. 396-402, 2004.

[4] N. Tintarev and J. Masthoff, “ A survey of explanations in recommender systems,” IEEE 23rd International Conference on Data Engineering Workshop, pp. 801-810, 2007.

[5] J. L. Herlocker, J. A. Konstan and J. Riedl, “ Explaining collaborative filtering recommendations,” 2000 ACM conference on Computer supported cooperative work, pp. 241-250, 2000.

[6] P. Resnick et al, “ GroupLens: an open architecture for collaborative filtering of netnews,” 1994 ACM conference on Computer supported cooperative work, pp. 175-186, 1994.

[7] G. Linden, B. Smith and J. York, “ Amazon.com recommendations: Item-to-item collaborative filtering,” IEEE Internet Computing, Vol. 7, No. 1, pp. 76-80, 2003.

[8] M. Bilgic and R. J. Mooney, “ Explaining recommendations: Satisfaction vs. promotion,” Beyond Personalization 2005, pp. 13-18, 2005.

[9] 三澤 遼理, 服部 俊一, 高間 康史, “ 価値観に基づくユーザモデルによる協調フィルタリングの拡張手法の提案,” 第 27 回人工知能学会全国大会 (JSAI2014), 1H4-NFC-01a-5, 2014.

[10] R. J. Bayardo Jr and R. Agrawal “ Mining the most interesting rules,” 5th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 145-154, 1999.

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-08

直感的な意味付けによる百分率と速さの問題のための学習システムの開発

Learning System for Percentage and Speed Questions by Intuitive Meaning Attachment

長田 佳倫 [*1]　　砂山 渡 [*1]
Yoshinori Nagata　　Wataru Sunayama

[1] 広島市立大学 情報科学部
Faculty of Information Sciences, Hiroshima City University*

Abstract: Elementary, junior high or high school students are required to understand way of thinking against questions that involve calculation in the process of solution. Students who know formula only have to remember all formula to solve various questions and cannot have ability to apply one to similar questions. In this study, a learning system that assists establishing way of thinking against percentage and speed questions by intuitive understanding iterations is proposed. According to experimental results, the proposed system that emphasis on way of thinking was effective to learning and its radication rather than the comparative system that emphasis on solution by formula.

1　はじめに

小学校，中学校，高校の学生や生徒が，計算によって解答を導き出す科目や分野の問題解決過程には，問題の解決方法を説明する「考え方」，考え方に基づいて問題を解けるようにする「問題の定式化」，そして定式化により与えられる式に基づいて答えを求める「計算」とがある．このような問題解決においては,「考え方」を理解して「問題の定式化」を行うプロセス（本研究では問題の本質と呼ぶ）が重要となるが，実際には「問題が定式化」された後のいわゆる「公式」を活用して「計算」を行えば試験で得点が取れるため，学生のレベルに応じてはこの本質部分が省略されることも少なくない．

しかし本質部分を除いた学習では，すべての公式やパターンを暗記する必要があることに加え，類似する他の問題への応用を考えることができなくなる欠点が生じる．本質部分の学習では「問題を解く際の考え方を理解すること」が必要となるが，文部科学省の「教科書の改善・充実に関する研究報告書」[1] によると，多くの小学生は現在の算数の教科書に対し，簡単な問題から急に難しくなるのでわからなくなること，また絵や写真は少なくてもいいがもっと詳しい図や解説がほしいという学生の感想が示されており，説明のわかりやすさと，段階的かつ具体的な手順の説明の必要性がうかがえる．

そこで本研究では，百分率と速さの分野を対象として，公式の丸暗記とその適用を目指す学習ではなく，問題の考え方から定式化へと至るプロセスに焦点を当てた学習を支援する．すなわち，考え方の理解と記憶を促し，段階的かつ具体的な手順の説明を備え，学習した考え方を定着させられる学習システムを構築する．本稿で提案するシステムは，学習対象分野の問題の考え方を理解していない，あるいは十分に考え方が定着していない学習者を主な対象とする．

*(連絡先) 砂山渡, 731-3194, 広島市安佐南区大塚東 3-4-1, 広島市立大学大学院情報科学研究科, sunayama@hiroshima-cu.ac.jp

2　関連研究

2.1　問題解決過程を支援する研究

一般的な数学の問題解決過程においては，一つの知識によって即座に解答に至ることは少なく，いくつかの論理展開をつないで理解を進めて行く必要がある．そのため，生徒の内発的な問いをつなげて問題解決を進展させる研究 [2] や，生徒のペアがお互いに対話しながら問題解決を図ることの意義に関する研究 [3] がある．しかし本研究では，複数の過程を要する問題を解く前段階として，一つの知識により導かれる式を用いて解答することができなければ，より複雑な問題に対応することはできないと考えた．

また，問題解決過程においては図を描くことが有効と示されており，特に図を描くことが困難な生徒に，図を描くきっかけを与えることが有効と示されている [4]．そこで本研究では,「比較量，基準量，割合」の 3 つ組からなる分野を対象として，その 3 つの数量の関係のみの把握から解決できる問題を対象とした上で，直感的に理解しやすいグラフを用いて表すことで，既知の量と求める量の関係を明確にして問題解決を促す．

問題解決を支援する図やその効果に関する研究は古くから行われている．数学教育においては，問題の状況を説明して問題解決を促す図に「情景図」があり，特に線分を利用して表したものを線分図と呼ぶ．これまでの研究においては，このような図の役割について調査した研究 [5, 6] や,「速さ」や「割合」などの特定の単元を対象として，その分野の問題を幅広く解決することを支援するものが多かった [7, 8]．しかし，理解力が十分でない生徒に対して，その場しのぎとならない根本的な問題解決能力を与えるためには，複数の図を用いたり，多様な問題を対象とする前に，単純かつ複数の分野に適用可能な考え方を教える必要がある．

そこで本研究では,「比較量，基準量，割合」の 3 つ組からなる分野に対して，線分図に相当するグラフを用いて，問題解決に必要な公式の考え方に着目し，特定の分野によらない問題解決支援を目指す．特に，提示するグラフを問題解決に向けた説明のためだけに用いるのではなく，学習者が問題を解く場面において，頭の中でグラフを再現して考えられるようにすることを目指す．

2.2 繰り返し練習により学習内容の定着を支援する研究

近年，PC やタブレット端末を利用した学習支援が広く行われるようになってきている．特に知識の定着を目的とした場合，計算ドリルや漢字ドリルなど類似問題を反復して解く学習が一般に行われており，コンピュータを活用したドリル型の学習支援について，その効果が確認されている [9]．コンピュータを活用した学習の利点にはいくつか考えられるが，単純な知識の定着という観点においては，短時間に繰り返し可能な練習問題を自動的に生成することで，集中的に反復が可能になる点が挙げられる．しかし，考え方の学習を目指したシステムでは，時間をかけた学習により知識の定着を目指すことが多く，学習意欲の維持が困難になり中途で投げ出してしまうなど，別の問題が生じる可能性があった．

そこで本研究では，公式への当てはめなど，機械的な作業の反復により「知識」を覚えさせる学習システムではなく，問題を解く過程を重視して，頭の中で考えることを短時間に繰り返す練習により，問題を解く際の「考え方」を身につけられる学習システムの構築を目指す．特に，紙や鉛筆を用いずに頭の中で考えさせること，また短時間で多くの問題をこなせるようにすることを重視して，暗算が可能な計算で解ける問題を用意し，考え方の理解と定着に特化したシステムを構築する．

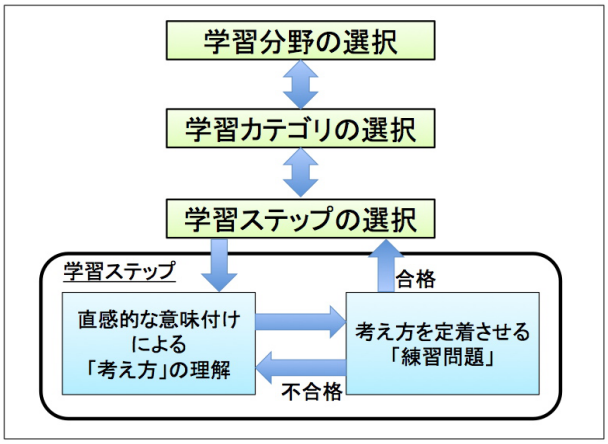


図 1: 直感的な意味付けによる学習システムを用いた学習の流れ

3 直感的な意味付けによる学習システム

本章では，直感的な意味付けにより，問題の考え方の理解を促し，理解した考え方を繰り返しの練習により定着させる学習システムについて述べる．

3.1 直感的な意味付けによる学習システムの構成

図 1 に，学習システムを用いた学習の流れを示す．提案システムを用いる学習者は，学習分野，学習カテゴリ，学習ステップを選択した後に，指定したステップでの理解目標となる問題について，考え方を理解した後，考え方を定着させる練習問題を解く．

3.2 学習分野

本システムでは,「比較量，基準量，割合」の 3 つ組からなる学習分野を対象とする．本研究では，特に「百分率」と「速さ」の分野を対象とした．百分率の分野は「比べる量，もとにする量，百分率」の 3 つ組，速さの分野は「道のり，速さ，時間」の 3 つ組から構成される．

表 1: 学習カテゴリ「比べる量」「もとにする量」「百分率」の学習ステップ，考え方で用意したページ数ならびに練習問題の種類数（ただし「もとにする量」ではステップ 4) を省略した 4 ステップ構成）

学習ステップと出題に用いる百分率	考え方	問題種類数
1) 100%	1 ページ	27
2) 50, 25, 10, 1%	4 ページ	108
3) 25, 10, 1%の複数倍	3 ページ	432
4) 100% + ステップ 2), 3)	3 ページ	540
5) 100% − ステップ 2), 3)	3 ページ	540

表 2: 学習カテゴリ「道のり」「速さ」「時間」の学習ステップ，考え方で用意したページ数ならびに練習問題の種類数

学習ステップと出題に用いる時間	考え方	問題種類数
1) 1 時間	1 ページ	27
2) 30, 20, 15, 10, 1 分	5 ページ	135
3) 1 時間, 20, 15, 10, 1 分の複数倍	5 ページ	513

3.3 学習カテゴリ

「比較量，基準量，割合」の 3 つ組からなる学習分野において，その 1 つ 1 つの構成要素を学習カテゴリと呼ぶ．本研究で対象とする学習分野「百分率」では，「比べる量」「もとにする量」「百分率」，学習分野「速さ」では,「道のり」「速さ」「時間」が学習カテゴリとなる．

3.4 学習ステップ

学習カテゴリで学習する内容に対して，段階的な学習を可能にする細分化を行ったものを学習ステップと呼ぶ．学習分野「百分率」ならびに学習分野「速さ」の学習カテゴリで用意した学習ステップを，それぞれ表 1 と表 2 の左部分に示す．両カテゴリのステップ 1) からステップ 3) の 3 つのカテゴリにおいては，扱う割合を，基準量と同一の値，基準量の $1/n(n$ は整数) で表される単純な値，基準量の $1/n$ で表される単純な値の複数倍，を扱った問題として構成した．これらに加えて，学習分野「百分率」では分野特有の出題傾向に応じて，基準量からの増減に関するステップを設けた．

本システムで対象とする問題は，暗算が可能な問題に限ってステップを構成している．これは本研究が，紙を使って計算することなく，頭の中だけで考えて解答を導き出す，考え方の理解に特化した学習を目指していることによる．

各学習ステップは，直感的な意味付けにより考え方の理解を行うための「考え方」部分と,「考え方」を定着させるための「練習問題」から構成される．表 1 と表 2 の右部分に，各ステップで用意した「考え方」を説明するページの数と，練習問題で出題される問題種類数を示す．学習者は，この「考え方」のページを閲覧することで，各ステップの問題を解く方法について学ぶ．その上で，学んだ考え方を実際の問題を解きながら定着させるための「練習問題」に進む．

