

利他的貢献度を用いた協調的意見交換支援システム

Cooperative Opinion Exchange Support by Using Altruistic Contribution Evaluation

渥美 峻^{1*} 砂山 渡¹
Takashi Atsumi¹ Wataru Sunayama¹

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: Recently, exchange of opinions with service on the online is performed widely. However, in the exchange of opinions that there is the unspecified number of participant, there are the problems that opinions to lack sense of cooperation increase. Therefore, We use unselfish contribution degree to let the participant of the exchange of opinions be conscious of another person and group total profit. We suggest the exchange of opinions support system which promotes the input of the cooperative opinion by doing so it.

1 はじめに

近年、オンライン上のサービスを用いたコミュニケーションが活発になってきており、BBS（電子掲示板システム）や twitter, mixi といった SNS が広く利用されてきている。それに伴い、これらのオンライン上のサービスでは不特定多数による意見交換が頻繁に行われている。意見交換は、他人の意見や提案に対して同意や反論や新たなアイディアの提示によって成り立っている。そうした意見交換では、テキストによるコメント等の文字媒体によって行われることが多い。しかし、意見交換の参加者は、そうしたオンライン上でのテキストによる意見交換では、意見交換の過程を把握するためにそれまでに入力されたテキストを読まなくてはならない。そのため、意見交換の過程を把握するために多量の時間がかかってしまうことや、意見交換の全体像を直感的に理解することが難しい。そうしたことから、他人の意見をほとんど把握できず、意見交換の流れを無視した発言や、他人の意図を読み取れずに反論してしまう意見もあり、協調性に欠ける意見が多くなってしまいう問題点がある。

そこで、本研究では、意見交換における自他の意見入力の過程を直感的に理解でき、ユーザ間の協調的な意見交換を支援するシステムを作成する。特に、想定する意見交換は、オンライン上の不特定多数を対象とし、結論となる可能性のある回答を列挙し、その中から結論を決定する問題を対象とする。また、意見を入力した参加者が意見交換の場において結論を出すにあたり貢献度の高い意見の入力を促すことで、協調的な意見交換の支援を行う。

2 関連研究

本節では意見交換支援システムの関連研究を、協調作業、コミュニケーション支援と情報視覚化の観点からまとめて述べる。

コンピュータネットワークを活用して協調作業を支援する CSCW は、人間の協調作業を支援し、時間と距離の差を埋める有効な方法として各種の研究と試行が報告されている。VoiceCafe[2] は、音声メッセージを共有・操作できる。このシステムは、非同期型システムであり、例えばリアルタイムで参加していなくてもあたかもリアルタイムで行われている仮想会議の臨場感を得ることができる。非同期型システム上で、あるものを共有する点では、本研究と類似するが、本研究では、ユーザ間での画像の配置と意図の明確化によって、あるタスクを協調的に解決することを目的としている。本研究では、この理論を用いて、協調的な意見交換を支援するシステムを作成する。

また、協調学習において自分の発言を他人が参照し、他人に貢献していることを気づかせることで、他人を意識した議論を行うシステムがある [3]。このシステムは、ネットワーク上の協調学習で周囲を意識することで知識の獲得や内容理解を目的としている。本研究では、この理論を用いて、協調的な意見交換を支援するシステムを作成する。

多次元データがある観点に基づいて二次元または三次元情報として視覚化した上で、時間軸を設けることでそのデータの変化の過程を表示する研究がおこなれている。この変化の過程は表すことで、不連続なデータ間の関係を明確にすることが出来る。例えば、web ペー

ジ間の関係をリンクとした3次元表示のグラフをユーザに提供する納豆ビュー [4] やグラフ間の対応関係はアニメーションにより明示する研究 [5] がある。ユーザの行動履歴を再現した研究に、ユーザの視線履歴を可視化して着眼スキルを伝達する研究 [6] がある。この研究では、行動履歴をそのまま再現することで、言葉では伝えられないスキルを伝達しており、人間は他人の行動を見て感じ取ることによって、その時の感覚や状況をつかめることを示唆している。すなわち本研究では、意見入力となる画像の移動履歴を再現することによって、各ユーザの意見入力の様子を感じ取ってもらうとともに、参加者全員の意見集合の履歴の変遷を視覚的に確認できる環境を構築する。

3 協調的意見交換支援システム

本節では、提案する利他的貢献度を用いた協調的意見交換支援システムの詳細について述べる。本システムは、先行研究 [1] で用いられた既存システムを改良し、協調的な意見交換支援の機能を追加したものである。

3.1 対象とする意見交換

本システムでは、不特定多数による意見交換で結論を導き出す意見交換を扱う。すでに正解があるもしくは将来的に正解が決定する問題に対して、正しい解答を導き出すための意見交換や、正解がないものに対して意見交換を行うことで結論として決定する場面を想定する。特に、意見交換で出された結論が参加者全体の利益につながるものが適当だと考える。これは、意見交換において自分と他の参加者の利益につながらない意見交換では、参加者が損得を考えずに自由に発言できてしまうため結論を導き出すことが難しくなると考えられる。例として、順位付け、カテゴリ分類、役割分担といった意見交換が考えられる。

3.2 協調的意見交換支援システムの構成

本システムでは初期入力として、意見交換のテーマ、4つの回答エリアの命題画像、回答の選択肢、回答の選択肢画像を事前に与える。ユーザ（意見交換を行う参加者）はシステム起動時に、意見交換支援インタフェース（システムクライアント）にIDとパスワードを入力し、ネットワークサーバに通信要請を行い、意見交換参加者の意見入力履歴をダウンロードする。そして、ユーザは意見入力前にアニメーションで再現された入力履歴を確認することで意見交換の推移を把握を行う。その後、ユーザはシステム上に表示された選択肢画像を、各回答エリアの命題に対して適当なエリアに移動させ



図 1: 協調的意見交換支援インタフェース

配置をすることで意見の入力を行う。配置後に入力者は賛否の選択、コメントの入力を行うことで自身の回答の意図を明確にする。

本研究は、先行研究ではオフラインでの動作のみだったが、ネットワークサーバとシステムクライアントにより、意見入力履歴のダウンロードと、意見入力履歴の保存を可能とし、オンライン上での動作を行う。また、オンラインで不特定多数による意見交換をするにあたり、先行研究では意見入力時に回答エリアに新規で選択肢画像が生成されてしまう。そのため、1つの回答エリア内で同じ選択肢画像を統合することで、ユーザが増えた場合でも意見交換が行えると考えられる。

3.2.1 協調的意見交換支援システムに対する初期入力

協調的意見交換支援システムを利用するして意見交換を行うにあたり、システムに対して初期入力をしておく必要がある。ユーザが本システムを利用する際に初期入力は予め行っているものとする。初期入力は、意見交換のテーマ、4つの回答エリアの命題画像、回答の選択肢、回答の選択肢画像となる。4つの回答エリアの命題画像は、意見交換において、回答に対する適切な命題を4つ定義し、その命題にふさわしい画像を入力とする。回答の選択肢は、4つの回答エリアの命題に対して結論となる可能性のあるものを列挙し、回答の選択肢画像は列挙した回答の選択肢だと判断できる画像を入力とする。

3.2.2 協調的意見交換支援インタフェース（システムクライアント）

図1に協調的意見交換支援システムで使用するシステムのインタフェースを示す。画面上部に、意見交換のテーマが表示され、画面中央に、結論となる可能性のある選択肢画像の集合が横一列に並べられる。左上、右上、左下、右下に4つの回答エリアが存在し、各回答エリアの端にそのエリアの命題を表す画像が表示される。画面下部には、履歴の再生をコントロールするボタン群が表示され、画面右下の端には意見入力を確認させ保存するボタンが表示される。

また、本インタフェースはネットワークサーバと通信を行うシステムクライアントとしての機能も持つ。ユーザの入力履歴は自動的に入力履歴がサーバシステムに送信される。意見入力が終了したらユーザは保存ボタンを押すことで、サーバ側で入力した意見の保存を行い、入力した意見を確定させる。

3.2.3 ネットワークサーバ

ネットワークサーバは、サーバとなるコンピュータでシステムを起動を行う。起動しているサーバシステムは、クライアントシステムがネットワークを経由してアクセスしてきた場合に、アクセス情報を元にログインを行い、クライアントごとに送信されるデータを受信し管理を行う。受信するデータは、クライアント毎に割り振られた個別の情報、意見交換を行っている際のデータがある。クライアントが意見交換を終了し、履歴保存の命令を受け取った場合のみ、サーバシステムは意見交換の履歴の追加を行う。

3.2.4 ユーザ操作

協調的意見交換支援システムにおいて、ユーザが行う操作は、ID・パスワード入力、画像選択、意見入力、意図入力、入力履歴の再生の5つがある。ID・パスワード入力は、協調的意見交換支援インタフェースの起動時に行い、ユーザに割り当てられたIDとパスワードを入力することで本システムにログインを行う。画像選択は、意見入力をするにあたり、回答する選択肢画像をマウスでクリックすることで選択する。意見入力は、画像選択で選択した画像を、各回答エリアの命題に対して選択肢画像を移動し配置することで行う。意図入力は、意見入力の後に、入力した意見の意図の賛否とコメントの入力を行う。入力履歴の再生は、これまで行われた全てのユーザの意見交換の履歴の再生を入力再生コントロールボタンを操作することで行う。

3.3 意見入力

本節では、意見入力について述べる。意見入力の処理は先行研究の機能を使用している。

3.3.1 ユーザの意見入力

意見入力は、4つの回答エリアと選択肢画像によって行われる。

ユーザは、選択肢の画像を各回答エリアに移動させ配置することで、選んだ選択肢の画像が配置した回答エリアの命題に対して適切と考えた意見と意図の入力を行う。一度入力した意見については変更と削除が可能となる。変更方法は、自分が入力して配置した意見を、再び移動させ配置することで行える。意見の削除は、配置した意見を、画面中央の選択肢画像が並んだ選択肢画像表示領域に移動させることで、入力した意見を削除することが可能となる。

3.3.2 ユーザの意図入力

ユーザが入力する意図は、賛否とコメントの2つからなる。賛否は、自分が配置した選択肢が、各エリアの命題に対して賛否かを表す。本システムでは、「賛成」、「反対」のどちらかを選択し入力を行う。「賛成」は、まだ提案されていない新しい賛成意見としての入力と、他人が提案した賛成意見に対して同意の意図での賛成意見の入力を行う場合に選択をすることで、賛成意見を入力する。「反対」は、他ユーザが入力した賛成意見に対して否定的な意見を入力したい場合に選択することで、反対意見を入力する。各選択肢につき各ユーザは賛成・反対の意見を1回ずつのみ行うことが可能であり、1つの回答エリアに対して1人のユーザの賛成意見が配置できる選択肢は1つのみとする。コメントは、配置した画像と選択した賛否について、4文字以上20文字以下で簡単なコメントを入力可能となる。

3.3.3 意見・意図入力時の画像処理

意見入力時に行われるシステムの画像処理には、選択肢画像の移動、選択肢画像の配置、賛否・コメントの表示、配置した画像に対する意図の反映の4つがある。

選択肢画像の移動は、ユーザが選択可能な選択肢画像をドラッグ操作をする際に画像の処理を行う。ドラッグ中に画像はマウスカーソルを追従する。また、ドラッグ中の選択肢画像が配置可能な場合は画像周囲に黄色い枠を表示し、配置不可能な場合は画像に赤いバツの

マークが表示される。

選択肢画像の配置は、選択肢画像の移動中に、ドロップ操作を行う際に画像の処理を行う。ドロップした選択肢画像が配置可能だった場合、回答エリア内のドロップをした位置に画像を表示する。

賛否・コメントの表示は、入力された意図を、最新コメント表示エリアに表示する。最新コメント表示エリアは左端に配置した選択肢画像が表示され、その右横に賛成意見ならば青色の長方形が、反対意見ならば赤色の長方形が表示される。その右横の黄緑色の領域には入力したコメントとコメントを入力したユーザのIDが表示される。

配置した画像に対する意図の反映は、意見の意図を入力した際、配置された画像に入力した意図によって画像の処理を行う。賛成意見を入力した場合は、自身の回答選択肢画像の左下に青色の正方形を、反対意見を入力した場合は、自身の回答選択肢画像の左下に赤色の正方形を表示する。また、本システムは、配置された画像の枠を、入力意図の賛否の割合によって、賛成を青色で、反対を赤色で表示する。

3.4 意見入力履歴の再生

ユーザが意見を入力する前に、現在までの意見入力履歴の再生をアニメーションにより行う。アニメーションにより意見交換の過程を表示することで、直感的に意見交換の過程が把握できると考えられ、意見交換の場の把握の支援を行うことを目指す。

3.4.1 ユーザによる意見入力履歴の再生

ユーザは、現在までに入力された履歴を再生して確認することができる。再生は、アニメーションにより行われる。この機能によりユーザは、各回答エリアに選択肢画像が配置され意見入力されていった過程が確認できる。この情報をアニメーションから得ることにより、直感的に意見入力の推移が把握でき、意見交換の結論を決定するに向けて、より効果的な意見入力の支援ができると考えている。

3.4.2 意見入力履歴の情報と再生処理

協調的意見交換支援システムでは、ユーザが意見交換を行っている間、ユーザが移動した画像情報を取得し履歴再生に使用する。取得する情報は、「システムを使用しているユーザのID」「移動中の選択肢画像のID」「入力開始からの時間」「移動中の選択肢画像のxy座標」「意図入力時に選択した賛否」「入力したコメント」となる。

意見入力履歴の再生には、保存されている意見入力履歴をもとに、各ユーザの意見入力履歴を再現したアニメーションを再生する。アニメーションは全ユーザの入力履歴を入力された順に行うため、入力履歴の増加に伴って再生時間も増加するが、再生速度の調整機能を設けることで、再生時間を各ユーザにとって適切に設定できる。

3.5 利他的貢献度

本節では、利他的貢献度について述べる。

3.5.1 利他的貢献度の定義

本研究では、利他的貢献度を「協調的な意見交換の場において、参加者各自が結論を導き出すにあたり、協調的に意見交換に参加し、有益な意見を提案し、貢献できたかの度合い」と定義する。意見交換において、多くの同意が得られる意見は、他人もしくは意見交換の場において有益な意見の可能性が高い。また、多くの賛同が得られる意見は結論を導き出すにあたり重要なファクターとなってくる。そうした、意見を入力した意見交換の参加者は意見交換の場において、利他的貢献度が高いとされる。逆に、意見交換において、多くの同意が得られない意見は、他人もしくは意見交換の場におい無益な意見の可能性が高い。また、そうした意見に対して反対意見が多く入力されている意見は、協調的な意見交換において不必要な意見だと考えられる。そのことから、そうした意見を入力した意見交換の参加者は利他的貢献度が低いとされる。

3.5.2 利他的貢献度の数値化

協調的意見交換支援システムでは、各ユーザの意見入力に対して利他的貢献度の数値化し、視覚化することで協調的な意見交換の支援を行う。本システムでは利他的貢献度を数値化したものを利他的貢献ポイントと呼び、意見入力と利他的貢献ポイントの相関を表1に示す。意見入力と利他的貢献ポイントの相関は以下の3つに分類される。

- (A) 自身が入力した意見に対する利他的貢献ポイント
- (B) (A) 群で行われた入力に対して他ユーザの入力に応じた利他的貢献ポイント
- (C) 自身が他ユーザに対して入力した場合の利他的貢献ポイント

表 1: 意見入力と利他的貢献ポイントの相関

操作	操作	ポイント
(A)	既出でない賛成意見を入力した	+1
	賛成意見の入力	+2
(B)	他人の賛成意見が存在する	(A) のポイント $\times 2$
	参加者の過半数の賛成意見が存在する	(A) のポイント $\times 3$
	賛成意見より多くの反対意見が存在する	(A) のポイント $\times 0$
(C)	賛成反対の両意見を入力した	(A) または (B) のポイント $\times 0.5$

(A) 群は、自身の意見入力に対して加算される利他的貢献度ポイントとなる。これは、結論を導き出すにあたり、意見を創出した場合にポイントを与えている。また、入力された意見が新しい選択肢画像の配置ではない場合も、意見を入力することで結論を導き出すために貢献をしていると考えられるためポイントの加算を行う。

(B) 群は、(A) 群で行われた賛成意見の入力に対して、他ユーザの入力に応じた利他的貢献ポイントとなる。これは、自身が入力した賛成意見に対し、自身以外の賛成意見が存在する場合は、意見を入力した自身以外の利益とも考えられるため、多くの利他的貢献ポイントを与えている。また、賛成意見が過半数以上存在する選択肢画像に対し、自身も賛成意見を入力している場合は、結論を出すにあたり、全体の利益となる意見を入力していると考えられるため、より多くのポイントを与えている。しかし、自身が賛成している意見に対して、より多数のユーザが反対意見を入力している場合は、同意が得られない意見を入力しているとして、その入力に対する利他的貢献ポイントは与えないこととしている。

(C) 群は、自身が他ユーザの賛成意見に対して反対入力をした場合の利他的貢献ポイントとなる。反対意見は、意見交換の場において協調的でなくなる可能性も考えられるので、入力に対してデメリットを設けている。具体的には、反対意見を入力する選択肢画像の賛成意見の利他的貢献度ポイントを半分にする。反対意見の入力は、反対意見に対する同意が得られない限りは他ユーザに影響がないため、反対意見を入力する場合も協調的な意見交換ができると考える。

3.6 利他的貢献度のランキングによる視覚化

数値化した利他的貢献度を視覚化し、利他的貢献度のランキングによる視覚化を追加したシステムの全体像を図2に示す。本システムでは、意見交換の参加者のIDと、利他的貢献ポイントが高い順に上から順位付けしたものを表示している。これにより、利他的貢献

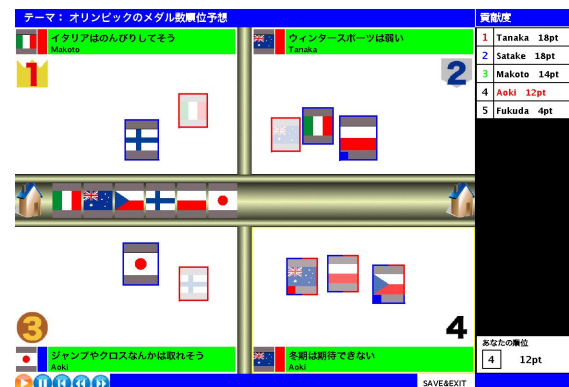


図 2: 利他的貢献度のランキングによる視覚化を追加した提案システム

ポイントの低いユーザに対して、他ユーザに対して利他的貢献ポイントが低いことを意識させることで、各自、高い利他的貢献ポイントを促す。また、順位付けしたIDと利他的貢献ポイントを表示する際、自信のIDと利他的貢献ポイントは赤色で表示する。さらにユーザは、自身の利他的貢献ポイントと参加者間での利他的貢献ポイントの順位を画面右下に表示することで、自身の利他的貢献ポイントを容易に把握できるようにしている。

3.7 利他的貢献度に関する入力意見の画像処理

本システムは、回答エリアに配置された選択肢画像に対して、利他的貢献度により表示を変更することで協調的な意見交換の支援を行う。

配置された選択肢画像に対して、利他的貢献度に関する画像処理は、選択肢画像サイズの調整と選択肢画像の明度調整の2つである。

選択肢画像サイズの調整では、賛成意見が入力により選択肢画像のサイズの拡大を行う。賛成意見が複数入力されている意見は、利他的貢献度も高くなる。そ

表 2: 実験テーマの詳細

テーマ番号	意見交換のテーマ	回答の選択肢
テーマ 1	ソチオリンピックのメダル数順位予想	イタリア・オーストラリア チェコ・フィンランド ポーランド・日本
テーマ 2	かわいい女優ランキングの作成	戸田恵梨香・佐々木希・新垣結衣 相武紗季・長澤まさみ・能年玲奈 堀北真希・榮倉奈々
テーマ 3	旅行するなら行っておきたい国ランキングの作成	イタリア・アメリカ合衆国・インドネシア オーストラリア・スペイン・ドイツ イギリス・フランス

のことから、画像サイズを大きくし目立たせることで、意見交換の参加者に利他的貢献度の高い協調的な意見の入力を促すことができると考える。

選択肢画像の明度調整では、入力意図の賛否の割合により選択肢画像の明度の変更を行う。配置された選択肢画像の中で、反対意見が多い選択肢画像は利他的貢献度も低くなる。そうした意見を目立たなくさせ入力を抑えることで、反対意見のない選択肢画像への入力を促すことにつながる。そうすることで、結論を導くにあたり、より利他的貢献度が高くなり、協調的な意見入力が増えると考えられる。

4 評価実験

本節では、提案システムを用いて意見交換を行ってもらい、利他的貢献度を用いた協調的な意見交換を支援する効果の有無を検証した実験について述べる。

4.1 実験内容

情報科学を専攻する、大学生、大学院生 12 名に 3 つのテーマについて、提案システムと比較システムを用いて、それぞれのテーマについて意見交換により結論を決定する実験を行った。12 人を無作為に 6 人ずつに分け、提案システムを使うグループと比較システムを使うグループで実験を行った。比較システムは、提案システムから 3.6 の利他的貢献度を視覚化する機能を省いたシステムとなる。システムは各自メールの添付ファイルにて配布し、ネットワークが利用できる場所から実験を行ってもらった。ネット上で集まった不特定多数の人が、全員で一つの結論を導き出す作業を想定し、その作業を協調的に行うことを支援する。また、被験者に対して意見交換の結果次第で景品を与えることで利益を付加する。景品は、各エリアにおいて被験者の意見と意見交換で決定した結論が一致している場合にポイントを付加し、総合したポイントにより景品を与えるものとした。実験は互いが特定できる情報を隠蔽した状態でを行い、実験期間は 1 週間とし、被験者

には期間内に最低 5 回はシステムを利用してサーバにアクセスを行ってもらった。

実験のテーマには、システムの利用を想定する場面として、組織内で匿名のまま意見交換を行い 1 つの結論を得る場面や、ネット上の SNS や掲示板で集まった人たちの意思決定を支援する場面が考えられる。その中で、用意された選択肢を用いた順位付けを行う問題について実験を行う。実験に用いたテーマの詳細を表 2 に示す。

テーマ 1 のソチオリンピックのメダル数予想問題は、将来的に正解が決定する問題であり、過去のデータや傾向、出場選手といった判断材料により意見交換を行うことが出来ると考えられる。客観的な視点による意見交換で、いかに協調が出来るかを目的としてテーマの設定を行ったテーマ 2 のかわいい女優ランキングの作成は、正解のない問題で判断材料が個人の価値観による問題となる。そうした、主観的な視点での意見交換で、多数の被験者が納得できる結論が決定できるかを目的としてテーマの設定を行った。テーマ 3 の旅行で行っておきたい国ランキングの作成も、テーマ 2 と同じく主観的なテーマに近い視点での意見交換となる。そうした主観的な意見と、各国の観光地等の客観的な判断材料から、いかに協調して結論を決定できるかを目的としてテーマの設定を行った。

4.2 実験結果・考察

4.2.1 各テーマにおける結論と賛成数、利他的貢献ポイントの推移による考察

各テーマの各回答エリアにおける結論と賛成数を、表 3 に提案システムについて、表 4 に比較システムに示す。また、各テーマの各エリアに配置された選択肢画像の種類の数を、表 5 に提案システムについて、表 6 に比較システムに示す。

表 3、表 4 から、各テーマの意見交換によって出された結論と賛成数について考察を行う。2 つの表より、提案システムは比較システムに比べ結論に対する賛成数が多い傾向にあることが分かる。この理由として、意

表 3: 各テーマの各エリアにおける結論と賛成数（提案システム）

	テーマ 1		テーマ 2		テーマ 3	
	結論	賛成数	結論	賛成数	結論	賛成数
回答エリア 1	フィンランド	3	新垣結衣	4	アメリカ合衆国	4
回答エリア 2	ポーランド	3	戸田恵梨香	2	フランス	3
回答エリア 3	日本	5	能年玲奈	2	フランス	3
回答エリア 4	チェコ	2	相武紗季	2	オーストラリア	2

表 4: 各テーマの各エリアにおける結論と賛成数（比較システム）

	テーマ 1		テーマ 2		テーマ 3	
	結論	賛成数	結論	賛成数	結論	賛成数
回答エリア 1	イタリア	2	新垣結衣	2	アメリカ合衆国	4
回答エリア 2	オーストラリア	3	長澤まさみ	2	イギリス	2
回答エリア 3	ポーランド	2	能年玲奈	2	フランス	3
回答エリア 4	チェコ	3	新垣結衣	2	オーストラリア	4

表 5: 各テーマの各エリアに配置された選択肢画像の種類数（提案システム）

	テーマ 1	テーマ 2	テーマ 3	平均
回答エリア 1	2	2	3	2.33
回答エリア 2	3	3	3	3.00
回答エリア 3	2	3	3	2.67
回答エリア 4	3	3	3	3.00
平均	2.50	2.75	3.00	2.75

表 6: 各エリアに配置された選択肢画像の種類数（比較システム）

	テーマ 1	テーマ 2	テーマ 3	平均
回答エリア 1	3	4	2	3.00
回答エリア 2	3	3	4	3.33
回答エリア 3	3	5	3	3.67
回答エリア 4	2	4	3	3
平均	2.75	4.00	3.00	3.25

見交換が進み結論を出すにあたり他の参加者と協調することで意見の収束を図ったためだと考えられる。一方で、比較システムの結論に対する賛成数は多いところもあるが、提案システムに比べ、少なくなっている。これは、表 6 の各エリアに配置された選択肢画像の種類数から、比較システムは提案システムに比べ、結論を出す段階で配置されている選択肢画像の数が多い傾向にあることが分かる。このことから、比較システムでは他種類の選択肢画像が配置されたことで、賛成意見が分散したことから結論に対する賛成意見が少なくなったことが分かる。

4.2.2 ソチオリンピックのメダル数順位予想における考察

ソチオリンピックのメダル数予想では、提案システムは表 3 から、回答エリア 4 以外は半数以上の賛成意見による結論となった。表 6 から、反対意見も少なく、他の選択肢画像も少なかったことから結論となった選択肢画像のサイズや明度の表示の変更により、意見が収束されたと考えられる。

4.2.3 かわいい女優ランキングの作成における考察

かわいい女優ランキングの作成では、表 3、表 4 から、結論に対する賛成数は提案システム、比較システムともに少ない傾向となった。これは、結論を選択する判断基準が個人の嗜好に委ねられるため、表 5、表 6 から各エリアに多数の選択肢画像が配置され、意見が分散したためだと考えられる。意図入力のコメントにおいても、「かわいい」といった明確な判断基準がないコメントも多かった。つまり、意見交換のテーマにおいて結論を導くための判断基準が個人の嗜好に大きく関与する場合、意見の変更・削除の操作が行われにくく、協調的な意見交換を行うことが難しかったと考えられる。

4.2.4 旅行するなら行っておきたい国ランキングの作成における考察

旅行に行っておきたい国ランキングでは、表 5、表 6 から、上記のかわいい女優ランキングと同じく、複数の選択肢画像が配置された。しかし、表 3、表 4 から結論に対する賛成数は参加者の半数を上回るものが多かった。これは、上記のかわいい女優ランキングと同

じく嗜好が関係するテーマではあるが、国の建造物や観光地、食事等の客観的な判断基準が明確なため、協調的な意見交換が行われたと考えられる。

本研究における実験では、被験者数が提案システムと比較システム、各6人ずつで行った。しかし、本研究で提案するシステムはオンライン上の不特定多数を対象としているため、実験で行ったより多くの人数を対象とすることが考えられる。参加人数が多くなるに連れて結果も変化することも想定され、1つは、配置される選択肢の種類が多くなることが考えられる。そうした場合、意見が分散し利他的貢献ポイントが上昇しにくくなると考えられる。また、本実験でも被験者により意見入力回数に差が見られたが、人数が多くなることで、意見入力を頻繁に行う人と行わない人により差が見られると考えられる。そうした場合、よりコンスタントに意見入力を行った人は、意見交換において貢献していると考えられる。