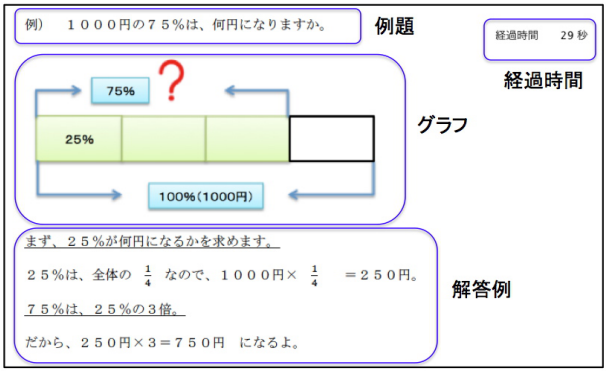


図 2: 「考え方」のシステム画面（「比べる量」 ステップ 3)）の例

3.4.1 直感的な意味付けによる考え方の理解

「考え方」を理解するページの構成を図 2 に示す．考え方を理解するページは，例題，グラフ，解答例，ページの閲覧開始からの経過時間の 4 つで構成する．各学習ステップにおいては，例題の数値を変更した別の例題の解き方を説明したページを複数用意する．

例題は，各ステップで出題対象となる数値を用いたものとして，最も簡潔な表現で必要な値を求めさせる問題とする．これは，問題理解過程ではなく問題解決過程を支援するために，もっともシンプルに理解できる問題表現を意図したことによる．

グラフは，例題を解く際に，頭の中に描かれるべきイメージを表す図として用意する．すなわち，問題の解き方を解説するためだけの図ではなく，後に同様の問題が出題された場合に，学習者が頭の中で再現しやすい図として，学習カテゴリ内で一貫性のある図を用いる．図 3 に，学習カテゴリ「比べる量」の各学習ステップのグラフの例を示す．

解答例は，問題を解く方法の説明となることに加えて，頭の中でグラフをイメージしたときに，どのように考えを進めていけばよいか，頭の中の思考過程を表す形で用意する．また，前のステップの考え方を理解している前提で，簡潔な表現となるようにする．

経過時間は，閲覧中の考え方のページを見てからの時間を表示する．これは，考え方を読み飛ばさずに理解してもらうために，次のページに進むための必要閲覧時間 (20 秒) を設けた上で，それを確認できるように用意する．

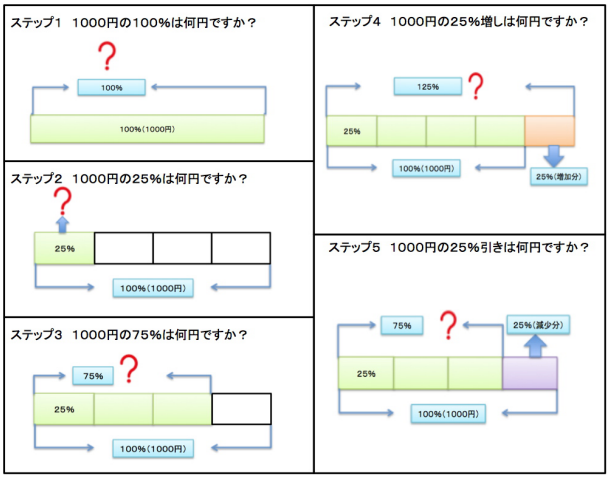


図 3: 学習カテゴリ「比べる量」の各学習ステップのグラフの例

表 3: 学習カテゴリ「比べる量」における学習ステップの学習目標

学習ステップの学習目標 (25%の値を利用する例による説明)
1) 100%は全体を表すことの理解. 100%と聞くと全体を表す図をイメージできる.
2) 25%は全体の 1/4 を表すことの理解. 25%と聞くと全体の 1/4 を表す図をイメージできる.
3) 75%は 25%の 3 倍となることの理解. 75%と聞くと全体の 1/4 が 3 つを表す図をイメージできる.
4) 25%増は全体に 1/4 を加えることの理解. 25%増と聞くと全体に 1/4 を加えた図をイメージできる.
5) 25%減は全体から 1/4 を減じることの理解. 25%減と聞くと全体に 1/4 足りない図をイメージできる.

これらの構成要素と，学習カテゴリを細分化した段階的な学習ステップは，頭の中でグラフのイメージを作りながら，徐々に問題を解くための考えを進めることにより,「考え方」を身につけられる設計を意図している．実際に用意している，学習カテゴリ「比べる量」における各学習ステップの学習目標を表 3 に示す.

3.4.2　考え方を定着させる練習問題

理解した「考え方」を定着させることを目的として，考えることを繰り返し練習するための問題を用意する．練習問題は，各学習ステップごとに 10 問 1 セットとして用意し，9 問以上の正解で合格とする．9 問以上の正解で合格とした理由は，学習内容を習得していれば全問正解できるはずの問題となっていることと，勘違いや操作ミスが起こる可能性も考慮して，1 問までの間違いは認めたことによる.

図 4 に練習問題の出力構成を示す．練習問題の出力は，問題，グラフ，5 択の選択肢，その問題の解答開始からの経過時間の 4 つで構成する.

練習問題で使用する数値の組合せとして,「百分率」

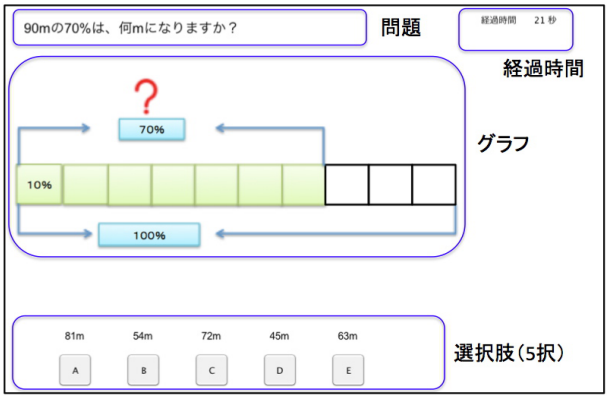


図 4:「練習問題」のシステム画面の例

表 4: 学習分野「百分率」の学習ステップで用意した練習問題の百分率のパターン（ただし「もとにする量」「百分率」ではステップ 4) を省略した 4 ステップ構成）

ステップと百分率のパターン	パターン数
1) 100%	1
2) 50, 25, 10, 1%	4
3) ステップ 2) の 2 倍から 9 倍 (100%未満)	16
4) 100% + ステップ 2), 3)	20
5) 100% − ステップ 2), 3)	20

と「速さ」の両分野において,「基準量」(「もとにする量」と「道のり」) には 10 から 10 ずつ 90 まで，100 から 100 ずつ 900 まで，1000 から 1000 ずつ 9000 まで，の 27 パターンを用意し,「割合」(「百分率」と「時間」) には，表 4 と表 5 に示すパターンだけ用意した．問題は，これらの可能なパターンの組合せの中から，毎回ランダムに出題する．またこれらの数値パターンは，解答時に暗算が容易になる値とする．具体的には，一桁の数のかけ算または割り算により，答えが 6 桁から小数点第 1 位までの値になる問題を取り扱う.

グラフは,「考え方」を理解する際に用いたものと同様のグラフを自動的に描画して表示する．ただし基準量の数値は表示せず，学習者に自分で当てはめてもらうことを意図した．学習者は解答時にもこのグラフを頼りにすることで，グラフを用いることが解答のための必要条件となることを理解させ，グラフの形を繰り返し確認することで独力でイメージしやすくなるように練習を行う.

4　直感的な意味付けによる学習システムの効果の検証実験

本章では，構築した直感的な意味付けによる学習システムを用いる学習者が,「百分率」と「速さ」の分野において,「考え方」を理解して，解答を導く能力を身につけられるかを検証した実験について述べる.

表 5: 学習分野「速さ」の学習ステップで用意した練習問題の時間のパターン

ステップと時間のパターン	パターン数
1) 1 時間	1
2) 30, 20, 15, 10, 1 分	5
3) ステップ 1), 2) の 2 倍から 9 倍	19

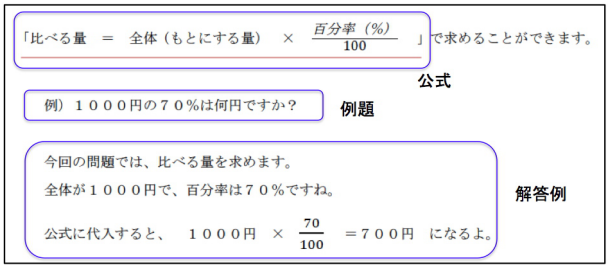


図 5: 比較システムの「考え方」ページの例

4.1 実験手順

実験は，ある学習塾の小中学生 16 名を被験者として，提案システムを用いて学習を行う提案群と，比較システムを用いて学習を行う比較群に 8 名ずつに分けて行った．被験者にはまず，提案群と比較群へのグループ分けと，後に学習成果を確認するために，事前テストを行ってもらい，テストの点数が同程度となるようにグループ分けを行った．その後，被験者は指定されたシステムを用いて，1 日 1 時間ずつ合計 4 日間の学習をしてもらった．最後に事後テストを行い，事前テストとの点数の比較，ならびに学習時のログと比較することで，システムの効果を検証する．

比較システムには，提案システムにおける「考え方」のページにおいて，グラフの代わりに教科書通りの公式を示し，解答例として公式を適用した解法を提示した．5 に比較システムの「考え方」ページの例を示す．また「練習問題」においては，グラフや公式は表示されない．これは比較システムにおいては，公式は 1 つの学習カテゴリ内で 1 つしかなく，考え方で繰り返し参照していることをもとに，練習問題を解く際に，公式を頭の中に思い出させることで，公式の利用法の定着を意図したことによる．

事前テストと事後テストは，各学習カテゴリの各学習ステップで扱う「割合」の種類について 1 問ずつの合計 72 問とし，1 問 1 点とした．

4.2 実験結果と考察

4.2.1 事前テストと事後テストに基づく学習効果

表 6 に事前テストと事後テストの正答率 (%) の被験者平均を示す．事後テストの結果から，提案群の被験者のみ事前テストの結果に比べて事後テストの点数が有意に増加していた（$t(7)= 3.50, p < 0.01$）．また提案群の被験者が，比較群の被験者に比べて点数が有意に高くなった（$t(14)=2.35, p < 0.05$）．このことから，公式を利用した解法により解答を導く学習に比べ，グラフを用いて考え方を理解させる学習の方が効果が高かったことがわかる．

表 6: 事前テストと事後テストの正答率 (%) の被験者平均（括弧内は標準偏差）

	提案群	比較群
事前テスト	47.7(17.8)	48.1(18.2)
事後テスト	66.7(16.4)	52.4(18.2)
増加正答率	18.9(14.3)	4.3(8.1)

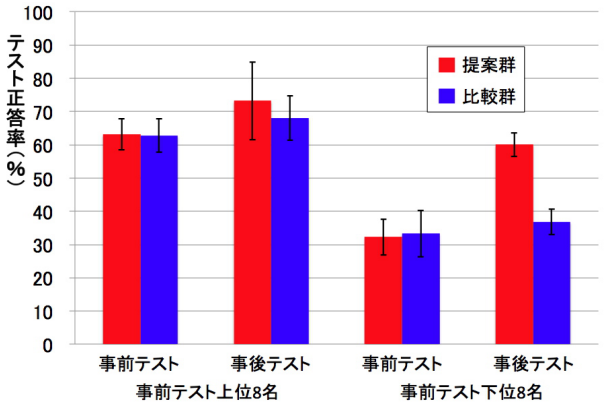


図 6: 事前テストの上位 8 名と下位 8 名の事前テストと事後テストの正答率 (%) の被験者平均と標準誤差

図 6 に，事前テストの上位 8 名と下位 8 名の事前テストと事後テストの正答率 (%) の被験者平均を示す．提案群の被験者のうち，特に事前テストの点数が低かった 4 名の正答率が大きく上昇していた（$t(3)=5.38, p < 0.05$）．このことから，学習分野における「考え方」の理解が進んでいなかった学生に対して，本システムの効果が大きかったことがわかる．

4.2.2 学習の達成度と学習内容の定着

表 7 に，被験者が学習により練習問題に合格したステップの数の被験者平均を示す．事前テスト上位の被験者 8 名について，提案群の被験者は全員が 23 ステップ全てで合格まで達したのに対して，比較群の被験者は，一人平均 3 つのステップを合格できなかった．このことから，「考え方」を十分には理解していなかったが，それなりに問題を解ける学生に対して，提案システムを用いた学習によって，曖昧だった「考え方」を明確にする学習効果があったことがわかる．