4.2.5 被験者の人数による考察

本研究における実験では、被験者数が提案システムと比較システム、各6人ずつで行った。しかし、本研究で提案するシステムはオンライン上の不特定多数を対象としているため、実験で行ったより多くの人数を対象とすることが考えられる。参加人数が多くなるに連れて結果も変化することも想定され、1つは、配置される選択肢の種類が多くなることが考えられる。そうした場合、意見が分散し利他的貢献ポイントが上昇しにくくなると考えられる。また、本実験でも被験者により意見入力回数に差が見られたが、人数が多くなることで、意見入力を頻繁に行う人と行わない人により差が見られると考えられる。そうした場合、よりコンスタントに意見入力を行った人は、意見交換において貢献していると考えられる。

5 結論

本研究では、先行研究に意見交換の場において結論を出すにあたり貢献を行った度合いを視覚化することで、オンライン上で不特定多数での協調的な意見交換を支援するシステムを提案した。

また、評価実験により、貢献の度合いを視覚化しないシステムと比べ、協調的に意見交換が行われ、多数の賛成意見により結論を導くことができるとわかった。

今後の課題として、提案システムを用いてより多くの被験者を対象に実験を行い、意見交換に参加している人数による分析を行っていくことを目標としていきたい。

参考文献

- [1] Yukihiro Tamura, Yuuki Tomiyama, and Wataru Sunayama : Opinion Exchange Support System by Visualizing Input History, IEEE Proc. of the 14th International Conference on Knowledge-based Intelligent Information Engineering Systems (KES 2010), Lecture Notes in Artificial Intelligence 6278, Springer, pp. 235-243, (2010).
- [2] 西本卓也, 北脇裕康, 高木治夫: 非同期型音声会議システム VoiceCafe, 情報技術レターズ (FIT2003 講演論文集), LK-005, (2003)
- [3] 林佑樹, 小尻智子, 渡邊豊英: 貢献アウェアネスに基づく協調学習対話インタフェース, 電子情報通信学会技術研究報告, ET, Vol.111, No.85, pp.27-32, (2011)
- [4] 塩澤秀和, 西山晴彦, 松下温: 「納豆ビュー」の対話的な情報視覚化における位置づけ, 情報処理学会論文誌, Vol.38, No.11, pp. 2331-2342, (1997)
- [5] 松下光範, 加藤恒昭: コンテキスト保持による探索的データ分析支援の枠組, 人工知能誌, vol.18, No.2, pp.251-264, (2006)
- [6] 藤本武司, 砂山渡, 山口智浩, 谷内田正彦: 視線行動の可視化による着眼スキル伝達支援, 人工知能学会論文誌, Vol19, No.3, pp.174-183, (2004)

ソーシャルネットワークユーザ間情報伝播量測定

Measuring the Quantity of Information Propagation of Users in Social Network

宮城 涼^{1*} 高木 友博¹

Ryo Miyagi¹, Tomohiro Takagi¹

¹ 明治大学大学院理工学研究科基礎理工学専攻

¹ Computer Science Course, Graduate School of Science and Technology, Meiji University

Abstract: These days information propagation method that transmits information to many people is proposed. Many previous method regard the network structure as important. However, these methods decide the quantity of information propagation without sufficiently considering a conversation topic of users. Therefore, We propose quantification method of information propagation considering communication between users in Social Network. We verified experimental results in twitter data.

1. はじめに

近年ソーシャルネットワークが台頭しており,その重要性が高まっている.国家レベルではアラブの春や東日本大震災でソーシャルネットワークの重要性が広く認知されるようになり,個人レベルではソーシャルネットワークを通じて今までは注目されていなかった人や物に注目が集まるようになった.企業レベルではソーシャルネットワークを積極的に活用しようという流れがあり,その活用方法が模索されている.企業のソーシャルネットワークを使用したマーケティング活動の一つにバイラルマーケティングがある.バイラルマーケティングとはソーシャルな影響を利用したマーケティングであり,口コミなどが人々の行動,意見,購買につながるという考え方に基づいている.このバイラルマーケティングに関連する研究は数多く行われており,ネットワーク構造を重視し,全体最適解を求める研究が特に活発に行われている.しかし,これらのネットワーク構造を重視した研究ではユーザの嗜好をあまり考慮していない.そのため,現実的な場面でネットワーク構造を重視した技術でバイラルマーケティングを試みようとしたとき,3つの課題が想定される.1つ目は,ターゲットユーザに向けてどういった情報を伝えたいのか分からない.2つ目は,ターゲットユーザにどのように情報が伝わっているのか分からない.3つ目は,ターゲットユーザが他ユーザにどのような情報を伝播しているのか分からない.これらの課題に対してユーザ間という局所に注目し,どのような種類

の情報がどのくらい伝播しているかを定量化した情報伝播量を測定し対処することが考えられる.情報伝播を扱った研究は存在するが,情報伝播量の測定値を十分検証した研究はなされていない.従って本稿では情報伝播量を測定し,測定値が実際の情報伝播を反映しているか検証する.具体的には, Twitter ユーザの各発言の嗜好を LDA によってトピックで表現し,それを時系列入力データとして移動エントロピー及びトピック移動エントロピーで情報伝播量を測定し検証する.

又,本稿で提案するトピック移動エントロピーを用いることによって,従来の移動エントロピーを用いた手法では測定できなかったトピックごとの情報伝播量を測定できることを示すと同時に,実際の情報伝播量をより反映した測定ができることを示す.

2. 関連研究

この章ではバイラルマーケティングに関連する研究について,ネットワーク構造を重視した研究とネットワーク内の局所に注目した研究に分類し説明する.

2.1. ネットワーク構造を重視の研究

バイラルマーケティングの主な目的は,口コミによってより多くのユーザに情報を伝播することにある.よって,より多くのユーザに情報伝播を行うことができるユーザを検知し,伝播させたい情報をそのユーザに流すことは理にかなっている.ソーシャルネットワーク内で多数のユーザからの被リンクを得ているインフルエンサーはこの条件を満たしている

*連絡先: 明治大学理工学研究科基礎理工学専攻
〒214-0034 川崎市 多摩区 東三田 1-1-1,
E-mail: miyagi.wsl@gmail.com

といえる.このような考えに基づき PageRank を拡張することによってインフルエンサーを検知する研究[1][2]が行われている.その中でネットワーク構造だけでなく,ユーザの嗜好をトピック分布で表現し PageRank を用いてトピック類似度を考慮した研究[3]がある.

又,ネットワークを重視した研究に影響最大化(Influence Maximization)がある.バイラルマーケティングに「影響力が強いと予測される少数のノード群をターゲットとし,社会ネットワーク上でのノードからノードへの情報伝搬を引き起こし,そして,それら情報伝搬の結果として,より多くのノード群にその情報を伝える」[4]という戦略が存在する.影響最大化はこの戦略に従い,ターゲットとするノード(ユーザ)の選択を目的としている.情報拡散モデルによって,他ノードに多くの情報伝播をできると判定したノードをターゲットノードとして選択する.情報拡散モデルの代表的なモデルとして,IC(independent cascade)モデルや LT(linear threshold)モデルなどがあり,これらを使用した研究[5]が行われている.又これらのモデルを拡張した様々なモデルが提案されており,その中で情報伝播確率を学習,予測するモデル[6][7]が提案されている.

2.2. ネットワーク内の局所注目の研究

ソーシャルネットワーク内の局所に注目した研究は数多く存在するがその中から,特定のソーシャルネットワークに関する研究,部分的なネットワークを使用した研究,ネットワークを使用しない研究を取り上げる.

様々なソーシャルネットワークを利用した研究が存在するが,特定のソーシャルネットワークのみに特化した研究がある.特に Twitter に関する研究において,Web Ecology Project[8]では10日間で12人の有名な Twitter ユーザのフォロワーについて調査している.Cha ら[9]はフォロワー数,リツイート数,メンション数を用いた影響力測定の比較をしている.

部分的なネットワークを使用した研究も行われている.Tang ら[10]はトピックベースの情報伝播の分析を行っている.Topical Affinity Propagation(TAP)という手法を用いて,ネットワークと各ユーザを表したトピック分布でメンバー間のトピックごとの情報伝播量を測定している.この研究はトピックモデルを用いてトピックごとの情報伝播量を測定しているという点で本研究と関連している.

ネットワークを使用しない研究も存在する.greg ら[11]は時系列なユーザの発言を LDA でトピック分布として表現し,それと移動エントロピーによってユーザ間情報伝播量を測定している.特に,本稿は彼

らの研究にならってシステムを提案している.彼らは条件付き相互情報量推定を用いて,移動エントロピーの推定値を使用している.また,ユーザ間情報伝播量のみ計算を行っている.それに対し,本稿ではトピック移動エントロピーを提案し,トピックごとのユーザ間情報伝播量を測定するという点で異なる.

3. 潜在的ディリクレ配分法

Blei ら[12]によって導入された潜在的ディリクレ配分法(LDA:Latent Dirichlet Allocation)はデータに隠れた潜在的意味(トピック)を推定する確率的トピックモデルの1つである.Hofmann[13]による確率的潜在意味解析モデル(pLSI:probabilistic Latent Semantic Indexing)におけるトピックと単語の多項分布それぞれにディリクレ事前分布を導入して拡張し pLSI の問題を解決した.一文書に複数トピックが存在することを表現でき,文書の生成過程をトピックに基づいて確率的に表現するモデルである.具体的には, LDA は指定したトピック数だけ各単語を割り振り,文書のトピック分布を出力する.

4. 移動エントロピー

この章では移動エントロピーについて説明した後,ソーシャルネットワーク情報伝播量測定でどのように使用するかを説明する.又,移動エントロピーを拡張したトピック移動エントロピーを説明する.

4.1. 移動エントロピー

移動エントロピー(Transfer Entropy:TE)はSchreiber[14]によって導入された.時間 t における確率変数 X, Y の要素をそれぞれ x_t, y_t とする. X に対する Y の影響を表す移動エントロピー $TE_{Y \rightarrow X}$ は式(1)のように定義される.

$$TE_{Y \rightarrow X} = \sum p(x_{t+1}, x_t^{(k)}, y_t^{(l)}) \log \frac{p(x_{t+1} | x_t^{(k)}, y_t^{(l)})}{p(x_{t+1} | x_t^{(k)})} \quad (1)$$

k, l は時間 t における状態から何個分遡った状態までを含めるかを表している.

4.2. 移動エントロピーを使用した情報伝播量測定

確率変数 X のエントロピー $H(X)$ を式(2)のように定義する.エントロピー $H(X)$ は X のあいまいさの量

を表す.

$$H(X) = - \int dx p(x) \log p(x) \quad (2)$$

このとき,相互情報量(Mutual Information)と条件付きエントロピーはそれぞれ式(3),式(4)で定義される.

$$I(X:Y) = H(X) + H(Y) - H(X,Y) \quad (3)$$

$$H(X|Y) = H(X,Y) - H(Y) \quad (4)$$

X の次の時刻の状態をX' とするとき,移動エントロピーは条件付き相互情報量 (Conditional Mutual Information)で式(5)のように表される.

$$TE_{Y \rightarrow X} = I(X':Y|X) = H(X'|X) - H(X'|Y, X) \quad (5)$$

2 行目の第 1 項は X を考慮したときのX'のあいまいさであり, 第 2 項は X と Y を考慮したときのX'のあいまいさである. よって,式(5)の 2 行目は Y を考慮したことによるX'のあいまいさの減少度を表しており,Y を考慮したことによるX'の予測性の向上度ともいえる.これを Y から X への情報伝播量とみなす.

又,実際の計算は $k=1, x_t^{(1)} = x_t, y_t^{(1)} = y_t$ として式(6)で計算する.具体的な計算方法については後述する.

$$TE_{Y \rightarrow X} = \sum_{x_{t+1} \in X'} \sum_{x_t \in X} \sum_{y_t \in Y} p(x_{t+1}, x_t, y_t) \log \frac{p(x_{t+1}, x_t, y_t) p(x_t)}{p(x_{t+1}, x_t) p(x_{t+1}, y_t)} \quad (6)$$

4.3. トピック移動エントロピー

本稿では,移動エントロピーを拡張したトピック移動エントロピー(Topic Transfer Entropy:TTE)を提案する.これによって特定トピックの情報伝播量を測定することができる.具体的には,Y からX' に伝わった情報が同一であったときのみ計算を行う. トピック移動エントロピーを式(7)で定義する.具体的な計算方法については後述する.

$$TTE_{Y \rightarrow X}(\text{topic}) = \sum_{x_t \in X} \{ p(x_t, Y = \text{topic}, X' = \text{topic}) \log \frac{p(x_t, Y = \text{topic}, X' = \text{topic}) p(x_t)}{p(X' = \text{topic}, x_t) p(X' = \text{topic}, Y = \text{topic})} \} \quad (7)$$

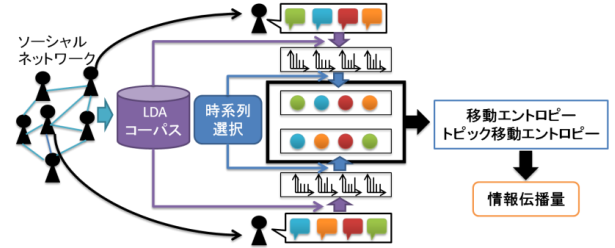


図 1 情報伝播量測定システム概要

5. 情報伝播量測定システム

この章では,LDA 及び移動エントロピー,トピック移動エントロピーを使用した情報伝播量測定システムについて述べる.ソーシャルネットワーク内のユーザの発言を LDA によってトピックとして表現する.その後,ユーザ間の発言を時系列に列挙し,その中からデータを選択する.選択したデータを入力として移動エントロピー,トピック移動エントロピーで情報伝播量を測定する.図1に情報伝播量測定システム概要を示しシステムの処理過程を説明する.

① LDA コーパス : 各発言のトピック分布

ユーザの発言をトピック分布で表現するまでの処理過程を説明する.形態素解析器 Kuromoji[15]を用いて発言文書から名詞を抽出し,発言を Bag-of-words で表現する.この Bag-of-words を LDA で解析することで各発言をトピック分布で表現する.