事前テスト下位の被験者 8 名について，学習ステップの数には，両群の間に有意差は見られなかった（$t(14)=1.68,$

表 7: 練習問題に合格した学習ステップの数 (全 23 ステップ) の被験者平均（括弧内は標準偏差）

	提案群	比較群
事前テスト上位	23.0 (0.0)	20.0 (2.2)
事前テスト下位	18.3 (0.4)	17.0 (1.2)
合計	20.6 (2.4)	18.5 (2.3)

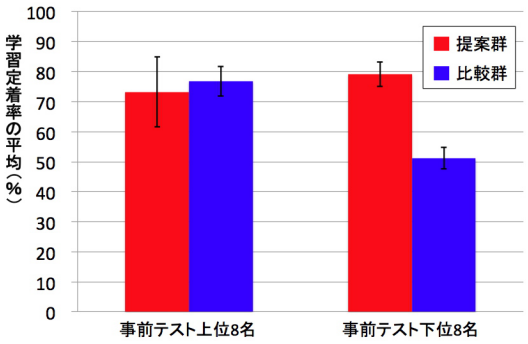


図 7: 学習定着率（事後テストにおいて練習問題に合格したステップの問題の正答率）の被験者平均と標準誤差

$p > 0.1$）．このことから，用いたシステムによらず，繰り返しによる学習によって，一時的に正しい解答を導ける状態に達したことがわかる．

学習定着率を，事後テストの問題のうち，練習問題に合格したステップの問題の正答率，と定義する．この学習定着率 (%) の被験者平均を図 7 に示す．事前テスト下位の被験者のうち，比較システムを用いた被験者の学習定着率が低い結果となった（$t(6)=4.80$, $p < 0.01$）．このことから，学習分野の理解が進んでいない学習者に対しては，グラフを用いた直感的にわかりやすい考え方を用いることが特に有効であり，提案群の事前テスト下位の被験者においては，学習時に理解した考え方を頭に残しやすくする効果があったと考えられる．

5 まとめ

本稿では，直感的にわかりやすく，記憶を促すグラフを用いた説明により考え方の理解を促し，学んだ考え方を定着させる繰り返しによる練習問題により，「比べる量，基準量，割合」の 3 つ組からなる学習分野の習得を支援する学習システムについて述べた．実験の結果，考え方を重視した提案システムの方が，公式による解き方を重視した比較システムに比べ，学習の効果，ならびに学習内容の定着が図れることを検証した．また，対象とする学習分野の考え方を十分に理解していない学生に，提案システムの効果が高いことを確認した．

今後は，「比べる量，基準量，割合」の 3 つ組からなる他の学習分野に対して本システムを適用することや，共通の考え方で問題を解くことが可能な複数の分野の学習において，複数の分野に共通する考え方を提示することで，ある分野の学習成果を，他の分野の学習に生かす方法について検討していきたいと考えている．

参考文献

[1] 教科書の改善と充実に関する研究報告書（算数），文部科学省 (2008)

[2] 清水祐子：問題解決過程における問い方の発達を促す支援：Scaffolding の考え方を取り入れて，数学教育論文発表会論文集，Vol.41, pp.117–122 (2008)

[3] 清水美憲：数学的問題解決の過程における対話の意義 (II)，数学教育論文発表会論文集，Vol.23, pp.48–54 (1990)

[4] 松田由香里：児童の問題解決過程における図の役割に関する研究 ： 小学校 3 年生に対する授業分析を通して，数学教育論文発表会論文集，Vol.35, pp.151–156 (2002)

[5] 菊地光司：算数の問題解決における図的表現の働きに関する研究，日本数学教育学会誌，Vol.78, No.12, pp.334–339 (1996)

[6] 山口耕：算数学習における絵図的表現の研究：表現レベルが文章題解決に及ぼす影響について，数学教育論文発表会論文集，Vol.38, pp.151-156 (2005)

[7] 水井裕二：「速さ」の非定型文章題の問題解決に有効な図的表現，日本教科教育学会誌，Vol.24, No.1, pp.1–10 (2001)

[8] 新堀栄：数学的道具としての概念形成を目指した教材構成に関する研究：割合指導の問題提示場面に用いられる図の視覚的影響，数学教育論文発表会論文集，Vol.32, pp.317–322 (1999)

[9] 王戈，熊谷倫子，沢井 佳子，坂元章：学習支援システムの使用が小学生の学力に及ぼす効果：パネル研究による評価，日本教育工学会論文誌，Vol.29, Supplement, pp.45–48 (2006)

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-09

日本語の係助詞「は」および主格の格助詞「が」の働きのグラフによる視覚化

A visualizing the function of Japanese particle topic marker ‘wa’ and subject marker ‘ga’ by using graphs

岡安　一壽
Kazuhisa OKAYASU

神奈川県立湘南高等学校
Kanagawa Prefectural Shonan High School

Abstract: **Ⅰ.** It is possible to visualize the function of Japanese particle topic marker ‘wa’ and subject marker ‘ga’ by using graphs. These graphs help us understand the role of ‘wa’ and ‘ga’. And they show the difference between Exhaustive listing and Neutral description.
Ⅱ. People or computers can gather information as 2-dimensional arrays by hearing many sentences of the type ‘A wa B ga C’. By using the arrays, they can say the sentences of the type ‘A wa B ga C’, and can give the information to the others.

1. 普遍的な様子を表すための「は」と非普遍的な様子を表すための「が」の働き

説明の都合上「ａはＡ」「ａがＡ」という形の文のａを主部，Ａを述部と呼ぶことにする。

1. 1　主部が１つしかないものの場合（主部が固有名詞等の場合）

●（いつでも）ａ――○――Ａ
　または
（どこでも）ａ――○――Ａ
という関係を伝えるためには「ａはＡ」と言う。
例.（毎日）　太陽――○――東から昇る
という関係を伝えたければ,「太陽」と「東から昇る」を「は」でつないで,
　　太陽は東から昇る　　(1-1)
と言えば良い。
図１－１は，文(1-1)で伝えたい内容をグラフで視覚化したものである。縦軸は伝えたいか否かを表すものとする。

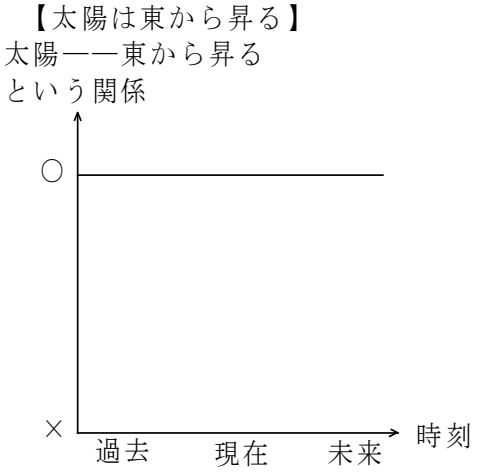


図１－１

●いつでも成り立つわけではないのだが,
　または，どこでも成り立つわけではないのだが
（その時の）　ａ――○――Ａ
または,
（そこでは）　ａ――○――Ａ
という関係を伝えるためには,「ａがＡ」と言う。
例.（今まで）太陽――×――東から昇る
（今）　太陽――○――東から昇る
という関係を伝えるためには,「太陽」と「東から昇る」を「が」でつないで,
（おい,）太陽が東から昇る（ぞ！）　(1-2)

と言えば良い。図１－２は，文(1-2)で伝えたい内容をグラフで視覚化したものである。

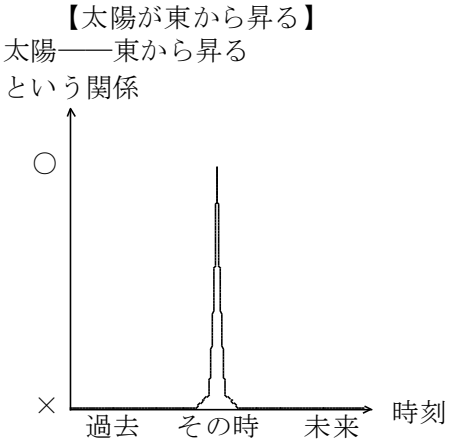


図１－２

1．2　主部がいくつかあるものの場合（主部が普通名詞等の場合）

●（どのａでも）ａ――○――Ａ
という関係を伝えるためには「ａはＡ」と言う。
例．（どのクジラでも）クジラ――○――大きい
という関係を伝えたければ，「クジラ」と「大きい」を「は」でつないで，

クジラは大きい　(1-3)

と言えば良い。

文(1-3)は，１頭のクジラについて時間的な恒常性を述べているのではなく，総称としてのクジラの普遍性を述べている。しかし，話し手にとっては，過去にクジラを見た時も，これからクジラを見る時も成り立つことを示しているのと同じである。

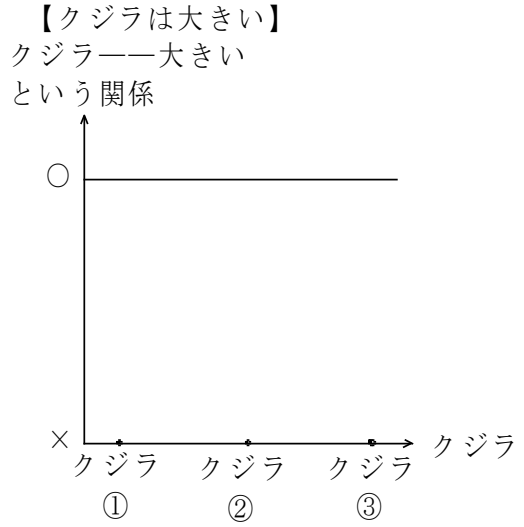


図１－３

図１－３は文(1-3)で伝えたい内容をグラフで視覚化したものである。

●どのａでも成り立つとは限らないのだが
（その）　ａ――○――Ａ
という関係を伝えるためには，「ａがＡ」と言う。
例．初めての海外旅行でアメリカへ行った人が，注文したハンバーガーの大きさに驚いて
（今まで見てきた）ハンバーガー――×――大きい
（目の前の）　ハンバーガー――○――大きい
という関係を伝えたければ，「ハンバーガー」と「大きい」を「が」でつないで，

ハンバーガーが大きい　(1-4)

と言えば良い。

文(1-4)は目の前のハンバーガーの特殊性，総称としてのハンバーガーの非普遍性を表しており，中立叙述と呼ばれる。

図１－４は文(1-4)で伝えたい内容をグラフで視覚化したものである。

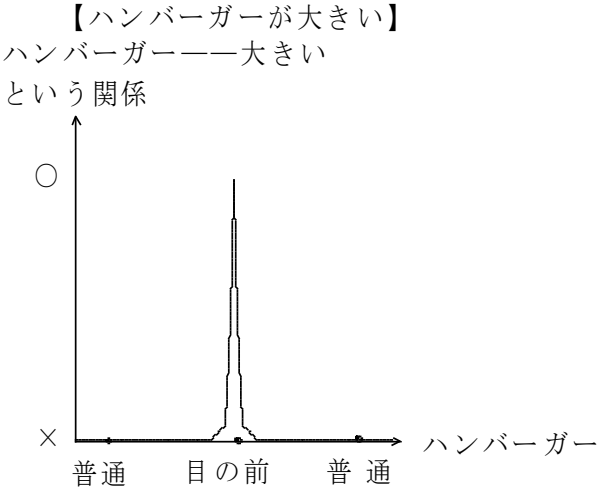


図１－４

2．総記の「が」と中立叙述の「が」

初めてアメリカへ行った人が，青い郵便ポストに驚いて，

郵便ポストが青い　(2-1)

と言ったとする。この場合の「が」は中立叙述の「が」であり，目の前の「郵便ポスト」の特殊性を表している。この意味をグラフで視覚化したものが
図２－１である。

それに対して，「アメリカについて答えて下さい。｛タクシー，郵便ポスト，ナショナルチームのユニフォーム｝の中で青いものはどれですか。」という質問に対して

タクシー　――×――　青い
郵便ポスト　――○――　青い
ユニフォーム　――×――　青い

という内容を伝えようとして

郵便ポストが青い　(2-2)

と答えた場合の「が」は総記の「が」と呼ばれる。この意味をグラフで視覚化したものが図２－１である。横軸は主部を含むグループである。

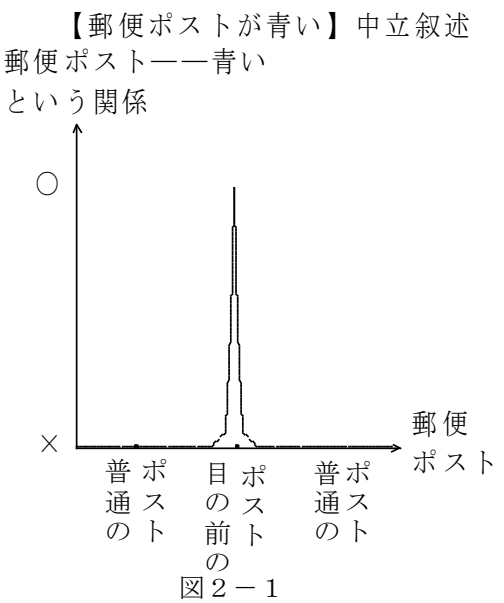


図２－１

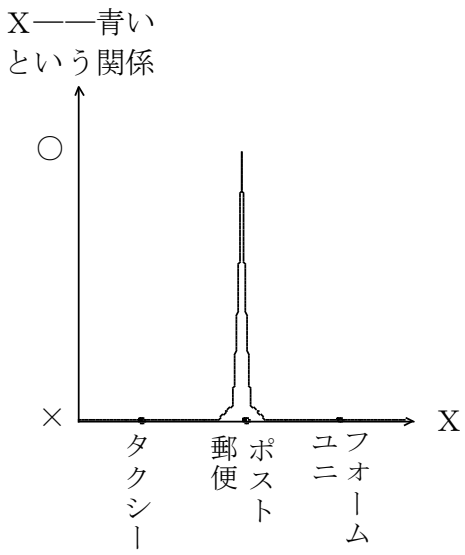


図２－２

3. 連体修飾節に含まれる中立叙述の「が」

ヨシオ――昨日の昼に公園でおにぎりを食べたという関係があったとする。

すると,「昨日の昼」を表すものとして

「ヨシオが公園でおにぎりを食べた時」

と言うことができる。これは

ヨシオ――○――公園でおにぎりを食べた

という関係が成り立つ時という意味であり,

ヨシオが公園でおにぎりを食べた　(3-1)

の部分は中立叙述である。

意味をグラフで表すと図３－１のようになる。

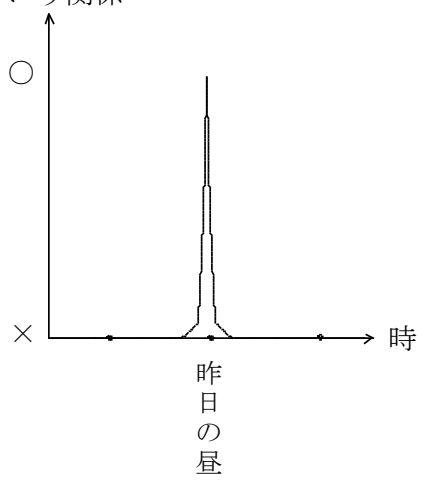


図３－１

また,「おにぎり」を表すものとして

「ヨシオが昨日の昼に公園で食べた食べ物」

と言うことができる。これは

ヨシオ――○――昨日の昼に公園で食べた

という関係が成り立つような食べ物という意味であり,

ヨシオが昨日の昼に公園で食べた　(3-2)

の部分も中立叙述である。

意味をグラフで表すと図３－２のようになる。

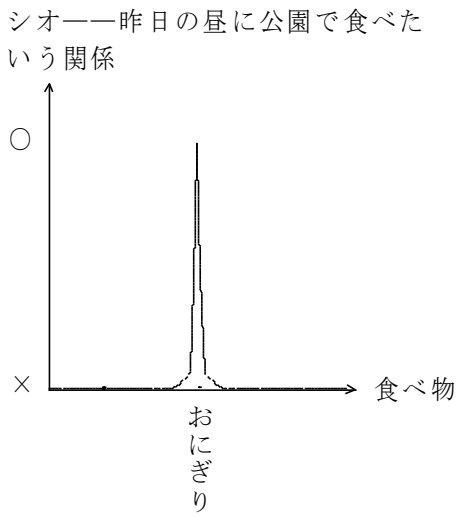


図３－２

また,「公園」を表すものとして
「ヨシオが昨日の昼におにぎりを食べた場所」
と言うことができる。これは
　ヨシオ――○――昨日の昼におにぎりを食べた
という関係が成り立つような場所という意味であり,
　ヨシオが昨日の昼におにぎりを食べた　(3-3)
の部分も中立叙述である。
意味をグラフで表すと図３－３のようになる。

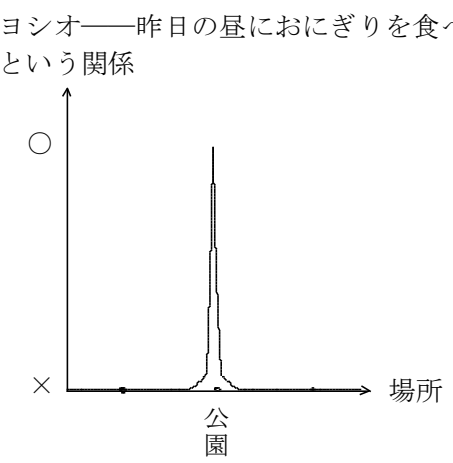


図３－３

　こうして，連体修飾節中の中立叙述の「が」の働きをグラフで見ると，被修飾語の表す集合が横軸になっており，それは述部の構成要素である。当たり前の話になってしまうが，主部を含む集合以外の集合の中で比較をしているのが中立叙述だということが判る。
　大野晋(1978)によれば，歴史的に見ると，「が」が現在の働きを持つに至ったのは，「自分が建てた家」「お玉が生まれた時」のように，「が」の上の体言と下の体言とを統合し一体化すること，つまり，連体修飾節の中での働きに始まる。そして，大野は，それこそが「が」の基本的性質であるとしている。

4. 述語節に含まれる総記の「が」と中立叙述の「が」

　大人が，子供に「象って首長いの？」と尋ねられた時に,
　　首――×――長い
　　鼻――○――長い
という内容を伝えようとして
　　象は鼻が長い　　(4-1)
と答えとする。この文の「鼻が長い」の部分は総記であり，意味をグラフで表すと図４－１のようになる。
　それに対して，
大人が，子供に「象って，どんな動物？」と尋ねられた時に,
（普通の動物）　鼻――×――長い
（象）　　鼻――○――長い
という内容，つまり，鼻というものは長いとは限らないが象については長いということを伝えようとして
　　象は鼻が長い　　(4-2)
と答えとする。この文の「鼻が長い」の部分は中立叙述であり，意味をグラフで表すと図４－２のようになり，横軸は動物の集合である。一般的に，
「～は～が～」という型の文の述語節が中立叙述である場合，意味をグラフで表すならば横軸は「は」の前の部分を含む集合になる。

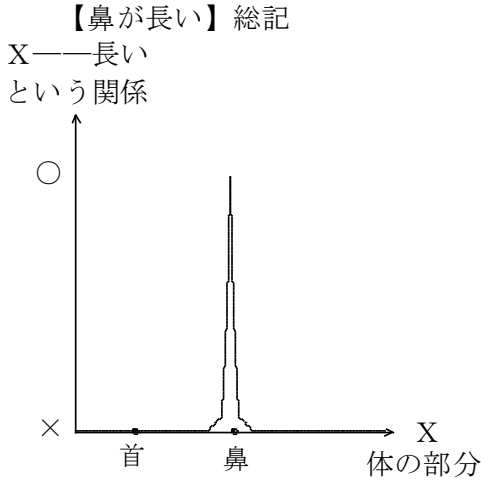


図４－１

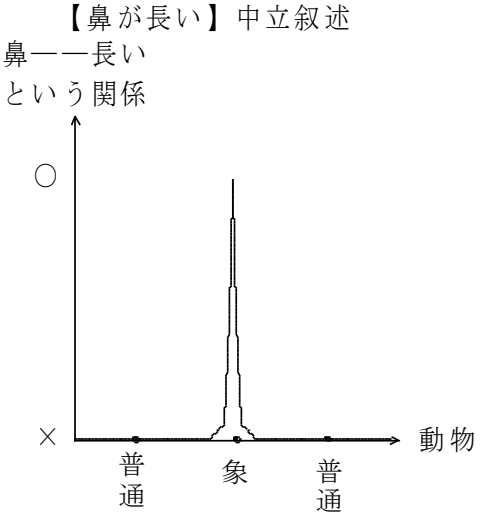


図４－２

人間またはコンピューターが
「象は鼻が長い」
「キリンは首が長い」

「アリクイは舌が長い」
といった文を聞いたり読んだりする度に，カウントをしたとする。これらの文は動物名と体の部位に関する情報なので，長いものの集合は表１のような２次元の配列としてまとめられる。表１の数値は聞いたり読んだりした回数の相対値で 10 が縦の列の最大になるようにしたものであると考えてもらいたい。聞いたり読んだりする回数を「長い」というイメージの強さに対応させるのは乱暴であるが１つのモデルであると了解してもらいたい。

長い	手	首	足	鼻	舌
ｷﾘﾝ	0	10	7	0	1
象	0	0	2	10	0
ﾃﾝｸﾞｻﾞﾙ	10	0	0	5	0
ﾀｺ	0	0	10	0	0
ｱﾘｸｲ	0	0	0	3	5
ｶﾒﾚｵﾝ	0	0	0	0	10

表１

総記の意味の(4-1)は表１の「象」の横の行で比較を行っている。
それに対して中立叙述の意味の(4-2)は鼻の縦の列で比較を行っている。
その比較に基づいて，人またはコンピューターは他者に「～は～が～」という型の文で情報を伝えることができる。

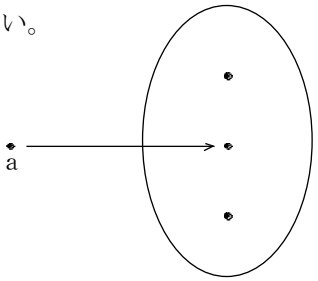
5. おわりに

「は」「が」でマークされるにしてもされないにしても，主部と述部の関係の大きな特徴は，主部と述部の性質を持つもの集合の要素の１つが対応されるということである。つまり，ａ――Ａという関係は

ａ∈{Ａという性質を持つものの集合}

であり，３者に共通である。「は」については，この関係を示す働きしかない。

ところが，「が」が，ある集合の中でａ――Ａという関係が特殊であることを示す働きを持つため，「は」は役割分担として普遍性を示したり対比のニュアンスを示したりするのだと考えられる。



Ａという性質を持つものの集合

ニュアンスの違い程度でしかないかも知れないが，このような「は」と「が」の働きの違いを認識していることは，コンピューターに日本語を聞き取らせたり話させたりする上で重要なだけでなく，情報を収集・蓄積する上でも有効だと考えられる。

参考文献

[１] 久野暲：『日本文法研究』, pp.27-47, 大修館書店 (1973)
[２] 大野晋:『日本語の文法を考える』,pp.36-38, pp.145-178, 岩波書店（1978）
[３] 庵功雄, 高梨信乃, 中西久実子, 山田敏弘:『初級を教える人のための日本語文法ハンドブック』, pp.259-268, スリーエーネットワーク（2000）
[４] 上林洋二:「Ⅳ.5.A 主題と主格(ハとガの表現)」,『日本語百科大事典』, 大修館書店（1988）.
[５] 井上和子,寺村秀夫:『日本文法小事典』, pp.151-163, 大修館書店（1989）
[６] 久保美織：JAPANESE SYNTATIC STRUCTURES AND THEIR CONSTRUCTIONAL MEANINGS, pp.10-41, ひつじ書房（1994）
[７] http://www.k3.dion.ne.jp/~okayasu1

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-10

地域特性を表すツイートの探索的閲覧支援システムの開発

Developing an Exploratory Tweet-Browsing System to Analyze Local Reputation Trends in Geographical Regions

森田 洋平[1]　白松 俊[1]　岩田 彰[1]

Yohei Morita [1], Shun Shiramatsu [1], and Akira Iwata [1]

[1]名古屋工業大学 大学院工学研究科
[1] Graduate School of Engineering, Nagoya Institute of Technology

Abstract: In this paper, we aim to develop an exploratory tweet-browsing system to analyze local reputation trends in geographical regions. We implemented functions for mapping the local reputation information, for extracting local feature words, and for exploratory browsing of local tweets. We conducted an experiment to evaluate whether users can extract local reputation trends about the snap election in 2014 by using our system. The experimental result indicated that users can extract local reputation trends about political topics in a particular prefecture by using our system. After this experiment, we also conducted a questionnaire survey about the usability of our system.