② 各ユーザ間の時系列トピック選択

各ユーザ間の時系列トピック選択までの処理過程を説明する. まず 2 人のユーザを選択し,ユーザ間の発言を時系列に列挙する.ユーザの発言は①のトピック分布で一番高い所属確率のトピック番号で表される.この時,一番所属確率が高いトピックが複数ある場合,移動エントロピーでは一番数が小さいトピック番号を選択し,トピック移動エントロピーでは全トピックを選択する.ところで,移動エントロピー,トピック移動エントロピーはどの発言がどの発言に対して影響を与えたかを定義しなければ計算できない.ここで,情報の受け手,情報の送り手について,任意の発言トピックを x_i, y_j ,発言数を m, n としたとき,時間の古い方から並べた時系列トピックは $X = \{x_0, x_1, \dots, x_m\}, Y = \{y_0, y_1, \dots, y_n\}$ と定義される.この 2 つの時系列に並んだ発言トピックを同一時系列上に時間の古い方から列挙し,以下のアルゴリズムに従いトピックを選択する.

1. x_0 を取得し,それ以降を探索
 2. y_j の次に x_i がくる組み合わせをすべて取得
- このアルゴリズムの具体例を図 2 に示す.

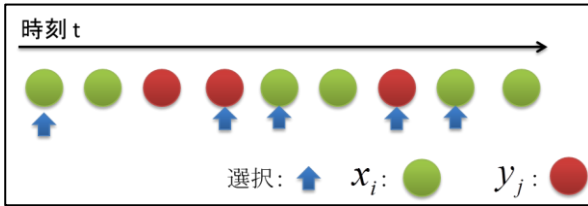


図 2 時系列トピック選択

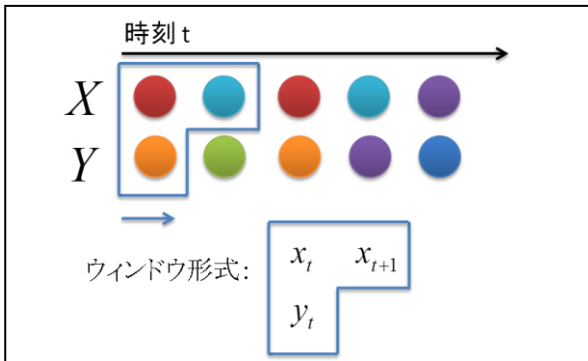


図 3 情報伝播量測定

$$p(x_{t+1}, x_t, y_t) = p(\text{blue circle, red circle, orange circle}) = \frac{2}{4}$$

$$p(x_t) = p(\text{red circle}) = \frac{2}{5}$$

$p(x_{t+1}, x_t, y_t)$ は4ウィンドウ中の2つが一致しているため、 $\frac{2}{4}$.
 $p(x_t)$ は5つ中2つが赤であるため、 $\frac{2}{5}$.

図 4 確率計算方法

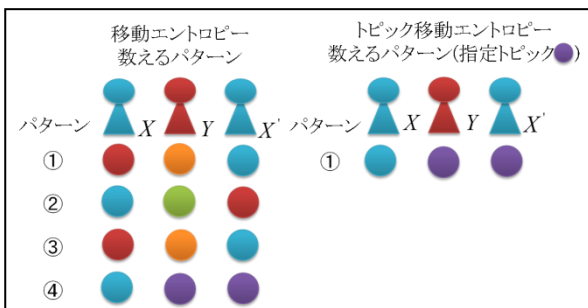


図 5 移動エントロピーとトピック移動エントロピーの差異

③ 情報伝播量測定

移動エントロピーを用いてユーザ間の情報伝播量を測定する処理を説明する.図 3 のように,ウィンドウをスライドさせて要素のパターン一致数から式(6)の各確率を計算する.例えば,図 3 のような X,Y の時系列トピックが並んでいる時, $p(x_{t+1}, x_t, y_t)$, $p(x_t)$ は図 4 の値になる.

トピック移動エントロピーも図 3 のように,ウィン

ドウをスライドさせて計算する.ただしトピック番号 topic を指定したとき, $x_{t+1} = y_t = \text{topic}$ のみ計算する.

移動エントロピーは Y のどのような発言も X に少なからずは影響を与えており, X' にその影響が反映されているという考えに基づいた情報伝播量を測定しているのに対し,トピック移動エントロピーは Y が X に影響を与えた時, X' は Y と同調的な発言をするという考えに基づいた情報伝播量を測定している.この差異を具体的に示すために,図 3 でウィンドウを左から右へずらした時の移動エントロピーが数えるパターンとトピック移動エントロピーが数えるパターンを図 5 に示す.ここで, X, Y, X' は X が Y に影響されて X' になったことを表している.移動エントロピーは Y の発言のトピックとそれに影響を受けた X' の発言のトピックの種類に関係なく伝播しているとみなし X, Y, X' の全発言パターンを数えることで情報伝播量を計算するのに対し,トピック移動エントロピーは Y と X' の発言トピックが同じとき,つまり X' が Y に影響を受けて同じような発言をするときのみ情報が伝播したとみなし情報伝播量を計算する.

6. 実験

6.1. 検証データ

情報伝播システムの検証として Twitter データを使用した.検証には「東京ディズニーリゾート PR 【公式】」のツイート情報に影響を受けているユーザが多いと見込まれる為, Twitter API を使用し「東京ディズニーリゾート PR 【公式】」をフォローしているユーザのうち,2,972 ユーザとユーザが発言した 421,336 ツイートを取得した.取得したツイートが作成された期間は 2009 年 5 月 8 日から 2014 年 3 月 3 日である.

特定のユーザ名から始まるツイートをリプライといい,そのユーザのタイムラインにのみ表示される.本稿では,情報の送り手のリプライは全て考慮せず,情報の受け手のリプライは情報の送り手に対するもののみ用いた.ツイート内の URL とハッシュタグは削除せず,形態素解析を行う.ここで,文書頻度の高い名詞はストップワードとして除去する.除去するストップワードは表 1 に示した通りである.

6.2. LDA の結果

検証データを入力として用いたときの LDA の結果を提示する.LDA のハイパーパラメータ α, β の値は 1.0, 0.3 とし,トピック数は 50, 75, 100, 125, 150 を定性的に評価し 75 とした.推論にはギブスサンプリング

を用い,繰り返し回数は1500回とした.トピック数75のLDAの結果は表2に示す.表2では各トピックに所属する所属確率が高い5つの単語を列挙している.

表1 ストップワード

http	一	二	三	ゼロ-zero	皆	四	九	八	五	月
------	---	---	---	---------	---	---	---	---	---	---

表2 「東京ディズニーリゾートPR【公式】」のLDA結果

解釈	topic番号	単語 トップ1~5				
カップル	0	男	馬鹿	女	嘘	LINE
ラブライブ	1	ライブ	キャス	七	chiaki	チアキ
?	2	方	訳	ファン	風	英語
?	3	放送	本日	女性	生	開始
天気	4	雪	雨	朝	天気	今日
自然災害 警報	5	県	市	駅	トウキョウ	運転
バイト	6	バイト	今日	パン	マン	伯父
?	7	絶対	駄目	love	筆	renai
?	8	七	六	恋	日	桜
原発	9	零	日本	ニュース	問題	新聞
ライブ	10	曲	映画	発売	決定	イベント
テレビ	11	姉	テレビ	山	疲れ	フジ
デザート	12	ケーキ	音	菓子	味	アイス
顔文字	13		月	ノ	日	
スバ	14	一緒	御腹	風呂	一杯	今度
?	15	所	今	名前	家族	人
お詫び・心配	16	御免	(´・`・´)	先輩	元氣	草
デート・遊び	17	今日	久々	練習	カラオケ	帰宅
六本木	18	ブログ	更新	応援	ロッポンギ	紹介
気持ち	19	気	最近	全部	所為	体調
軽い返事	20	笑い	本真	蓋	917	ため
金	21	円	万	金	以上	そう
プレゼント	22	フォロー	当選	プレゼント	ツイート	方法
動物	23	七		夜	猫	六
家族	24	子	母	いけメン	娘	ママ
ジョジョ	25	前	目	普通	適当	能力
歌詞	26	中	頭	電車	番	涙
冠婚葬祭	27	画像	卒業	結婚	式	花
セル	28	出演		監督	バーゲン	安売り
恋愛	29	事	為	勝手	喧嘩	魔
ハリーポッター	30	六	mymelody	sayamero	hide	七
携帯	31	顔	メール	携帯	絵	先
家	32	内	奴	家	婆	旦那
ソーシャルゲーム	33	ゲーム	参加	七	スマホ	リアル
曜日	34	日	人	誕生	診断	友達
?	35	六	live	chigatv	stream	momoclo
友人	36	友達	感じ	女子	男子	最近
DVD	37	七	話	六	箱	(´・`・´)
テスト勉強	38	物	勉強	一人	テスト	(´・`・´)
?	39	七	六	sakura	(´・`・´)	showma
Twitter拡散呼びかけ	40	人	フォロー	全員	夢	おた
嵐	41	嵐	ドラマ	ショウ	ジュン	役
季節旅	42	度	最高	何	春	夏
ジェジュン	43	今	個の学習書	ジェジュン	時代	ファン
運動・体	44	上	体	ライン	ロ	下
1年回想	45	年	今年	;	数	来年
Twitter	46	bot	Twitter	定期	ツイート	サイト
?	47	笑い	aya	mentaiko	☺	爆笑
感情	48	暇	意味	間	びっくり	本気
?	49	祭り	チーム	ゆき	ユウ	メンバー
占い	50	今日	仕事	運	健康	恋愛
テラスハウス	51	格好	ハウス	テラス	舞台	ケン
子育て	52	子供	爆笑	部屋	部屋	親
?	53	手	https	旅行	aruaru	動画
Twitter:ジャニーズ	54	方	声	物	最後	希望
Facebook	55	写真	無理	泣き	撮影	理由
出勤	56	時	分	頃	出勤	半
女性恋愛	57	本当	幸せ	気持ち	感謝	互い
ディズニー	58	ディズニー	トウキョウ	クリスマス	disney	ランド
心理	59	事	自分	人	言葉	心
休日	60	明日	休み	朝	学校	機
学校	61	金	先生	学校	高校	大学
?	62	六	七	班	今	場所
正月	63	予定	御節	オオサカ	年	料理
マラソン	64	回	七	結果	何	最終
?	65	笑い	706	aaii	4337	miuhiro
予定	66	楽しみ	疲れ	次	明日	(´・`・´)
くつろぎ時間	67	時間	後	週間	人生	睡眠
出前寿司	68	イン	弁当	予約	寿司	個人
バレンタイン	69	チョコ	♥	バレンタイン	無し	七
疲労	70	今日	昨日	疲れ	・	連
ご飯	71	毎日	御飯	箇月	ラーメン	カレー
風邪	72	感	脳	風邪	会社	病院
プロスポーツ	73	日本	選手	世界	試合	マオ
アニメ	74	パス	歌	鳥	キス	マイ

表3 移動エン트로ピーの情報伝播量上位10 ユーザ

Xの名前	総ツイート数	TE
千葉県 の担当	19	0.30103
しょーくん	173	0.30103
ちょこ	25	0.30103
中川美里	23	0.30103
(怜美@櫻宮.° —°)	8	0.30103
41.0℃プロジェクト	11	0.30103
yuka(*´`*)♡	50	0.30103
ごみ人間ごみ	12	0.30103
ちわ	29943	0.291807
はじまりの国さいごの話	159	0.286273

表4 トピック移動エン트로ピーの情報伝播量上位10 ユーザ

Xの名前	総ツイート数	TTE(58)
doughnut@甘党	58	0.15904
FK531(for D)	15	0.15904
ロマンスタ	14	0.15904
はな88	38	0.150515
じゅんじゅん@D垢	40	0.150515
ドナもち	86	0.150515
ゆうGoofFTW@D垢	78	0.150515
kendisney	19	0.150515
ペヌ吉	95	0.150515
安藤裕介	658	0.11928

表5 トピック移動エン트로ピーの情報伝播量下位3 ユーザ

Xの名前	総ツイート数	TTE(58)
ぼおと	76	-0.029035
HIKACHU	71	-0.029349
Minami	22	-0.058697

このアカウントは、千葉県のアカウントさんを、リスト @Chiba_eyes/lists に、カテゴリー分けしながら、フォローして居マッスル！ by ほっとあいず @hot_eyes ♪(^o^)♪ ツ(^o^)シ

図6 「千葉県 の担当」のツイート

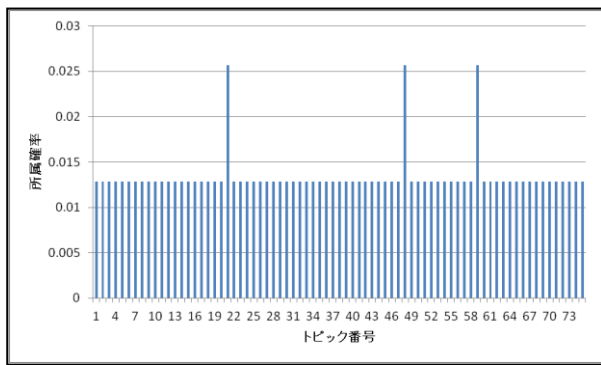


図 7 複数のトピックが一番高い所属確率である発言トピック分布

6.3. LDA の考察

表 2 中の「解釈」という項目はトピックの中身を観察し,主観で付けたタグである。「解釈」が「?」になっているトピックは,トピック中の単語から意味を解釈できない,または複数の意味が含まれており一意に解釈できないことを示している.表 2 のようにほとんどのトピックの意味は解釈することができる.特にトピック 58 は東京ディズニーリゾートと判断できる.

6.4. 移動エントロピー, トピック移動エ

トロピーの結果

表 3 は移動エントロピーを用いた情報伝播量上位 10 ユーザの結果であり,表 4 はトピック番号 58 のトピック移動エントロピーを用いた情報伝播量上位 10 ユーザの結果である.トピック 58 はディズニーに関するトピックであり,ユーザ「東京ディズニーリゾート PR【公式】」は 402 ツイート中 299 ツイートがトピック 58 のツイートだった為,トピック 58 のトピック移動エントロピーの情報伝播量のみ結果を示す.

6.5. 移動エントロピー, トピック移動エン

トロピーの考察

移動エントロピーとトピック移動エントロピーを用いた情報伝播量上位 10 ユーザの結果に関する考察を行い,測定した情報伝播量が実際の情報伝播を反映できているかという精度について述べる.評価方法は高い情報伝播量のユーザ間で情報伝播が行われているか主観で評価した.