1. はじめに

近年，マイクロブログなどのソーシャルメディアが急速に普及してきており，個人が自分の意見を発信する機会が増えてきている．そのため，他の多くの人の意見や評判情報を得たいと思った時に手軽に収集可能となってきている．マイクロブログ上の意見や評判情報は投稿量が多く，様々な年齢，性別，地域の人々が投稿するため，アンケートやレビューでは得ることのできない情報を得られる可能性がある．本研究では特に，マイクロブログから地域に特有の意見や評判情報を得られる可能性に着目する．

毎日新聞では，立命館大学との共同研究プロジェクトで政党・政治家や有権者の Twitter 上のつぶやきやリツイート数，つぶやかれた単語などを分析するとともに，従来型の世論調査による結果内容との比較し，ネットでの呼びかけによる影響やユーザの関心などを調査している[1]．このように，新聞や雑誌においては世論調査，企業においてはマーケティング等にも利用されるなど，一般市民の日常の意見を抽出するための対象として非常に関心が持たれている．しかし，マイクロブログの利用方法が多様化していく中で，情報を収集・分析・可視化まで行えるアプリケーションというのは未だ少数である．また，世論調査やマーケティングなどにおいては，地域によって人々の意見が異なるため地域ごとに分析を行うことで新たな発見があると考えられるが，我々の知る限り，そのような研究は多くない．

そこで本研究では，日本国内での利用者も多く，大量の情報発信が行われているマイクロブログの一つである Twitter を対象に，世論調査や，マーケティング等に利用できるようにツイートを収集し，地域ごとに分析を行い，可視化するアプリケーションの開発を目指す．

2. 関連研究

2.1. 評判を抽出するための可視化の研究

Twitter を用いて意見・評判情報を可視化するアプリケーションとして，ヤフー株式会社が提供する，Twitter 上の投稿を検索できる「Yahoo!リアルタイム検索」において公開されている，つぶやき感情分析[2]がある．つぶやき感情分析は、検索したキーワードについてユーザがどのような感情を持っているかを、「ポジティブ」「ネガティブ」の割合でグラフ表示する機能である．

図 1 に示すように，ユーザが検索ボックスに語句を入力し検索すると，Twitter からその語句を含むツイートを収集し，画面左にストリーミング形式で表示する．画面右側には上から順に，ツイート数の時系列推移，ポジネガの割合，ポジネガの割合の時系列推移，トレンド語句が表示される．

また，膨大なツイートの中から評判情報を効率的

に取得するための可視化の研究として，村上ら[3]は任意の検索語句に対する意見がポジティブなものか，ネガティブなものかの評価をポジティブ度・ネガティブ度のキャラクターを用いた可視化と，評価極性ごとのツイート集合の表示，関連ワードの表示をするアプリケーションの開発を行っている．また，糸川ら[4]は，Twitter 上である話題に関する発言を分析，評価するための探索的ツイート閲覧システムについて提案している．

図 1　Yahoo!リアルタイム検索結果

2.2. 関連研究における課題・開発目的

前節で述べた関連研究のアプリケーションでは，政党に対する評判などの地域によって評判傾向が異なる題材に対しては人口が多く，投稿数の多い地域の評判傾向が全体の評判傾向として強く現れてしまい投稿量の少ない地域の特徴的な意見が埋もれてしまう危険性がある．

そこで本研究では，ある対象に対する評判を地域ごとに分析し，地域の特徴的な意見を抽出できるシステムの開発を目指す．

3. システム構成とインタフェース

本研究で開発したシステムの全体の流れを図 2 に示す．本システムでは，Twitter から TwitterAPI を用いてツイートを収集し，ツイートに対して前処理を行い，データベースに保管する機能を持つデータベースサーバと，クライアントから受けとったクエリーを含むツイートをデータベースサーバから受け取り，ブラウザに表示する情報の計算を行う，アプリケーションサーバで構成されている．

3.1. ツイートの前処理

3.1.1.　位置情報の取得

地域別で分析を行うため，ツイートの位置情報の取得が必要である．Twitter では，ユーザアカウントの設定からツイートに位置情報を付加することができる．しかし，筆者が収集したツイートに対して調査したところ，位置情報が付加されたツイートの数は 1%にも満たなかった．分析に用いるツイートのデータ数が少ないと意見に偏りが出る可能性や，有用な情報を得られない可能性がある．そこで本研究では，「ツイートの文章」・「ユーザプロフィール用の居住地」・「ユーザプロフィール」のテキストから位置情報の取得を試みた．

テキストからの位置情報の取得方法として，GeoNLP プロジェクトが開発している，ジオタギングツール GeoNLP[4]を使用した．GeoNLP は，自然言語テキスト内に含まれる地名や住所，施設名などにタグ付けを行い，そのタグに緯度・経度の情報を埋め込むツールである．

本研究で GeoNLP を用いて地名を解析する範囲は「ツイートの文章」と，ツイートした位置が，ユーザの居住地と近い可能性が高いことから「ユーザプロフィールの居住地」，「ユーザプロフィール」の 3 箇所とした．位置情報の優先度は，「ツイートに付加されている位置情報」＞「ツイートの文章」＞「ユ

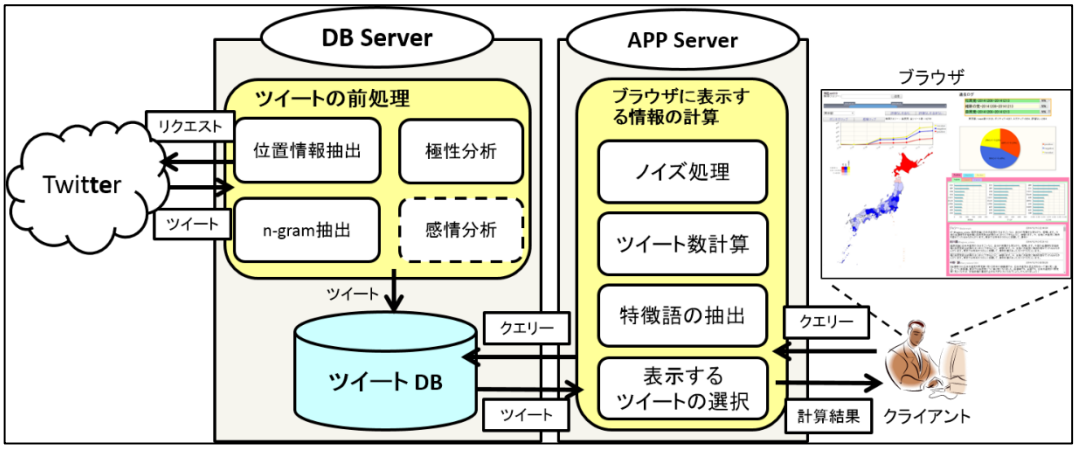


図 2：開発システム概要図

ーザプロフィールの居住地」＞「ユーザプロフィール」とした．また，地名語が複数あった場合は，得られた位置情報全てで，つぶやかれたツイートとして分析を行った．

本研究で収集したツイートに対して GeoNLP で位置情報の取得を試みたところ，全体の 30%に位置情報が付加された．

3.1.2. 評価極性分析

ある対象に関する評価を含むツイートを効率よく抽出するため，ツイートを評価極性で分類をする評価極性分析を行った．評価極性とは，その評価情報が「良い」や「好き」に代表されるポジティブな意味を持つのか，「悪い」や「嫌い」に代表されるネガティブな意味を持つのかを表す．本研究ではツイートをポジティブ・ネガティブ・評価なしの 3 種類に分類を行った．

評価極性の判定には，株式会社エクシングの提供する言語解析 WebAPI[6]を使用する．言語解析 WebAPI では「文単位」，「単語単位」のポジネガを判定することができる．

本研究では，ツイートを文で区切り，1 文単位で API の入力として，以下の条件で「ポジティブ」，「ネガティブ」，「評価なし」の 3 つの評価値にツイートを分類した．

1. ポジティブ・ネガティブに分類された文がない場合，ツイートの評価極性を「評価なし」と分類する
2. ポジティブ・ネガティブに分類された文がある場合，文数が多い方をツイートの評価極性とする
3. ポジティブ・ネガティブに分類された文の数が同数の場合，経験則により，より後半に出現する文の評価値を優先し，その評価値をツイートの評価極性とする

筆者が「ポジティブ」，「ネガティブ」，「評価なし」に分類した各 30 件，合計 90 件のツイートの文章に対し，言語解析 WebAPI を用いて分類した結果,正答率はポジティブが 76.7%，ネガティブが 80.0%，評価なしが 63.3%となった．

3.1.3. n-gram の抽出

n-gram とは，隣り合って出現した n 形態素を単語の単位とすることである．特に n=1 である n-gram をユニグラム，n=2 である n-gram をバイグラム，n=3 である n-gram をトライグラムと呼ぶ．本研究では，ツイートの文章がそれほど長くないことや，計算量を考慮してトライグラムまでを収集の範囲とした．n-gram の抽出のための形態素解析に MeCab[7]を用いた．

文章中に含まれる全ての n-gram を取得した場合，おそらく，その文章がどのような話題について書かれているか（政治に関するものか，スポーツに関するものか，など）を考える場合に有益とは思えない「は」，「が」，「です」といった助詞や助動詞，「ある」，「いる」，「あれ」，「それ」といった動詞や名詞でも，どのような話題の文章にも登場するような単語が特徴語として提示されると考えられる．このように，話題の種類と関連を持たないと考えられる単語のことをストップワードと呼ぶ．

本研究では，ユニグラムの場合はストップワードとなる語を全て除去し，バイグラム，トライグラムの場合はストップワードが語頭，または語尾に来る場合に除去した（一部例外を除く）．

また， 全ての n-gram に共通で接尾辞は一つ前の形態素と繋げて一つの形態素とした．接尾辞とは，「～さん」，「～的」といった単独では用いられず，常に他の語の下について，その語とともに 1 語を形成するもののことである．また，語尾の品詞が動詞であった場合は，原型に直して取得を行った．

3.2. インタフェース

ブラウザに表示されるインタフェースは図 3 である．インタフェースは，目的となる地域の特徴的な意見を含むツイートを探索的に閲覧できる用にするため以下の様な構成になっている．

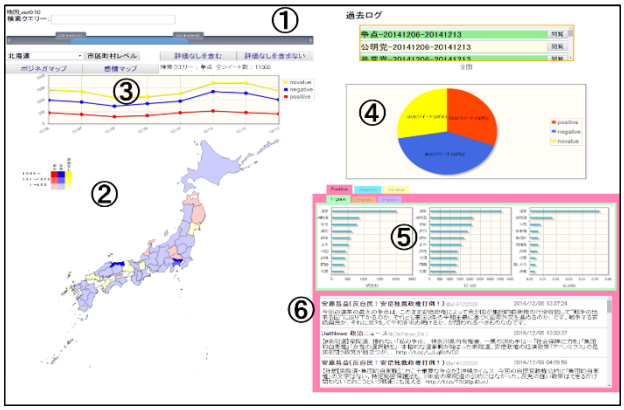


図 3:インタフェース全体図

① クエリー入力部

ユーザが分析したい情報のクエリーを入力する部分である．分析したいツイートの日付範囲を下部のタイムスライダーで入力することができる．一度分析したクエリーと日付範囲に対しては過去ログに追加され，分析を再度行うことなく閲覧することができる．過去ログには「クエリー」－「開始日」－「終了日」の順で表示されている．

② 評価極性マップ

評価極性ごとで一番割合の高いものを色で，ツイート数の多さを色の濃度で表現し，日本地図に可視化することができる．評価極性の色は，ポジティブが「赤」，ネガティブが「青」，評価なしが「黄」となっている. ツイート数は色の濃度で表現している．各都道府県をクリックすることで，クリックした都道府県のみのツイートを分析した結果（ツイート数の時系列グラフ，評価極性割合グラフ，特徴語グラフ, タイムラインビュー）を閲覧することができる．

また，都道府県選択後に「市区町村レベル」ボタンを押すことで市区町村レベルでの分析結果も閲覧することができる．

③ ツイート数の時系列グラフ

日付ごとのツイート数を，積み上げ折れ線グラフで閲覧することができる．横軸は日付，縦軸はツイート数，色は評価極性マップと同様である．また，グラフ上のマーカーをクリックすることで，クリックしたマーカーの日付のみのツイートを分析した結果（評価極性割合グラフ，特徴語グラフ，タイムラインビュー）を閲覧することができる．

④ 評価極性割合グラフ

現在分析しているツイート集合の評価極性ごとのツイート数と割合を閲覧する事ができる．色は，評価極性マップと同様である．

⑤ 特徴語グラフ

評価極性ごとのツイート集合における特徴語を 3 つの指標で抽出した結果閲覧できる．特徴語グラフは左から頻度法，TF-IDF，情報利得と自己相互情報量を組み合わせた指標の順で並んでおり，TOP10 まで表示されている．情報利得（以下 IG と呼ぶ）は，ある特徴が出現したか否かという情報が，クラスに関する曖昧さをどの程度減少させるかを表す尺度であり，トピックの検出などにおいて広く用いられている．ここでは，システムで収集した全ツイート集合$\mathbf{C}$に対して，入力された期間における，クエリーで入力した単語を含むツイート集合c^+を正例，それ以外のツイート集合c^-を負例とする．特徴量を求めたい語f_iの IG の計算式を以下に示す．f_i^+は語f_iが現れるツイート集合，f_i^-は語f_iが現れないツイート集合を表す.

$$IG(C|F_i) = H(C) - H(C|F_i)$$

$$H(C) = -p(c^+)\,log\,p(c^+) - p(c^-)\,log\,p(c^-)$$

$$H(C|F_i) = -p(c^+, f_i^+)\,log\,p(c^+|f_i^+) - p(c^-, f_i^+)\,log\,p(c^-|f_i^+)$$
$$-p(c^+, f_i^-)\,log\,p(c^+|f_i^-) - p(c^-, f_i^-)\,log\,p(c^-|f_i^-)$$

$$C = \{c^+, c^-\} \quad F_i = \{f_i^+, f_i^-\}$$

IG の値が高い語は，入力された期間における，クエリーで入力した単語を含むツイート集合に現れやすいか，あるいは現れにくいかのどちらかの特徴を持っていると考えられる．そこで，本研究では，IG と自己相互情報量（以下 PMI と呼ぶ）の組み合わせにより，クエリーで入力した単語とともに現れやすい特徴語の抽出を試みる．PMI は 2 つの確率変数の共起のしやすさを計る尺度である．語f_iにおける PMI の計算式を以下に示す.

$$PMI(c_0^+, f_i^+) = log\,\frac{p(c_0^+, f_i^+)}{p(c_0^+)p(f_i^+)}$$

PMI の値が正となる語f_iのうち，IG の値が上位である語をクエリー入力で取得したツイート集合における特徴語として提示する．

また，上部の「ポジ」,「ネガ」,「評価なし」のタブをクリックすることで各評価極性のツイート集合の特徴語を閲覧できる．また，上部の「1-gram」,「2-gram」,「3-gram」タブで n-gram ごとの結果を表示させることができる．特徴語グラフをクリックすることで，クリックした特徴語を含むツイートのみをタイムライン上に表示させる事ができる．

⑥ タイムラインビュー

評価極性ごとの収集したツイートを閲覧することができる．特徴語グラフと同様，上部のタブをクリックすることで評価極性の切り替えができる．表示されるツイートは，特徴語グラフで表示される単語をより多く含んでいるものほど重要なツイートであると考え，ビューの上位に表示される．

4. 評価・考察

本システムを使用して, 被験者 10 人に対して下記の問題 2 問の解答結果と，システムに関するアンケートを実施することで，開発目的を達成できているか確認した.

問題1. 昨年 12 月 14 日に行われた第 47 回衆議院議員総選挙に関して，全国的に争点となっているもの，沖縄県で争点となっているものは何か．それぞれ答えよ．（複数解答可）

問題2. 沖縄県での「辺野古への新基地移設」に対する各政党・県民の立ち位置（賛否や支持政党とその理由など）をまとめてください．（沖縄県での候補者は自民 4 人，共産・社民・維新・生活・無所属から各 1 人出馬している）

前提条件として，本システムではクエリーを入力してから分析結果を表示するまでに時間がかかるため，クエリーは「争点」,「自民党」,「共産党」,「社民党」,「維新の党」,「生活の党」,「無所属」の 7 点が入力され，検索日付範囲は選挙の投票日となる 12 月 14 日の前日である 13 日から一週間前の 6 日であ

ると仮定して，分析をあらかじめ行っておき，過去ログからクエリーを選択してもらうことで実証実験を行った．

4.1. 問題 1 の解答結果

問題 1 の解答を集計した結果の一部を表 2 に示す．

表 1:問題 1 解答結果の一部

全国		沖縄県	
争点	人数	争点	人数
集団的自衛権	10	辺野古への新基地移設	10
原発再稼働	8	米軍基地	5
アベノミクス	5	集団的自衛権	3
消費税増税	5	建白書	2
特定秘密保護法	4	アベノミクス	1
社会保障	4	安倍政権の是非	1

全国的な争点としては「集団的自衛権」や「原発再稼働」，「経済政策」などが特に多く解答されていたが，沖縄県の争点としては「辺野古への新基地移設」を全員が解答していた．これは，沖縄県での最大の争点となっていたのが「辺野古への新基地移設」であり[8]，それに関わるツイートが多く提示されていたためである．
以上のような結果から，地域ごとで分析をすることで地域の特徴的な意見の抽出が行えていると考えられる．

4.2. 問題 2 の解答結果

問題 2 に関して，実際の事実を，箇条書してまとめると以下のようなことが言える．

①. 自民党は辺野古への新基地移設を推進している
②. 共産党，社民党，生活の党，無所属の候補者は移設に対して反対派である
③. 維新の党の候補者は移設について知事選では「県民投票」を呼びかけていたが，衆院選では中止，撤回を求めている
④. 維新の党本部は辺野古への新基地移設を推進している[9]
⑤. 共産党，社民党，生活の党，無所属の候補者が沖縄の全区で当選したことから県民は基地の移設には反対で，共産党，社民党，生活の党，無所属を支持する人が多い

①，②に関しては解答者全員が正しくまとめられており，自民党と共産党，社民党，生活の党，無所属の 4 党が対立関係にあることが記入されていた．
③，④に関して，維新の党が移設に対して推進派と記入した人は 5 人，反対と記入した人は 2 人，解答がなかった人が 3 人という結果であった．維新の党は，党本部と候補者で意見が違っていたことが原因で，推進派と反対派で意見がわかれてしまったと考えられる．また，維新の党本部と候補者で意見が異なっている，というところまでまとめることできた人はいなかった．
⑤に関して，解答者全員が「県民は反対派が多く，共産党，社民党，生活の党，無所属を支持」と記入していた．理由としてあげられていたものには以下のものがある．

- 県知事選で自民が推薦していた仲井真知事が基地建設に関わる承認をし，県民の怒りをかったため，自民以外を支持する人が多い
- 歴史的経緯から基地の撤去をもとめている
- オスプレイの危険
- 県民・市民ではない人が増えることへの不満

ツイートを実際に閲覧することで，「怒りをかった」や，「危険」といった恐怖心，「不満」など県民がどのように感じているかの意見情報が読み取れていた．

4.3. アンケートと解答結果

アンケートの内容は以下の 6 問である．解答には「1.大いにそう思う」，「2.少しそう思う」，「3.どちらとも言えない」，「4.あまりそう思わない」，「5.全くそう思わない」の 5 段階で解答してもらった．ただし，質問 6. は自由記述である．

質問 1. 本システムを用いることで地域の特徴的な意見を抽出するのに役立つと思いますか
質問 2. 本システムを利用することで効率的に意見をまとめることができると思いますか
質問 3. 本システムは意見の抽出に必要な機能が揃っていると思いますか
質問 4. 本システムは意思決定支援やマーケティング，世論調査などに役立つと思いますか
質問 5. 今後，本システムを使いたいと思いますか
質問 6. 意見を読み取れる上で役に立った機能、使いづらい機能、また、あると便利だと思う機能があれば記入してください（自由記述）

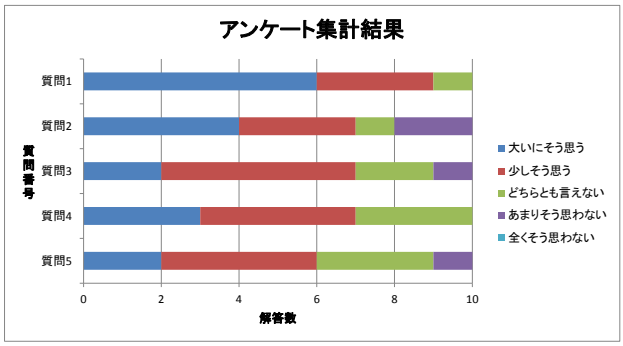


図 4:アンケート集計結果

質問 1 に対して肯定的な解答を示した人が 90%であることから、地域の特徴的な意見を抽出するのに有用なシステムの開発が達成できていると考えられる．これは問題 1 の解答結果で，全国の争点では出現していない「辺野古への新基地移設」や「建白書」などの沖縄独自の問題が争点として上げられていることや，問題 2 で県民の意見を正しく読み取れていたことからも推測できる．

質問 2,3,4,5 に対して肯定的な解答を示した人が 60～70%で質問 1 と比較すると少し減少した. 質問 4 に関しては，本実験では世論調査のためのみが対象であったため，意思決定やマーケティングにも有用であるかがわからないことが原因と考えられる．質問 2,3,5 に関しては, 自由記述にレイアウトや欲しい機能などに関する意見が多かったことからユーザビリティの低さが原因と考えられる．

4.4. 考察

本実験で，地域ごとに分析を行い，探索的にツイートを閲覧していくことで地域の特徴的な意見の抽出が可能であると示唆する結果を得た．しかし，今回出題した沖縄県の地域特性に対して，元々被験者が知識を持っていた可能性が拭えず，その場合，実験結果に有利に働いた懸念が残される．本来は，もっと認知度の低い地域特有の題材を出題するのが理想であった．しかし，今回対象にした選挙では，沖縄県以外で地域特有の争点が顕著でなく，認知度の低い地域特有の題材を多くツイートした地域を発見できなかった．今後は，そのような認知度の低い題材に対する評価実験を行う必要がある．

また，Twitter の利用者は若者が多いが，選挙での年齢別投票率は 20 代が一番低く，投票率の高い中年・高齢者層の人の意見を抽出できない可能性があり，題材によっては実際の事実とは異なる結果になる可能性がある[10]．そのため，Twitter の利用者層を考慮したシステムの利用が必要である．

5. おわりに

本稿では，Twitter から意見・評判情報を抽出する研究では十分に扱われていなかった，地域ごとの人々の意見・評判を考慮し，地域の特徴的なツイートの探索支援システムについて述べた．

また, 本システムを用いて第 47 回衆議院議員総選挙を題材に特定の地域の意見・評判情報を抽出することが可能であるか確認を行った．実際の選挙結果と比較して，地域の人々の意見を正しく抽出出来ていることを確認し，システムに関するアンケートにて開発目的を達成出来ていることを確認した．しかし，考察で示したように，被験者の元々の知識が結果に影響していた可能性もあるため，新たな題材でのシステムの評価をすることが今後の課題である．