① 移動エントロピー

移動エントロピーを用いた情報伝播量上位 10 ユーザ中 9 ユーザがディズニーに関するツイートを行っていないかった.また,「東京ディズニーリゾート PR【公式】」から影響を受けて発言したというような事は観察できなかった.唯一ディズニーに関してツイートをしていた「yuka(*~*)♡」も「東京ディズニーリゾート PR【公式】」から影響を受けて発言したというような事は観察できなかった.このような結果になった原因として,スパムユーザとトピック選択の問題がある.以下で両問題を具体的に説明し,対処方法を述べる.

上位 10 ユーザの特徴は同じようなツイートを繰り返しており,総ツイート数が少ない傾向があるスパムのようなユーザだった.具体的には「千葉県 の担当」は図 6 のようなツイートを繰り返しており,すべてトピック 5 に分類される発言である.このようなユーザはスパムユーザと判定して除去するという対策が挙げられる.

移動エントロピーを用いるとき,事前処理の時系列トピック選択で一番高いトピックが複数ある場合,一番数が小さいトピック番号のみを選択している.実際に発言トピック分布を観察すると図 7 のような複数のトピックが一番高い所属確率である発言がほとんどであった.図 7 の発言はトピック 20,48,58 が同じ所属確率であるにもかかわらず,現在の手法はトピック 20 と判定してしまう.このようなことから,適切な情報伝播量を測定できていないと想定できる.この問題に対して,greg ら[11]のように発言トピック分布全体を入力として条件付き相互情報量推定を用いて情報伝播量測定を行うことで対処することが挙げられる.

② トピック移動エントロピー

トピック移動エントロピーを用いた情報伝播量上位 10 ユーザ中 9 ユーザがディズニーに関するツイートを行っていた.特に「FK531(for D)」,「じゅんじゅん@D 垢」,「ドナもち」,「ゆう GoofFTW@D 垢」,「kendisney」は積極的にディズニーに関するツイートをしていた.このことから,トピック移動エントロピーはトピックを考慮することにより,移動エントロピーよりも正確な情報伝播量が測定できているといえる.ただし,トピック移動エントロピーも 2 つの問題が存在する.

1 つ目はディズニーに関するツイートをしていなかった「ペヌ吉」の情報伝播量が高くなったという問題である.原因は「ペヌ吉」の発言の多くがトピック 58 を含んでいた為である.それらの発言は図 7 のように一番高い所属確率が複数存在し,トピック 58 を含む.このようなあいまいな発言が多く存在し,ト

ピック 58 の発言として複数回加算する為、情報伝播量が高くなってしまった。このようなあいまいな発言を判定し除去することで上記のような問題に対処することができると思われる。

2 つ目は表 5 のように積極的にディズニーに関するツイートをしているユーザの情報伝播量が負の値になり、「東京ディズニーリゾート PR【公式】」へ影響を与えていると測定されてしまった問題である。しかし、「東京ディズニーリゾート PR【公式】」が Twitter 上の 1 つの発言に影響を受けるとは考えにくい。この問題に対しては不明瞭な点が多く今後の課題として挙げられる。

7. むすび

本稿では、確率的トピックモデルである LDA を用いてユーザの発言の嗜好をトピックで表現し、ユーザ間の情報伝播量を測定した。移動エントロピーを用いた場合、スパムやトピック選択の問題が存在した為、測定した情報伝播量は実際の情報伝播を十分に反映することができなかった。本稿で提案したトピック移動エントロピーを用いた場合特定のトピックのみ計算することで、測定した情報伝播量は実際の情報伝播をより反映し、ユーザ間でどのような種類の情報がどのくらい伝播しているかを定量化できることを示した。

考察では様々な問題点を挙げたが、それ以外の課題としては、計算時間や大規模ネットワークへの対処である。実際にバイラルマーケティングを行う際、この 2 点は非常に重要であるが、本稿ではこれらについて考慮していない。

今後の展望としては、可視化、影響最大化、推薦エンジン、リンク予測、スパム判定、ユーザの影響力への応用が挙げられる。

参考文献

- [1] Kwak, H., Lee, C., Park, H., and Moon, S.: What is Twitter, a social network or a news media?, Proceedings of the 19th international conference on World wide web, pp. 591-600, (2010)
- [2] Ding, Z. Y., Jia, Y., Zhou, B., Han, Y., He, L., and Zhang, J. F.: Measuring the spreadability of users in microblogs, Journal of Zhejiang University SCIENCE C, pp. 701-710, Vol. 14, No. 9, (2013)
- [3] Weng, J., Lim, E. P., Jiang, J., and He, Q.: Twitterrank: finding topic-sensitive influential twitterers, Proceedings of the third ACM international conference on Web search and data mining, pp. 261-270, (2010)
- [4] 木村昌弘, 斉藤和巳, 中野良平: 社会ネットワーク上の情報伝搬における強影響力ノード抽出の効率化, 電子情報通信学会論文誌, pp. 1004-1015, (2008)
- [5] Kempe, D., Kleinberg, J., and Tardos, E.: Maximizing the Spread of Inuence through a Social Network, Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137-146, (2003)
- [6] Goyal, A., Bonchi, F., and Lakshmanan, L.V.S.: Learning Influence Probabilitie in: Social Networks, Proceedings of the third ACM international conference on Web search and data mining, pp. 241-250, (2010)
- [7] Saito, K., Nakano, R., and Kimura, M.: Prediction of information diffusion probabilities for independent cascade model, Knowledge-Based Intelligent Information and Engineering Systems, pp. 67-75, (2008)
- [8] Leavitt, A., Burchard, E., Fisher, D., and Gilbert, S: The influentials: New approaches for analyzing influence on twitter, Web Ecology Project, Vol. 4, No. 2, pp. 1-18, (2009)
- [9] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K.: Measuring User Influence in Twitter: The Million Follower Fallacy, 4th Int'l AAAI Conference on Weblogs and Social Media, pp. 10-17, (2010)
- [10] Tang, J., Sun, J., Wang, C., & Yang, Z.: Social influence analysis in large-scale networks, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 807-816, (2009)
- [11] Ver Steeg, G., & Galstyan, A.: Information-theoretic measures of influence based on content dynamics, Proceedings of the sixth ACM international conference on Web search and data mining, pp. 3-12, (2013)
- [12] Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent dirichlet allocation. the Journal of machine Learning research, Vol. 3, pp. 993-1022, (2003)
- [13] Hofmann, T.: Probabilistic latent semantic indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50-57, (1999)
- [14] Schreiber, T.: Measuring information transfer, Physical review letters, Vol. 85, No. 2, pp. 461, (2000)
- [15] Atilika. kuromoji - japanese morphological analyzer. (オンライン) (引用日: 2014 年 3 月 6 日) <http://www.atilika.org/>

Bilibili 動画サービスにおける感情コメント分析

A study of emotional user comments analysis for bilibili.tv

呉 沢臣¹ 伊東 栄典²

Zeichen Wu¹, and Eisuke Ito²

¹九州大学大学院システム情報科学府

¹Dept. of ISEE, Kyushu University

²九州大学情報基盤研究開発センター

² Research Institute for Information Science, Kyushu University

Abstract: On user generated contents sites, user's comments include the user's impression to the content. The authors believe that user's comments can be used as the data mining resource to evaluate the contents. In this paper, the authors focus on bilibili.tv, a Chinese video sharing site, and analyze user's emotional comments on the site. They also show co-relation between comments and popularity such as number of replay and bookmark.

1 はじめに¹

近年, YouTube (youtube.com), ニコニコ動画 (nicovideo.jp) などの利用者投稿型動画共有サービスが人気である. 中国でも優酷 (优酷 youku.com) や土豆 (tudou.com), 搜狐 (sohu.com) および哔哩哔哩動画 (哔哩哔哩动画 bilibili.tv) が人気である. これらのサイトには膨大な動画が蓄積されている. 膨大な動画から, 利用者が求める面白い動画を探すためには検索システムが必要である.

我々は日本のニコニコ動画を対象に, 動画への視聴者コメントに基づく動画ランキング手法を提案してきた[1,2]. ニコニコ動画は視聴者投稿型の動画共有サービスで, 独自のコメントシステムで人気を博している. 本研究では, 中国の動画共有サイトである哔哩哔哩動画 (以下 bilibili と表記) 対象に, 視聴者コメントに基づく動画の品質評価を検討している.

Bilibili は, AcFun の会員である@bishi 氏が 2009 年 6 月 26 日に設立したファンサイト「Mikufans」を前身としている[3,4]. Mikufans は 2010 年 1 月 24 日から現在のサイト名「哔哩哔哩 (bilibili)」に改名され, 現在まで続いている. Bilibili は動画の実データを自サイトに保存しておらず, 別の動画サイトが保持する動画を使い, それに利用者向けのメタデータ (主に中国語) 提供している. また他サイトにある

動画再生に同期して, 弹幕と称する視聴者コメントを動画画面上に表示する機能や, 動画についての利用者のお気に入り情報の保持, およびファン向けのグループ作成機能などを提供している. これらの機能から, Wikipedia[3,4]には bilibili は寄生型サイトと表記されている.

近年, 日本の近代文化であるゲーム・漫画・アニメ, J-POP・アイドルなどのポップカルチャーをクールジャパン推進の一部として海外に積極的に展開する動きがある. 既にゲーム・漫画・アニメ・J-POP・アイドル等は海外に根強いファンが居るとの報告もある[5]. Bilibili も日本発の文化を受けたサイトの 1 つであると言える. 実際, Bilibili は日本のニコニコ動画およびニコニコ動画から発した Vocaloid 文化に強く影響されている. 前身サイトの Mikufans の Miku は初音ミクであるし, またサイトの構成はニコニコ動画に類似している. 動画再生に同期しての動画画面上への弹幕コメント表示は, ニコニコ動画のコメント表示と同じ機能である. 更に, ニコニコ動画サイトへの投稿動画を bilibili 内で表示する場合も多い. なお, Bilibili にはニコニコ動画以外にも, 米国や韓国系の動画も多い.

我々は, 動画に視聴者が付与するコメントのうち, 感情的なコメントを定量化し, その値を動画検索時のランキングに用いる事を検討している. 本論文では Bilibili を対象に, 視聴者コメントの統計的な分析を行った. Bilibili における感情コメント「233」について述べ, 再生数やブックマーク数などの人気度と感情コメント「233」の相関を調査する.

² 連絡先: 九州大学情報基盤研究開発センター,
〒812-8581 福岡市東区箱崎 6-1 0-1
Email: ito.eisuke.532@my.kyushu-u.ac.jp

2 Bilibili 動画タグとコメント

2.1 Bilibili のメタデータ

最初に, Bilibili に投稿されている動画メタデータを 2013 年 11 月から 2014 年 1 月の間に収集した. Bilibili では動画に「AV ID」と呼ぶ識別子を付与する. AV ID は文字 av と数字で構成される. そのため, 数字部分を 1 から 99 万まで 1 つずつ増やしてアクセスした所, 440,849 個の動画についてメタデータ (HTML 形式) を収集できた.

図 1 に示すように 1 つの動画のメタデータには, 題名, 投稿者 (登録者名) の他に, 動画再生回数 (播放), コイン数 (硬币数量), マイリストおよびブックマーク数 (收藏), 弹幕コメント数 (弹幕) の情報が有る. 動画再生に合わせて画面上を流れるコメントを「弹幕」と名付けられている.

时间	评论	发送日期
00:27	哥来组成弹幕!!!	08-09 01:09
00:30	锵锵锵!!!	08-09 01:10

図 1 Bilibili での動画利用状況情報 (av689970)

図 2 に動画のタグの例を示す. タグは視聴者により動画へ付与されている. ニコニコ動画と同様に, bilibili でも最高 10 個のタグを動画に付与できる. 図 2 の動画は, ニコニコ動画で「sm21443197」の識別子を持つ動画であるため, その情報がタグに記載されている.

タグ
VOCALOID
初音ミク
MITCHIE_M
CORT
神调教
下载4L
神弹幕
VIVA HAPPY
视频左下原画面

図 2 Bilibili でのタグ情報 (av689970)

2.1 タグの個数と頻度

集めた全動画メタデータからタグを抽出した. 一意なタグは 345,140 個あった. 1 つの動画にしか付与されていないため, 出現頻度が 1 回のタグは 25 万個あった. これは全体の 72.4% である. 表 1 に, 頻度 1 ~ 5 回のタグ個数を示す.

表 1 一意なタグの個数

頻度	タグ数	割合
1 回	249,825	72.4%
2 回	34,013	9.9%
3 回	14,595	4.2%
4 回	8,579	2.5%
5 回以上	38,128	11.0%
総数	345,140	100.0%

2.3 Bilibili の視聴者コメント

Bilibili には 2 種類の視聴者コメントが有る. 一つは動画再生に合わせて画面上に流れるコメントである. これは「弹幕」と呼ばれている. 弹幕コメントは匿名で, 投稿者を特定できない. もう一つは会員名付きのコメントである. これを本稿では「長コメント」と呼ぶ. 図 3 に示すように長コメントの例を示す.

用户名	评论内容	发送日期
抱枕萌宠	抱枕萌宠 (7/21) /	2014-02-13 17:10
哥哥	哥哥 (7/21) /	2014-01-19 19:47
Tres	Tres (7/21) /	2014-01-08 01:49
雪之	雪之 (7/21) /	2014-01-07 11:22
星夜	星夜 (7/21) /	2013-12-14 22:10
万年的观众	万年的观众 (7/21) /	2013-12-13 11:51

図 3 Bilibili での長コメント (av689970)

3 感情コメント「233」

我々はコメントに含まれた感情を定量化することで, 視聴者の感情に基づく動画ランキングを作りたいと考えている. その第一歩として, 弹幕コメントおよび長コメントにおける, 感情コメントを分析した.