謝辞

言語解析 Web API をご提供頂いた株式会社エクシング様に深謝します．本研究の一部は，JSPS 科研費若手研究(B) (No.25870321)の助成を受けたものです．

参考文献

[1] 選挙毎日: ネット選挙 ツイッター分析 http://senkyo.mainichi.jp/2013san/analyze/20130731.html

[2] Yahoo! Japan リアルタイム検索, http://search.yahoo.co.jp/realtime

[3] 村上奈緒，尼岡利崇，“Twitter 上で任意の検索語句に対するネガポジ度を判定し可視化するアプリケーションの開発と研究”，エンタテインメントコンピューティングシンポジウム (2014)

[4] Itokawa, S.,Shiramatsu, S., Ozono, T., and Shintani, T., "Estimating Feature Terms for Supporting Exploratory Browsing of Twitter Timelines," In *Proc. of IIAI-AAI 2013*, pp. 62-67 (2013)

[5] 北本 朝展, 相良 毅, 有川 正俊, “GeoNLP：自然言語文を対象とした高度なジオタキングに向けて”, In *Proc. of CSIS Days 2011*, No. D10 (2011)

[6] 言語解析 WebAPI, 株式会社エクシング, http://bigdata.joysound.com/about.html

[7] MeCab, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[8] 琉球新報，“衆院選(衆議院議員選挙)”, http://ryukyushimpo.jp/news/storyid-235313-storytopic-125.html

[9] 維新の党マニフェスト, https://ishinnotoh.jp/election/shugiin/201412/pdf/manifest.pdf

[10] 衆議院議員総選挙における年代別投票率の推移, http://www.soumu.go.jp/senkyo/senkyo_s/news/sonota/nendaibetu/

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-11

RDF を用いた名刺情報の構造化および可視化による人脈マネジメントシステムの提案

Personal Connections Management System by Structured Business Cards Using RDF and Visualization

小林 沙綾夏[1*]　井上 林太郎[2]　松下 光範[1]　笹嶋 宗彦[3]　高岡 良行[3]
Sayaka Kobayashi[1], Rintaro Inoue[2], Mitsunori Matsushita[1], Munehiko Sasajima[3], Yoshiyuki Takaoka[3]

[1] 関西大学総合情報学部
[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
[3] 株式会社ワイエムピー・ムンダス
[3] YMP-Mundus Corporation

Abstract: 本研究では、電子化された名刺情報を構造化し、ユーザが所有する情報の一部を各利用者間で共有することで、人脈リポジトリとして利用できるシステムを提案する。これまでの名刺を電子化して管理するソフトウェアでは、コンテンツ・ナレッジマネジメントが行われておらず、名刺情報はそれぞれ独立した情報として蓄積されている。提案システムでは、人物の持つ氏名や特徴などの情報を RDF を用いて記述し、人物間と人物を特徴付ける情報とのつながりを表現することができる。また、これらの人脈リポジトリを可視化することによって、直観的な操作と新たな関係性の発見を促し、人材の仲介や情報収集などの支援を可能にする。

1　はじめに

多くの社会人にとって、名刺の管理は重要な問題である。名刺をフォルダにまとめる手法では、名刺の枚数が多くなるほど管理が煩雑となり、必要な情報の検索に時間がかかる。このような問題に対して、様々な名刺管理ソフトウェアが開発されている。これらのソフトウェアでは、光学文字認識により名刺から氏名や所属などの情報を取得し、データベースに蓄積する。このデータベースを利用することで名前や登録日時によるソートや、タグ付け機能による整理・検索などを可能としている。電子化した情報を活用する手法の一つとして、Linked Data のコンテンツ・ナレッジマネジメントがある。この手法は、文書、人物、キーワード、イベントなどをデータとして管理し、関連するデータ同士をひも付ける (リンクさせる) ことで、多様な探し方や情報の提供をできるようにするというものである [1]。名刺管理ソフトウェアでは、名刺情報の電子化による名刺の整理は行われているが、情報は全て独立したものとしてデータベースに蓄積されており、その情報の活用は名刺の整理やデータを会社内などで統一するなどに留まっている。そのため、電子化された名刺情報にはさらなる活用の余地がある。

本研究では、名刺交換したというつながりを人脈と捉え、名刺情報を従来の名刺管理ソフトウェアのように名刺リポジトリとして利用するだけではなく、より有効に活用するためのデータベースとして利用するシステムを提案する。さらに共有した人脈ネットワークを可視化することで、検索の直感的な操作と新たな関係性の発見を促す。名刺情報と人脈の一元的な知識管理による業務支援の実現を目指すものである。

2　RDF を用いた名刺情報の人脈共有マネジメントシステム

知識を持つ専門家を見つけるためのシステムとして、組織内での知識の共有、活用のために研究・開発された Know-Who システムがある [2]。これらのシステムでは、従業員の誰がどのような知識を持つのかを文書やメールなどから収集、統合し検索可能にすることで専門家を探すことが可能である。

*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:k373668@kansai-u.ac.jp

名刺交換は、今後関わる機会がある人物とコネクションをもつための重要なきっかけとなる行動である。このことから、本研究では「名刺を交換した」というつながりを「人脈」と定義した。例えば、仕事の中で専門的な知識や技術が必要になった場合、自身の人脈を仲介として人材の紹介を受けることがある。知人が持つ人脈に対する Know-Who 検索が可能であれば、適切な人物に仲介を依頼できると考えられる。これを実現するために、井上らはグループで人脈を共有し、お互いが持つ人脈を検索可能にするプロトタイプシステムを実装した [3]。プロトタイプシステムでは、RDF (Resource Description Framework) を用いて人物の氏名や所属などのメタデータの記述を行う。RDF は主語、述語、目的語の三つ組みのデータモデルをとっており、主語と目的語の関係が述語で記述される。また、人物のメタデータを RDF で記述した FOAF (Friend of a Friend) では、知人の知人といった連鎖する関係を表現できる [4]。プロトタイプシステムでは、この FOAF の枠組みを参考に、名刺情報を RDF を用いて記述し、人脈ネットワークを作成している。さらに、名刺データを RDF 化して、どのような人物の名刺を持っているのかを Know-Who データベースに蓄積している。ユーザは、この RDF 化した人脈リポジトリを共有するためのグループを作り、共有された人脈リポジトリから適当な人材もしくはコネクションをもつユーザを見つけて、仲介を依頼することができる (図 1 参照)。人物のメタデータには、予め名刺に記載された情報に加え、人物の特徴や出会ったときの状況など、ユーザが入力したメモが利用される。このメモから人物を特徴付ける単語をタグとして登録し、タグごとに他ユーザからの検索の可否を設定できる。

このシステムを使用したユーザの評価から、Know-Who 検索機能において、適切な仲介人を繰り返し検索できるなどの検索を補助する機能の実現と、関係性を視覚的に分かりやすく表示するなどの改善が必要であることが分かった。

3 デザイン指針

本研究では、前述したプロトタイプシステムに基づいて、ユーザの利用の観点からデザイン指針を作成した。プロトタイプシステムの名刺データベースビューアでは、キーワードで名刺の検索を行い、検索結果の一覧がリストで表示される。名刺の詳細な情報は、個別の名刺のプロファイルページへ画面を遷移して表示させていた。これにより、ユーザが自身の人脈を整理するためには検索と画面遷移を繰り返す必要があり、人物間の関係が分かり難いという問題があった。また、ユーザがより適切な人材を探すとき、必要としている情報

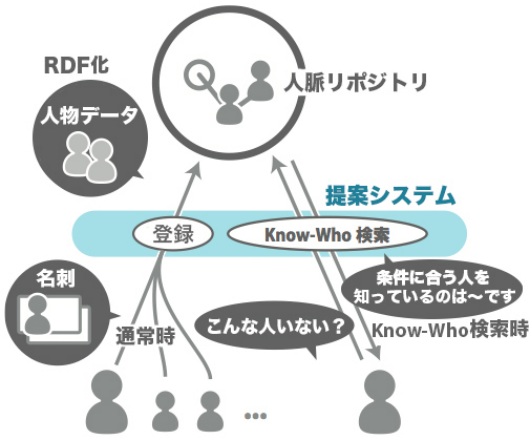


図 1: プロトタイプシステム利用の概念図

や技術を持つ人物を見つけるため、タグによる検索結果の人物に付与された他のタグ情報や、マッチした人脈を多く持つ共有ユーザの他の人脈を調べることが想定される。このような場合、ユーザは自身の求める人材およびその仲介が可能な他ユーザとタグ情報とのつながりに着目して検索を行うことが想定される。

これらのことから、人脈リポジトリにおいて、人物間と人物に関する情報 (タグ) のつながりが一見して把握できるような情報提示が求められる。また、キーワードを入力して検索することで、RDF によって関連付けられた人物とタグ情報のつながりを表示する機能を提案する。これを実現させるため、本研究では人物とタグをノードで、人物間または人物とタグのつながりをエッジで表し、人脈ネットワークを表現する。このようなグラフ表示を行うことで、ユーザの持つ人脈の規模などが直感的に把握できると考える。さらに、RDF のデータモデルを活用し Linked Data の形式で表現することで、同義のタグをひも付けることが可能となる。これにより、同義のタグが付与された人物間のタグを始点としたつながりを表現できるため、つながりに着目した人材検索が期待される。図 2 に可視化インタフェースの概念図を示す。図では、人物 B と C に付与されたタグ X が 1 つのノードで表示され、それぞれの人物へとエッジが伸びている表現をしている。

4 実装

4.1 構成

前述のデザイン指針で述べた可視化インタフェースを実現するために、本研究では、JavaScript の可視化ライブラリである D3.js[1] を用いてプロトタイプシステムの実装を行った。図 3 に、可視化した人脈ネットワー

[1] http://d3js.org/

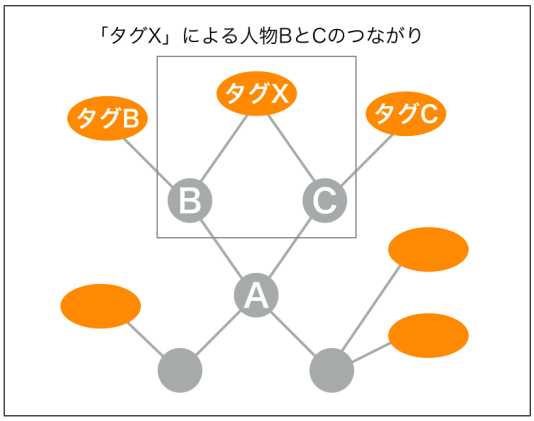


図 2: 可視化インタフェースの概念図

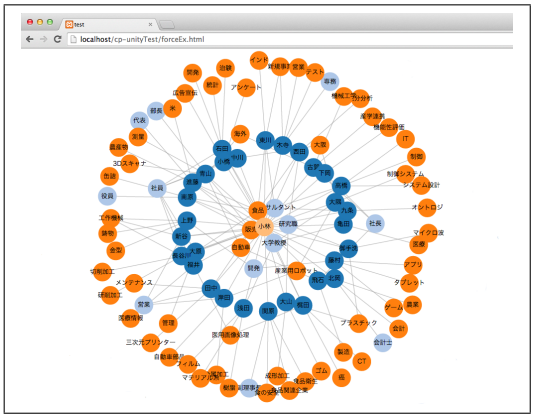


図 3: 人脈ネットワーク可視化の様子

クの初期状態を示す。この図は、「小林」の持つ名刺情報から構成した人脈ネットワークを可視化したものであり、人物ノードから各人物に付与されたタグノードへとエッジが伸びている。また、D3.js では複雑なグラフを描画するために、レイアウトオブジェクトとよばれるテンプレート機能が提供されている。提案システムでは、D3.js の Force Layout と呼ばれる力学モデル [5] を用いてグラフを描画している。これにより、ノードのドラッグやグラフの動的な変化など、直観的な動作を可能としている。

4.2 検索インタラクション

検索フォームに検索ワードを入力することで、共有されている人脈リポジトリ上の役職と公開タグが検索され、検索クエリを含むキーワードがプルダウンで表示される (図 4 参照)。図では、検索フォームに「食品」というキーワードを入力したとき、「食品」の文字を含む「食品衛生」「食品関連企業」のキーワードがプルダウンに表示されている。検索ボタンをクリックすると、検索クエリにマッチしたノードとそのノードが繋がっ

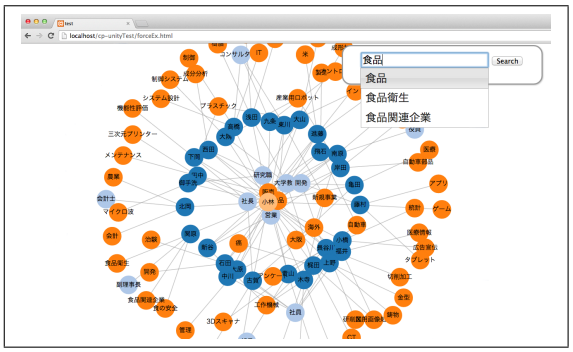


図 4: 検索フォームに「食品」という文字を入力した様子

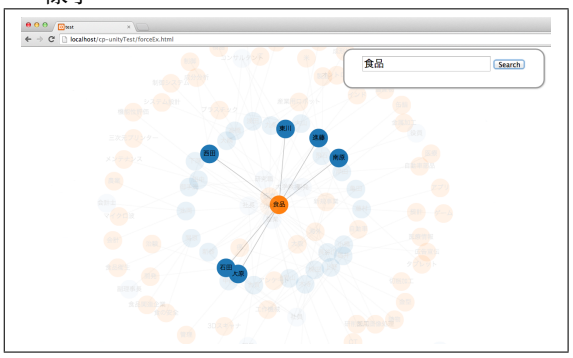


図 5: 「食品」をキーワードにつながり検索をした結果

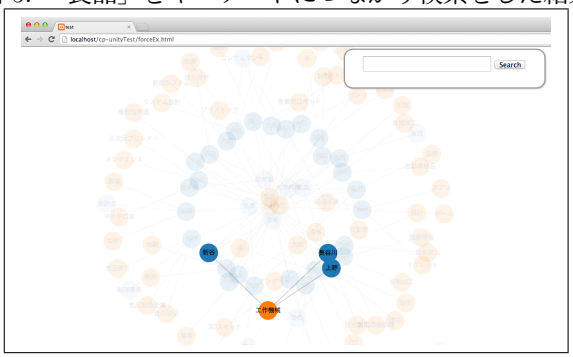


図 6: 「工作機械」のタグノードをダブルクリックした結果

ている人物ノードが強調表示される (図 5 参照)。図では「食品」というクエリで検索を行ったとき、「食品」のタグノードと 6 つの人物ノードとのつながりが浮き出るように表示されている。また、特定のタグノードをダブルクリックすることによっても、そのノードと繋がった人物ノードを強調表示することができる (図 6 参照)。図では、「工作機械」というタグノードをダブルクリックしたとき、3 つの人物ノードとのつながりが浮き出るように表示されている。

5 可視化インタフェースの機能検証

前章で述べたシステムの可視化インタフェースにおける問題点や改善点を洗い出すために、検証を行った。

5.1 構成

前章で述べたシステムでは、登録した名刺情報をRDFに変換し人脈リポジトリを作成することで、情報のつながりに重視した可視化を行うことができる。本研究では、この可視化インタフェースにおいて、ユーザが自身の人脈を整理・俯瞰する場合と、ユーザが共有した他ユーザの人脈から人材を検索する場合の2種類の検証を行った。前者では、架空ユーザであるAさんの人脈リポジトリのみを可視化させたインタフェースにおいて、Aさんがシステムを利用する背景と、Aさんが自身のもつ名刺を整理する際のインタラクションを想定した複数の検証シナリオを設定した。後者では、架空の共有ユーザであるBさんとCさんの人脈リポジトリを可視化させたインタフェースにおいて、架空ユーザであるAさんがシステムを利用する背景と、Aさんが Bさんと Cさんの人脈から人材を検索する際のインタラクションを想定した複数の検証シナリオを設定した。また、検索欄を使用した検索の有用性や問題点を確認するため、後者のインタフェースでのみ検索欄の使用を可能とした。そして、これらのシナリオを達成するために満たすべき要件から、検証協力者への質問項目をそれぞれ設定した。検証協力者は、情報系学部に在籍する大学生5名 (女性3名、男性2名) である。検証協力者には、シナリオに該当した質問に対してインタビュー形式で Yes または No で回答してもらった。これらの質問項目の内 No と答えた場合、その理由について検証協力者にコメントするように促した。なお、検証協力者の要望にあった際は、Yes の場合でもコメントしてもらった。また、検証の最後に事後アンケートとして、本システムを用いた操作について意見がある場合には回答してもらった。

5.2 検証結果

可視化インタフェースにおける、人脈整理・俯瞰と人材検索の観点からシナリオを想定し、検証を行った結果、検証協力者による知見から、いくつかの問題点と改善点が得られた。

まず、実装した可視化インタフェースに関して、検証協力者からつながりのノードを見る際の操作の煩雑さに対する指摘があり、人物とタグのつながりを強調する表示のため操作について改善が求められた。また、ノードを誤ってドラッグしてしまった場合に、グラフ全体が動いてしまい、始めから検索をし直す様子が見られ、ノードの不確定な動きに対する指摘があった。可視化した名刺情報の描画に関しても、「情報からつながりを辿っている際に、今まで見ていたノードを見失ってしまった」、「ノードが他のノードと重なっているとノードが隠れて見つけにくい」といった意見が得られ、誤操作や発見の阻害の問題点が指摘された。

人脈整理・俯瞰の検証では、「持っている名刺の中に特定の情報について詳しい人物が多いことに気付いた」という想定シナリオに対して、複数の検証協力者からシナリオを達成できなかったという結果が得られた。シナリオ達成のための操作では、「「食品」「食の安全」「食品衛生」などの、似た情報のノードは近い場所に位置されていてほしい」といった意見が得られ、人物に付与されたメタデータの分類に関する問題点があげられた。また、検証終了後には「登録した名刺情報が多くなった場合、同様の検索や整理ができないのでは」という指摘があった。

人材検索の検証では、シナリオ達成のための要件と、シナリオ共に全ての検証協力者が操作可能であるという結果が得られた。しかし、人物からタグ、タグから人物へと辿っていくような検索を必要としたシナリオでは、検証協力者の検証の様子から、何度も同じノードのつながりを表示させるなど、重複した操作が見られ、共有ユーザの発見までに時間がかかった。検証協力者からは、「一度検索した履歴を残しておきたい」といった意見が得られ、検索の軌跡が見えないことへの問題点があげられた。また、「共有する名刺情報が増えた場合、検索欄を駆使した検索が中心になるのではないか」といった意見が得られた。

6 考察

システムの可視化インタフェースでは、検証の結果から、操作における問題点が明らかとなった。これは、タグノードとそれにつながる人物ノードを強調表示した後、その人物ノードに付与された他のタグノードを確認する際に、人物ノードを1つずつダブルクリックしていくという繰り返しの操作の煩雑性によるものであると考察する。またシステムでは、2つのノードを一度に強調表示することが不可能であるため、つながった人物ノードを強調表示すると始めのタグノードの強調表示が解除されてしまうことからも、操作性に問題があると考える。この問題に対する解決策として、ノード同士のつながりの強調表示を2ホップ先まで表示できるようにする方法が考えられる。始めのノードをダブルクリックで強調表示させた状態から、1ホップ目のノードにマウスオーバーなどの操作をすることで、その先のつながりを強調表示ができれば、少ない手順で

検索が可能になると考える。その他に、可視化インタフェースのデザインにおいて、多数のノードが混在しているために起こるノードの誤操作や、ノード同士が重なって見えなくなるといったノードの描画の問題があげられた。また、人脈整理・俯瞰の検証シナリオにおいて、ノードを辿っている途中でノードが紛れてしまい、シナリオを達成するまでに至らなかった検証協力者がいることから、人脈グラフの表示の仕方に関する問題があると考える。さらに、今回の検証で使用したシステムでは、共有ユーザを含んだ各ユーザの登録した名刺情報はそれぞれ約 30 枚に設定した。しかし実際にシステムの実用化を想定したとき、ユーザが登録する名刺はより膨大な量になると予想される。可視化される情報が多いほどノードが混在し、グラフが膨大になるため、グラフの操作性に問題が生じることが考えられる。これらの問題を解決するためには、人脈の俯瞰や検索のための操作を損なわないような人脈グラフの表現が求められる。ノード同士が重ならないように跳ね返る処理を施したり、広がっている人脈構造を折りたたむことでグラフを最小化できるような機能が考えられる。

ユーザが自身の名刺を登録し、可視化することを想定した検証シナリオについて、可視化のデザインに関する問題点がいくつかあげられた。まず、想定シナリオに対して、複数の検証協力者からシナリオを達成できなかったという結果が得られた。これは、検証協力者が特定のタグノードとそれにつながる人物ノードの数ではなく、周りに見えている他のタグノードの情報の種類や多さに着目したことから、生じた判断だと考える。このような問題の解決策として、情報が似通った複数のタグノードを近い場所に位置させるなど、タグとして分類しているメタデータの自動的なカテゴリ分別や体系化の整備が必要である。これにより、人物に付与された情報をより簡便に俯瞰することができるため、人脈整理だけでなく、人材検索にも活用できると考える。また、検証終了後の指摘や、人材検索の検証で得られた検索欄に対する意見からは、ユーザの人脈整理のためのインタフェースにおける検索欄の必要性が示唆された。

人材検索を想定した検証シナリオでは、情報から人物、人物から情報へとノードを辿っていくような検索を行う必要があった。検証では、複数の検証協力者から、何度も同じタグノードを強調表示するといった操作が見られ、「自分が見たノードがどれだったか分からなくなる」といった指摘を受けた。この問題に対して、一度検索したノードや、強調表示したノードを固定した上で操作を続けられる機能や、検索の過程で強調表示したノードの履歴を残すなどの機能が必要であると考える。また、複数の情報に関して詳しい人物を検索するための想定シナリオでは、検証協力者が複数のキーワードを並べて検索を行おうとする様子が見られた。実装システムでは 1 つのキーワードでしか検索を行えないため、検証協力者は始めにキーワード 1 つ 1 つを検索したあと、強調表示されたノードの位置を思い出しながら検索を続けていた。このような問題に対して、より自身の希望に沿った人材を迅速に発見するために、複数のキーワードで一括検索できるような機能が求められる。また、検索結果のノードとそのつながりが強調される表示の仕方についても、検索した複数のキーワードがヒットした場合、キーワード同士のつながりが分かるような工夫が求められる。

7 おわりに

本論文では、RDF 化した名刺情報をもとに構築した人脈リポジトリを可視化することで、名刺の俯瞰や検索を可能とし、専門知識を持つ人とそれを必要とする人とを繋ぐためのシステムを提案した。人脈リポジトリの可視化方法と俯瞰・検索のための機能を検討し、提案システムのデザイン指針を策定した。デザイン指針をもとに試作システムを実装し、システム利用の想定シナリオを用いて検証を行った。今後は、実用化に向けたインタフェースの改善を行い、人脈を俯瞰する場合と、人材を検索する場合のための機能拡充を目指す。

参考文献

[1] 井形伸之, 西野文人, 粂照宣: Linked Data を用いた情報統合・活用技術, Fujitsu, Vol. 64, No. 5, pp. 464-470 (2013).

[2] 大山実, 東條弘, 榎本俊文, 佐藤哲司, 徳永裕史: コールセンタのための情報共有システム: Know-who 検索システムの適用, 信学技報, Vol. 100, No. 176, pp. 37-42 (2000).

[3] 井上林太郎, 松下光範, 笹嶋宗彦, 高岡良行: RDF を用いた名刺情報の構造化による人脈マネジメントシステムの提案, 第 28 回人工知能学会全国大会論文集, 2E3-5in (2014).

[4] Graves, M., Constabaris, A. and Brickley, D.: FOAF: Connecting People on the Semantic Web, Cataloging & classification quarterly, Vol. 43, Issue 3-4, pp. 191-202 (2007).

[5] Eades, P.: A heuristics for graph drawing, Congressus numerantium, Vol. 42, pp. 146-160 (1984).