3.1 コメント数

集めた動画メタデータからコメント情報を抽出した. 表 2 にコメントの件数を示す. 文字数が多い長いコメントも, 1 文字で終わる短いコメントも 1 件

と数えている。

表 2 視聴者コメント数

種類	コメント数	平均
弹幕	72,622,153	167.4
長コメント	5,173,862	11.92
総数	77,796,015	179.33

3.2. 弹幕における感情と「233」

視聴者が付与するコメントには、視聴者が動画に対して保持する感情が含まれている。本研究が対象とするコメント、特に弹幕コメントは、短いフレーズや絵文字が多い。

実際、Bilibili の動画を調査したところ、面白い動画には「233」や「2333333」のように 3 が多く続く弹幕コメントが多く与えられている。233 は中国語圏でのネットスラングで、大爆笑の意味で用いられている。この由来は、中国の BBS の「猫扑」における笑顔の絵文字である。猫扑では 233 番の絵文字が笑顔であった。そこから 233 が笑いの感情を表す文字として多用されるようになった。23333 のように 3 が多くなると笑いの感情が強い。

233 の用法は、ほぼ日本語のネットスラング「www」と同じである。日本の「www」は笑い(warai)の頭文字が由来である。日本のニコニコ動画では、w が多いほど笑いの感情が強い事を示している。

感情コメント分析の最初として、笑いの感情を表す「233」を分析する。ここで、弹幕コメントに「233」が出現する動画を 233 動画と呼ぶことにする。全 440,849 個動画の中で、233 動画の数は 134,411 個である。30.5%の動画が 233 動画である。

表 3 233 動画の統計

種類	動画数
全動画	440,849
233 動画	134,411

文字列「233」を含む弹幕コメントは 1,403,180 個である。233 動画では、平均 10.44 回「233」が出現する。

表 4 233 コメントの統計

	233 コメント数	平均
233 動画	1,403,180	10.44

図 4 に、縦軸に「233」の出現回数、横軸に出現回数による順序をプロットしたグラフを示す。図 4 に

示すように、233 コメントの分布は冪分布(power law)に従っている。

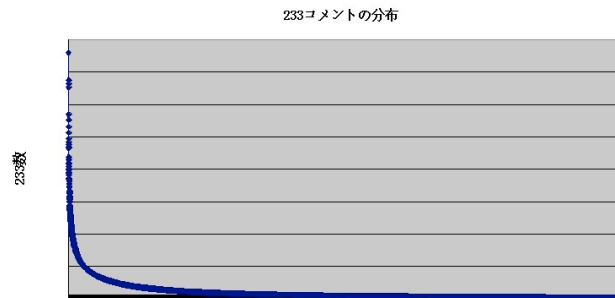


図 4 全 233 コメントの分布図

4 相関分析と考察

4.1 人気の度合い

コメントに 233 が多い動画は、笑える動画であろうと。大笑いできる動画は、人気の高い動画であると考えられる。その実証のため、人気の度合いと、233 コメント数についての相関を調べる。

人気の度合いに使える数値に、再生数(放送数)、長コメント数、弹幕コメント数、マイリスト数(ブックマーク者数)、評点、コイン数がある。これらの値は多いほど人気があると言える。

評点とコイン数はBilibiliが提供するユニークな視聴者評価システムである。この二つの数値は登録会員だけが与えることができる。登録会員が与えられる評点およびコインは、視聴時間および接続時間に比例する。ログインすると 1 日に 40 評点が貰え、また視聴時間が 1 時間につき 0.1 コイン貰える。1 人の会員は、シナリオ動画に最大 2 個コインを付与できる。

再生数は、会員でない人の視聴や、機械アクセスでも増える。一方、評点およびコイン数は、会員でないと与えられず、かつ利用度合いに比例した値しか付与できない。このため評点およびコイン数は視聴者から動画に対する支援度合いを表すと言える。

4.2 233 コメントと各人気度合いの相関

233 コメントの多い動画が、人気の高い動画であることを確認するため、各人気度合いと 233 コメントの相関を計算した。以下の表 5~8 に相関係数を示す。表では、P:再生数、N:長コメント数、D:弹幕コメント数、M:マイリスト数、H:評点、C:コイン数、233:233 コメント数である。

表 5 を見ると、再生数と評点の相関が高いことが分かる。マイリスト数と、評点・コイン数の相関も

高い。一方、コインと評点の相関は 0.56 で低くはないものの、高くもない。233 コメント数は、弹幕コメント数との相関は高いものの、他の数値との相関は少ない。233 コメント多いからといって人気があるとは言えない。

表 5 全 233 動画での相関係数

	P	N	D	M	H	C
N	0.38					
D	0.35	0.22				
M	0.61	0.37	0.26			
H	0.78	0.43	0.30	0.70		
C	0.47	0.51	0.21	0.70	0.56	
233	0.31	0.16	0.56	0.18	0.23	0.19

我々は、コイン数が最も人気度合いを反映すると考えている。そこで、コイン数の多い順に、上位 100, 1000, 10000 個の動画に限定した相関も算出した。その結果を表 6～8 に示す。

表 6 コイン順 top10000 動画 233 コメントの相関係数

	P	N	D	M	H	C
N	0.28					
D	0.12	0.11				
M	0.50	0.27	0.05			
H	0.70	0.34	0.04	0.61		
C	0.37	0.45	0.10	0.63	0.47	
233	0.15	0.06	0.50	-0.02	0.03	0.06

表 7 コイン順 top1000 動画 233 コメントの相関係数

	P	N	D	M	H	C
N	0.17					
D	-0.05	0.03				
M	0.31	0.16	-0.13			
H	0.61	0.30	-0.17	0.49		
C	0.21	0.45	0.01	0.47	0.34	
233	0.12	-0.003	0.47	-0.17	-0.05	-0.03

表 8 コイン順 top100 動画 233 コメントの相関係数

	P	N	D	M	H	C
N	0.23					
D	-0.17	-0.04				
M	0.60	0.14	-0.31			
H	0.70	0.51	-0.35	0.71		
C	0.33	0.54	0.01	0.41	0.53	
233	-0.06	-0.06	0.53	-0.23	-0.22	-0.12

やはり再生数と評点の相関が高い。またマイリスト数と、評点・コイン数の相関は高い。233 コメント数は、弹幕コメント数との相関は高いものの、他の数値との相関は少ない。こちらも、233 コメント多いからといって人気があるとは言えない。

5. おわりに

動画サイトや商品サイトにおける、利用者からのコメントは、その対象についての感情や意見を含む。我々は利用者が付与するコメントに基づく検索や推薦を検討している。本論文では、中国の bilibili 動画を対象に、視聴者コメントの感情を分析した。

感情分析の第一歩として、笑いの感情を表す「233」について分析した。各種の人気度合いと、233 コメントの数について相関を算出した所、233 コメント数は、弹幕コメント数との相関は高いものの、他の数値との相関は少ないため、233 コメント多いからといって人気があるとは言えないことがわかった。

本稿は第一段階の分析結果報告であり、今後多くの詳細分析が必要である。233 を含む様々な感情コメントを用いたランキング手法を提案し、それについて有効性を検証していきたい。

謝辞

本研究はJSPS 科研費 2350099 の助成を受けたものである。

参考文献

- [1] Naomichi Murakami, Eisuke Ito, Emotional video ranking based on user comments, Proc. of iiWAS2011, ACM, pp.499-502, 2011.
- [2] 村上直至, 伊東栄典: 動画コンテンツの視聴者コメントに基づくランキングとその評価, 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM2012), 日本データベース学会, F8-3, 2012.
- [3] Bilibili (Jan.20,2014). In Wikipedia: The Free Encyclopedia. Retrieved from <http://zh.wikipedia.org/wiki/Bilibili>
- [4] 哔哩哔哩 (Jan.20,2014). In Wikipedia: The Free Encyclopedia. Retrieved from <http://ja.wikipedia.org/wiki/%E5%97%B6%E5%93%A9%E5%97%B6%E5%93%A9>
- [5] 櫻井孝昌: 日本が好きすぎる中国人女子, PHP 新書, ISBN-10: 4569812422, 2013.

興味の移ろいを誘発するための横断的な情報アクセスの支援

Supporting Cross-Modal Information Access to Facilitate Transience of Interest

広川英智^{1*} 山下諒² 松下光範²
Eiji Hirokawa¹ Ryo Yamashita² Mitsunori Matsushita²

¹ 関西大学大学院総合情報学研究科

¹ Graduate School of Informatics, Kansai University

² 関西大学総合情報学部

² Faculty of Informatics, Kansai University

Abstract: This paper proposes a system that intends to facilitate inducement of a user's interest by showing various types of information related to photographs. Our proposed system focuses on meta-data (e.g., date, place) attached to each information and uses them to link with another information. For instance, this system connects a news article and music with a photograph by using meta-data on "date." Similarly, the system connects two news articles in which a common named entity appears are connected each other, or two songs which the common singer has sung are connected each other. With the system, we conducted a user study to observe a user's behavior under unrestrained information exploration task. Based on the results obtained, we found that the user's knowledge and interests influence the difference in the frequency of access to information that has different modality, and that every user often access to information that has common date.

1 はじめに

写真は撮影するだけでなく後日見返すことができる、我々にとって身近なメディアの一つである。イベントや旅行先で撮影した写真を後日一緒に見返すことで、友人との会話が弾み、楽しく撮影した当時の出来事や雰囲気振り返る事ができる。例えば、友人と野球観戦に行ったときに撮影した写真を一緒に見ている、当時の記憶を想起し「この年の優勝チームはどこだったっけ」という興味から会話が進んで、実際にスマートフォンで検索することがある。そこから、各チームの順位の推移について検索を続けていくうちに、「そういえばこの年ってどんなニュースがあったんだろう」といった、これまでとは違う興味を新たに持つことも考えられる。

しかし、写真がきっかけとなって始まる情報探索におけるユーザの情報要求は明確なものとは限らず、上手くクエリを生成できないことがある。また、Web 上から検索可能な情報は、テキスト、図表、画像、音楽、動画など多岐に渡るため、これらのモダリティの違いに関わらず、ユーザが求める情報に円滑にアクセスで

きる支援が必要である。本研究では、ユーザが撮影した写真と関係のある様々なモダリティの情報を結びつけることにより、写真をトリガとした“興味の移ろい”の誘発を目指している。それを踏まえて、写真が持つメタ情報の中から撮影日と撮影場所に着目し、これらを共通要素に持つ情報をユーザに提示することによって情報アクセスを円滑にするシステムを実装した。また、提案システムを用いたユーザ観察を行い、写真をトリガとした情報探索行動の振る舞いについて報告する。

2 関連研究

現在、一般的に用いられている情報検索技術は、キーワード検索エンジンが普及している。キーワード検索ではユーザが自らの要求をクエリとして入力することで解答が得られる。この方法は、ユーザの情報要求が明確である場合には有効に機能するが、漠然とした要求しか持っていない場合は、検索エンジンに入力するクエリを上手く生成できず、結果として適切な回答が得られないことも多い。このような曖昧な情報要求に基づく情報探索のモデルを Exploratory Search[1] と呼ぶ。Exploratory Search は探索過程で収集した情報か

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

ら知識を蓄えていき、情報要求を変化させながら情報アクセスを繰り返していく情報探索のモデルである。

Exploratory Search と同様に、情報探索の過程で情報要求が変化していくことを考慮したモデルとして、Bates が提案した Berrypicking[2] がある。Berrypicking は、探索を進めていく上で得られる文書や情報に基づいて新たなクエリを生成していく。また、その過程で探索者が持つ考えや情報要求を変化させながら、目的の達成や問題の解決を試みていく。

写真を閲覧したことで生じる情報要求も明確でないことが多く、写真をトリガとした情報探索もこれらのモデルに当てはめることができる。写真には、人や建物といった被写体や当時の出来事といったコンテンツ情報と撮影日や撮影場所といったメタ情報が含まれており、これらの情報から興味が生じることで情報探索が始まる。そこで、写真のコンテンツ情報やメタ情報を利用し閲覧や情報探索を支援する研究をいくつか述べる。

Yee ら [3] は、大量の画像群の各画像にメタ情報を付与し、そのファセットに基づき画像を探索できるシステムの Flamenco を提案している。このシステムは、画像データに付与されたメタ情報によるファセット検索とキーワード検索を用いることで、データの概観と詳細な検索をシームレスに行うことができる。

白鳥ら [4] は、大量の写真を撮影場所や撮影日に基づいて地図上に配置し、撮影者の移動経路とともに表示する PLUM を提案している。このシステムでは、ユーザが撮影した写真を撮影場所や撮影日に基づいてクラスタリングし、代表画像を選択することでクラスタに属する画像群を閲覧することができる。

田中ら [5] は、写真を情報探索のトリガとして扱い、時系列情報にアクセスする行為を考慮した Phickle を提案している。これは、写真に付随する撮影者の体験や活動に関する時系列情報と、写真のメタ情報やコンテンツに関連する時系列情報を結びつけることを意図したシステムである。

これらの研究は、写真をトリガとした横断的な情報アクセスについて検討していく上で有用なものである。しかし、これらの研究はアクセスの対象としている情報として写真に重点を置いており、写真のみに限った探索においては有効であると考えられるが、様々なモダリティの情報を対象としたアクセスについては考慮されていない。

以上を踏まえて、本研究では写真のメタ情報である撮影日と撮影場所に着目し、それらとつながりを持つ様々なモダリティの情報を結びつける。加えて、写真とのつながりだけでなく、同じモダリティの情報同士や違うモダリティの情報同士のつながりも考慮し、繰り返し行われる横断的な情報アクセスを円滑にする支援を試みる。

3 対象とするインタラクション

本章では、本研究で対象とする写真をトリガとした情報探索の例を挙げ、提案システムに必要な要素の検討を行う。

A さんは、甲子園で野球観戦した時の写真を見ていて、「この時以外にも、甲子園に行った写真があったはず」と興味を持った。撮影した写真の一覧から探していると見たい写真が複数枚見つかった。その中から A さんは一枚の写真に興味を持ち、その試合が阪神タイガースが延長戦でサヨナラ勝ちしたことを思い出した。その試合は 2008 年 9 月にあったことが写真の撮影日からわかったが、A さんは当時のことをあまり思い出せなかったので、同じ時期にどんなことが起きたのか調べ始めた。すると同じ月にリーマン・ブラザーズが経営破綻し、リーマン・ショックが起きた月と同じだということがわかった。そしてリーマン・ショックに興味を持ち、詳細を知るためにさらに詳しく調べ始めた。

この例では、A さんの興味は甲子園で野球を観戦した時の写真から同じ場所で撮影した別の写真に、阪神タイガースがサヨナラ勝ちした時の写真から同じ時期に起きたニュースに、最終的にリーマン・ショックに変化している。この探索過程において、A さんは写真の撮影場所や撮影日といった、閲覧している情報と共通の要素を持つ情報に興味を持ち探索行為を繰り返していることが伺える。本研究では、情報が持つ共通要素に着目し、ユーザが閲覧している情報と共通要素をもつ情報を提示してユーザの興味を誘発することで、円滑な情報アクセスを支援するシステムの実装を目指す。

4 システムデザイン

ここで、システムを実装する際に用いる情報について整理する。提案システムでは、ユーザが情報探索を行うトリガとして写真を用いる。写真はコンテンツ情報とメタ情報から構成される。コンテンツ情報とは撮影した写真に写っている人、建物、風景といった被写体や、撮影した時に起きている出来事などのことである。メタ情報とは、EXIF 情報¹(e.g., 撮影日、機種名、画素数)のことを指す。また、GPS 機能を搭載したデジタルカメラやスマートフォンでは撮影場所の経緯度を EXIF 情報に付与することができる。提案システムではこれらのメタ情報から撮影日の年と月、撮影場所を共通要素として扱う。

¹http://www.cipa.jp/exifprint/contents_j/01exif1_j.html



図 1: 写真を選択したときの画面



図 3: ニュース記事を選択したときの画面

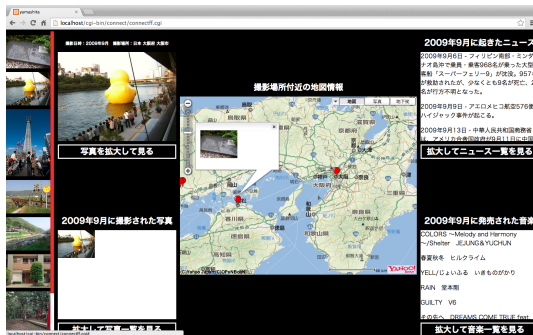


図 2: 地図情報を閲覧しているときの画面

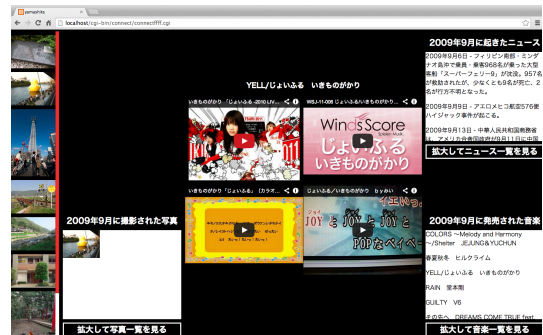


図 4: 音楽を選択したときの画面

本研究では写真以外にユーザの記憶を想起させる情報としてニュース記事と流行した音楽を利用し、それぞれの情報から抽出された共通要素を用いてお互いの情報を結びつけることで横断的な情報アクセスの支援を実現する。ニュース記事は、Wikipedia から 2000 年から 2013 年までの各年の出来事を月ごとにまとめたものをシステムに組み込んだ。ニュース記事をトリガとして、同じ名詞が出現するニュース記事にアクセスできるように、Wikipedia の見出しをユーザ辞書に登録し、固有名詞の抽出の精度を上げた形態素解析エンジンの MeCab² を利用してニュース記事の文章から名詞を抽出し、抽出した名詞が出現するニュース記事を提示するようにした。音楽情報は、2000 年から 2013 年 シングル CD オリコンランキング³ から 100 位以内の楽曲を発売日の月ごとに歌手名と曲名でまとめたものをシステムに組み込んだ。

5 提案システムの実装

本章では、提案した 2 つのバージョンのシステムに実装された機能について述べる。Ver.1 システムは違うモダリティの情報に渡り歩くことに重点を置き、Ver.2

システムは情報が持つ共通要素をユーザに提示することに重点を置いてデザインしている。

5.1 Ver.1 システム

Ver.1 システムの初期画面は、ユーザが撮影した写真の一覧が画面左側に表示される。写真を選択することで図 1 のように画面中央に選択した写真の拡大画像が表示され、写真の撮影日や撮影場所を共通要素に持つ他のモダリティの情報が写真の周囲に表示される。

画面左上には写真の撮影場所を中心とした地図が表示されおり、拡大して閲覧している状態が図 2 になる。地図上に配置されたピンをクリックすることでその場所で撮影された写真を閲覧することができる。

画面左下には選択した写真と同じ撮影日に撮影した写真が表示されており、選択することで図 1 と同じように拡大して表示することができる。

画面右上には選択した写真と同じ撮影日に起きたニュース記事の一覧が表示されており、スクロールすることですべての記事閲覧することができる。ニュース記事の一覧の中から興味を持った記事を選択すると、図 3 のように選択した記事が中心に表示され、それと共通する要素を持つ情報群が周囲に配置される。また、選択したニュース記事から抽出された名詞ごとに同じ

²<http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

³<http://www.oricon.co.jp/rank/d/>



図 5: 写真を選択したときの画面



図 7: 音楽を選択したときの画面



図 6: ニュース記事を選択したときの画面

名詞が出現するニュース記事の一覧を閲覧することができる。

画面右下には選択した写真と同じ撮影日に発売した音楽の一覧が表示されており、選択することで図 4 のように該当する YouTube⁴ の動画が 4 つ表示され、音楽を視聴することができる。

5.2 Ver.2 システム

Ver.1 システムではシステムに組み込んだ写真のデータとして、EXIF 情報を利用したが、EXIF 情報が欠如した写真を扱う場合や、海外で撮影した写真を扱う場合、地図情報を用いて写真を閲覧することが困難になると考えたため、Ver.2 システムでは EXIF 情報を利用せず、ユーザが手動で写真のデータベースに入力する手法を採用した。ニュース記事から抽出された名詞をわかりやすくユーザに提示するため、Yahoo! が提供している日本語形態素解析 API⁵ を用いて地名、人物、その他のキーワードに分類し、ニュースが起きた場所で撮影した写真を提示できるように実装した。また、選択した音楽情報から、同じ歌手が歌っている別の音楽にアクセスできるように、歌手ごとに歌っている

曲を楽天ブックス CD 検索 API⁶ で収集し歌手のデータベースを新たに作成した。

初期画面に表示された写真を選択することで、図 5 のように画面左側に写真が拡大されて表示される。画面右側には、写真の撮影日や撮影場所を共通要素に持つ情報が「撮影日つながり」、「撮影場所つながり」とまとめられて表示される。

ニュース記事を選択すると、図 6 のように画面左側に選択したニュース記事が表示され、画面右側には選択したニュース記事から「地名つながり」、「人物つながり」、「キーワードつながり」や、ニュースが起きた「日付つながり」が共通要素として表示される。

音楽を選択すると、図 7 のように画面左側に該当する YouTube の動画が表示され試聴することができる。画面右側には選択した音楽情報と共通要素を持つ情報が「歌手つながり」、「発売日つながり」として表示される。Ver.2 システムでは、どの画面からでも初期画面の写真の一覧を閲覧できるように、画面左下に初期画面に戻るホームボタンを設置した。

6 実験

6.1 実験の目的

本実験の目的は、提案システムを用いた情報探索におけるユーザの振る舞いを観察することである。そこで、本研究ではユーザの振る舞いを観察するために、(1) 提案した Ver.2 システムを用いた情報探索の推移、(2) 各モダリティの情報を閲覧する回数、(3) どの共通要素を利用して次の情報にアクセスするのか、の 3 点に着目してユーザ観察を行った。本ユーザ観察では、探索者の思考を推定するため、発話思考法 (think aloud) を用いた。実験協力者には、課題として「20 分間、自らが撮影した写真を組み込んだ提案システムを操作する」というタスクを課し、その探索過程を観察した。実験

⁴<https://www.youtube.com/>

⁵<http://developer.yahoo.co.jp/webapi/jlp/ma/v1/parse.html>

⁶<https://webservice.rakuten.co.jp/api/bookscdsearch/>

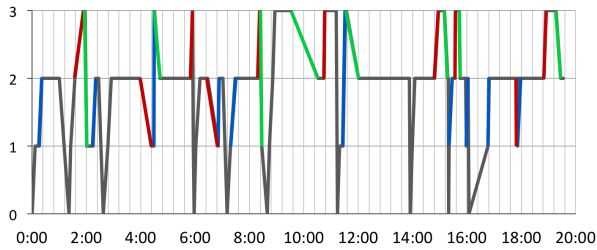


図 8: 提案システムを利用した情報探索の推移

協力者には発話しながら提案システムの操作を行ってもらった。

6.2 実験手続き

実験協力者は情報系の大学に通う学生 6 名（男性 2 名、女性 4 名）である。これらの実験協力者は、事前に自分で撮影した写真を 15 枚、撮影日や撮影場所を記述したデータと併せて提出すること、という条件を設けて募集した。本実験では、発話思考法を用いて行った。課題中は、実験協力者の様子をビデオカメラ、ボイスレコーダ、デスクトップキャプチャで記録した。その間、実験者は実験協力者の行動や発言に関して逐次メモを取り、その後のインタビューに備えた。インタビューでは、実験協力者に対してインタビューを行い、ユーザ観察の全工程を終了した。

6.3 実験結果

6.3.1 提案システムを利用した情報探索の推移

提案システムを利用したユーザの中から実験協力者 F の情報探索の推移を図 8 に示す。このグラフの縦軸の数字はそれぞれ、0 が提案システムの初期画面、1 が写真を閲覧している状態、2 がニュース記事を閲覧している状態、3 が音楽を閲覧している状態を指している。グラフの青い線は写真からニュース記事と音楽にアクセスしている部分を、赤い線はニュース記事から写真と音楽にアクセスしている部分を、緑色の線は音楽から写真とニュース記事にアクセスしている部分を示しており、実験協力者 F は 3 つのモダリティの情報を横断的にアクセスしていることがわかる。そして、探索過程で表示された情報に興味を持てなくなると、ホームボタンを利用して初期画面に戻り、再び撮影した写真をトリガに情報探索を始める様子が全てのユーザから観察された。

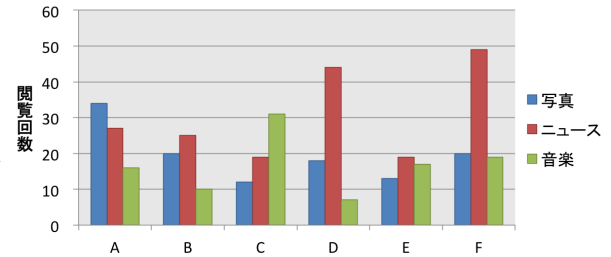


図 9: 各モダリティの情報を閲覧した回数

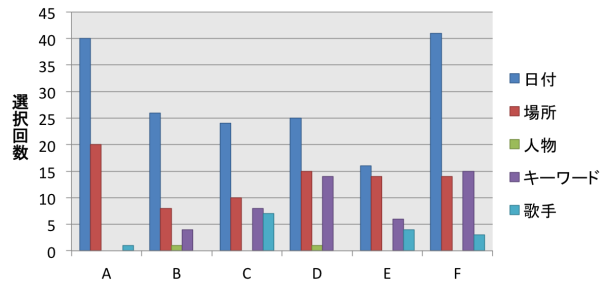


図 10: 各共通要素を利用して情報アクセスした回数

6.3.2 各モダリティの情報を閲覧した回数

提案システムを利用したユーザの各モダリティの情報を閲覧した回数を表したグラフを図 9 に示す。

実験協力者 A は写真を閲覧した回数が他のモダリティを閲覧した回数よりも多かった。このことについてインタビューで質問した結果、実験協力者 A は普段から日本や世界で起きた出来事についてあまり知識がなく、写真を選択した時に表示されたニュース記事について興味を持てなかったことがわかった。

実験協力者 B、D、F はニュース記事を閲覧する回数が他のモダリティの情報を閲覧した回数よりも多かった。

実験協力者 C は音楽を閲覧する回数が他のモダリティの情報を閲覧した回数よりも多かった。このことについてインタビューで質問した結果、実験協力者 C は日頃から音楽を聴くことが好きで、音楽を聴くことによって流行っていたその当時の出来事を思い出すことが多いと発言していた。これらの結果から、ユーザが事前に持っている興味や知識によって、閲覧するモダリティの情報が変化し情報探索の振る舞いが変化することがわかった。

6.3.3 各共通要素を利用して情報にアクセスした回数

提案システムを利用したユーザがどの共通要素を利用して次の情報にアクセスするのかを表したグラフを図 10 に示す。このグラフにより、全てのユーザが「撮影日つながり」、「発売日つながり」、「日付つながり」

といった、日付を共通要素にして次の情報にアクセスする回数が多いことがわかる。次に「撮影場所つながり」や「場所つながり」といった場所を共通要素に次の情報にアクセスする回数が多い。これらの結果は、今回システムに組み込んだ3つのモダリティの情報が時間情報を多く含んでおり、場所に関する情報も写真やニュース記事に含まれることが多いことが要因だと考えられる。

7 提案システムの到達点と課題

はじめに、図8の提案システム利用したユーザの情報探索の推移を示すグラフから、ユーザは何度も写真をトリガに情報探索を始めていることがわかる。このことから、写真をトリガに情報探索を始めるための最低限必要な機能は搭載できたと考えられる。

次に、図8のグラフを見ると、一つのモダリティの情報から他のモダリティの情報にアクセスしている様子が頻繁に観察されていることがわかる。図8のグラフから、全てのユーザが提案システムに組み込んだ3つのモダリティの情報を閲覧していることがわかる。図10のグラフから特に日付によって結びついている情報にアクセスする回数が多いことがわかる。このことから、提案システムでは3つのモダリティの情報を横断的にアクセスする支援を実現できたと考えられるが、日付を共通要素とした情報アクセスに偏ったシステムであることも考えられる。

最後に、Ver.1システムの地図上に撮影した写真を配置する手法をVer.2システムでは廃止したが、これによってEXIF情報を持たない写真もシステムに組み込むことができるようになった。さらに、図10のグラフから日付を共通要素にした情報アクセスに次いで、場所を共通要素にした情報アクセスの回数も多いことから、地図情報を用いなくても撮影場所から次の情報にアクセスすることができるようになっている。

Ver.2システムでは、ニュース記事から抽出した名詞を地名、人物、その他のキーワードに分類する精度が不完全であったため、適切に情報を表示できていない場面もあった。名詞の抽出の精度を向上させることによって、ユーザに適切な情報提示ができると考えている。また、提案システムで用いた音楽について、発売日と歌手を共通要素にしてが、音楽は彼女が聴いていたから自分もその曲を知って聴くようになった場合など、発売した時期とユーザが聴いていた時期とのずれが大きい場合がある。写真やニュース記事に比べ、音楽は時間情報との関係性が小さいため、今後は写真やニュース記事と同列に扱うのではなく、時間情報との関係性を考慮し補足情報として音楽を表示することによって適切な情報提示ができると考えている。

8 おわりに

本研究では、写真をトリガとした横断的な情報アクセスの支援を目指して“興味の移ろい”を誘発させることにより、情報探索を円滑にする手法について検討した。そのために、トリガとなる写真が持つメタ情報に着目し、様々なモダリティの情報に横断的なアクセスができるようにニュース記事や音楽も利用し、これらの情報が持つ共通要素を結びつけることで情報アクセスを円滑にするシステムを実装した。発話思考法を用いた実験を通じて提案したシステムを用いた情報探索時のユーザの振る舞いを観察した。ユーザ観察の結果から、ユーザが事前に持っている興味や知識によって、閲覧するモダリティの情報に差異があることや、日付を共通要素にして次の情報にアクセスすることが多いということがわかった。

謝辞

本研究の遂行にあたり、文部科学省科学研究費(課題番号: 24650040)の助成を受けた。記して謝意を表す。

参考文献

- [1] Marchionini, G.: Exploratory Search: From Finding To Understanding, *Communications of the ACM*, Vol. 49, No. 4, pp. 41–46 (2006)
- [2] Bates, M. J.: The Design of Browsing and Berrypicking Techniques for the Online Search Interface, *Online Information Review*, Vol. 13, No. 5, pp. 407–424 (1989)
- [3] Yee, K. P., Swearingen, K., Li, K. and Hearst, M.: Faceted Metadata for Image Search and Browsing, *In Proc. SIGCHI Conference on Human Factors in Computing Systems*, pp. 401–408 (2003)
- [4] 白鳥 佳奈, 伊藤 貴之, 中村 聡史: PLUM: 地図配置型の写真ブラウザの一手法, 情報処理学会研究報告. *HCI*, ヒューマンコンピュータインタラクション研究会報告, Vol.141, No.12, pp.1–6 (2011)
- [5] 田中和広, 松下光範: Phickle: 写真をトリガとした横断的な情報アクセスを支援するシステム, 第3回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, pp. 63–70 (2013)

利用者の状況に応じた用語解説抽出システムの提案とその実現に向けた検討

Proposal and Consideration for Extraction of Term Explanation Depending on Context

本間 康允^{1*} 渋谷 英潔² 森 辰則²
Kosuke Homma¹ Hideyuki Shibuki² Tatsunori Mori²

¹ 横浜国立大学大学院環境情報学府

¹ Graduate School of Environment and Information Sciences, Yokohama National University

² 横浜国立大学大学院環境情報研究院

² Faculty of Environment and Information Sciences, Yokohama National University

Abstract: In this paper, we propose and discuss extraction of term explanations depending on context. When we read some documents, we encounter some unknown terms. Although there are various explanations of a term, it is important to give users good explanations suitable for their context for understanding the unknown terms. We pay attention to brace expressions which appears frequently, and study those expressions in explanation of them. We also discuss a system that extracts explanations for terms from the Web.

1 はじめに

我々が各種文書を読み進める際、未知の用語に遭遇する機会は少なくない。我々は未知の用語に遭遇したとき、辞書で調べる、Web で調べるといった手段で解決を図る。しかし辞書に存在するのは、未知の用語に対する定義文が主であり、それが常に文書を読み進めるときの助けになるとは限らない。一方で、Web 上には各種用語に対して定義文は勿論、様々な言及が存在し、用語解説として利用出来るものも多い。近年のWebの発展により、我々が未知の用語に遭遇した際、検索エンジンは用語に対する言及を容易に調査することを可能にした。しかし、それらの言及の数は膨大で、様々な文脈で述べられている。更に言及の内容も多種多様で、用語の定義を述べている言及、用語の別名を述べる言及、用語が指し示す事物の具体例を述べる言及などが存在する。それらに対して、利用者は読み進めている文書に適した言及を見つけ出す必要がある。例えば「前駆体」とは三省堂大辞林によると「一連の生化学的反応過程の中で着目したある物質よりも前の段階にあって、一ないし数段階の反応によってその物質に変わりうる物質。」と説明される。しかし、文書中で「ダイオキシンの前駆体」が現れた際に、上記説明では

「前駆体」の定義しか分からない。一方で、Web 上の文書には「ダイオキシンの前駆体」に対して「クロロフェノール、クロロベンゼンなど」「T3CB」といったダイオキシンの限定した「前駆体」への言及が存在する。利用者はその時々に応じて、適した言及を選択する。しかし利用者が保有する知識と読み進めている文書によって、利用者に適した言及は変化するため、用語と利用者の状況に応じた言及は一対一対応で決めることが難しいと考えられる。

そこで本稿では、未知の用語の理解支援を目的として、Web 上に存在する用語に対する言及を用語解説とみなし、それらを抽出し、整理して提示するシステムを検討する。また、その実現に向けての検討を行う。用語解説は、定義を述べる用語解説だけでなく、書き手によって注釈づけられた補足的な用語解説も存在し、それは括弧表現で現れるときもあれば、文書の最後に用語に対する注釈を記述しているときもある。我々は、書き手によって注釈付けられた記述が未知の用語への理解を深めるのに有効であると考えた。本稿では、用語解説が現れる表現の中でも、注釈付けを行う典型的な表現として、括弧表現による注釈記述に注目する。しかし、括弧表現による注釈記述もまた様々であり、その現れ方の調査を行う必要がある。

本稿では想定するシステムの実現のために、まず括弧表現による注釈記述を調査し、その現れ方と用語に対する注釈記述の役割について分析を行う。次に、想

*連絡先： 横浜国立大学大学院環境情報学府
〒 240-8501 神奈川県横浜市保土ヶ谷区常盤台 79-1
E-mail: ammh.k@gmail.com

定するシステムの概略を示し、上記の調査と分析に基づいて、想定するシステムに必要な処理と本稿で実験を行う処理を述べる。その後実験を行い、その結果について考察を行う。

2 関連研究

Web 上から用語解説を抽出する研究は、これまでも多数行われている。

土橋ら [1] の WWW 検索エンジンを利用した用語解説文抽出システムでは、「X とは Y である」といった説明文抽出テンプレートを複数作成し、検索エンジン Google を用いて得られた文書群に対して適用することで、用語説明文の獲得を試みている。特徴的な表現を利用して用語に対する補足を得ようとする点では、本稿で述べるシステムと同じだが、利用者が読み進めている文書には注目していない点に違いがある。

また Web から用語に関する説明情報を収集して事典的コンテンツを生成する研究に、藤井ら [3] の Web マイニングによる事典的コンテンツの構築と多様なアクセス手法がある。藤井らは Web の事典的利用を目的とした検索システム CYCLONE を提案しており、獲得した説明情報を分野で分類することで多義語の説明の区別や、用語の関連語を視覚的に提示することを可能にしている。本稿で述べるシステムと非常に似ているため、CYCLONE については 4.3 節にて詳細に述べることにする。

括弧表現による注釈記述に関する研究としては、中山ら [6] の括弧表現の抽出・分類に関する研究がある。中山らは丸括弧や鉤括弧といった括弧表現の用法をそれぞれ 16 種類と 3 種類に分類し、用語に対する括弧表現の位置や用法ごとの特徴によって自動分類を試みている。

括弧表現内に存在する多様な用法の中でも、言い換え表現に着目した研究には、岡崎ら [7] の言い換え可能な括弧表現の抽出法がある。岡崎らは新聞記事を対象に括弧表現を言い換え可能性の観点から分類を行っており、その中で、括弧表現による注釈記述には言い換えが成立しない関係が存在し、文脈に応じて多様な注釈記述が見られると述べている。本稿では、中山ら [6] 同様に括弧表現の現れ方や用語との関係を言い換えに限らず調査する。また特に括弧表現内の記述に重点を置いて考察を行う。

3 括弧表現の調査・分析

本節では、用語への注釈付けとして典型的な括弧表現に着目し、用語とそれに付随する括弧表現による注釈記述の現れ方について調査を行う。

3.1 調査対象

括弧表現内に現れる内容は多種多様であり、括弧表現の抽出と分類を行った中山ら [6] の研究では頭文字、場所、所属といった 16 種類の用法で丸括弧を分類した。また鉤括弧については会話、強調、題目の 3 つに分類している。本稿で我々が検討するシステムの目的は用語に対する用語解説を抽出して、整理して提示することである。そのため、用法が限られる鉤括弧については触れないこととし、括弧表現の中でも用法の多い丸括弧表現に限定して調査を行うことにする。

中山ら [6] の研究からも、また我々が持つ括弧表現に対する経験的な知見に拠っても、括弧表現の使われ方は様々であることは明らかである。またその使われ方は話題によって異なることが推測され、使われ方それぞれの数も話題によって変わると考えられる。

上記の考察から、特定の話題について記述している Web 文書を話題毎に数個集めた文書群を調査対象とし、そこに現れる括弧表現とそれにより説明が加えられる語や句に対して調査を行う。調査を行った Web 文書を表 1 に示す。以降に現れる文章の例は、特に記載がない限り、これらが引用元である。

表 1: 調査対象文書

国立環境研究所 地球温暖化第一部「地球温暖化とは?」: http://www.nies.go.jp/escience/ondanka/ondanka01/index.html
国立環境研究所 オゾン層の破壊-過去・現在・未来- : http://www.nies.go.jp/escience/ozone/index.html
レアメタルリサイクル技術 - 環境技術解説 環境展望台: http://tenbou.nies.go.jp/science/description/detail.php?id=62
日本年金機構 年金について 年金制度全般: http://www.nenkin.go.jp/n/www/service/index.jsp
厚生労働省年金局 年金財政ホームページ: http://www.mhlw.go.jp/topics/nenkin/zaisei/
あなたの企業年金、お忘れではありませんか?—企業年金連合会: http://www.pfa.or.jp/nenkin/callcenter/
わかりやすい政治課題解説 法律-日本政治.com: http://nihonseiji.com/policy

3.2 括弧表現の現れ方と機能

ここでは具体例を交えて、括弧表現の現れ方と機能进行分析する。調査対象の Web 文書群に存在する括弧表

現の例を表 2 に示す。括弧表現は書き手による注釈

表 2: 括弧表現の例

塩素原子をまったく含まない「 <u>ハイドロフルオロカーボン (HFC)</u> 」で、第 2 世代の代替フロンとよばれています。
これらを燃やすとダイオキシンの生成を助ける塩素が発生します。これが、ダイオキシンの <u>前駆体</u> （クロロフェノール、クロロベンゼンなど）と結びついてダイオキシンが合成されるのです。
横浜金属（株）
これに対し日本自由法曹団は「 <u>(公益及び公の秩序は)</u> 『公益の福祉』とは異なり、抽象的な価値を...

記述であるため、語や句が現れた後に現れることが多い。しかし、前に現れて語や句に対して説明をつける場合も存在する。本稿では括弧表現内の記述と関係を持つ語や句を用語と定義して扱うこととする。また文末に現れて、注釈を加える場合もあるが、本稿ではこれは対象外とし、用語の前後に現れる括弧表現について調査を行う。

次に括弧表現内の記述について詳しく述べる。ハイドロフルオロカーボンの例では、括弧表現内の記述は直前の用語に対しての言い換え表現を与えている。これは中山ら [6] の研究の分類では、外来語の頭文字に当たる用法である。次に前駆体の例を見ると、括弧表現内の記述は、前駆体として考えられる化学物質のうちダイオキシンの前駆体となる化学物質だけに限定して具体的な例を与えている。そのためこの記述は、前駆体が現れる他の文章において、常に前駆体を正しく説明するとは限らない。つまり、この括弧表現内の記述は、それが含まれる文書の文脈に依存していることが分かる。それに対してハイドロフルオロカーボンの括弧表現内の記述である HFC は単なる言い換え表現であるため、こういった文脈でもハイドロフルオロカーボンの言い換えとして機能する。

以上のように、括弧表現内には補足説明となる記述が存在することが多いが、その一方で、括弧表現内の記述には定型表現が入る場合もある。横浜金属の例がそれに当たる。他に（有）や（独）、電話番号などが存在するが、このような定型表現は補足説明ではない。

以上の観察から、括弧表現内の記述について次のことが言える。

1. 括弧表現内の記述には、補足説明とそれ以外が存在する
2. 用語と補足説明の関係は複数の種類が存在する

3. 補足説明の中には文脈に依存したものが存在する

次の小節で、2 について調査対象の文書群に見られた用語と補足説明の関係を述べる。3 については、5 節において詳細に述べる。

3.3 用語と括弧表現内の記述の関係

調査対象の文書群を観察した結果、用語と括弧表現内の記述が持つ関係は少なくとも次のように分類された。そのうち補足説明になるものは以下の通りである。

1. 同一

補足説明が用語に対して同一の事物を指す関係を持つことを、本稿では同一と呼ぶことにする。塩素原子の例では括弧表現内の補足説明には

フロンは安定な物質で対流圏では分解されません。しかし、成層圏にまで達すると、強い紫外線によって分解され、塩素原子（Cl）を放出します。

一般永住者（一定の要件を満たして永住許可申請をし、許可され、日本国に永住している外国人のこと）には参政権が与えられていない。

言い換え表現が与えられ、一般永住者の例では括弧表現内の補足説明には用語の定義が与えられている。

2. 属性値

本稿では、補足説明が用語のある属性に値を与えているような関係を属性値と呼ぶことにする。岡崎ら [7] の研究においても、これらの括弧表現に対して同様の分析を行っており、これらを属性と呼んでいる。豊羽鉱山の例では、括弧表現内

レアメタルは、他の鉱物の副産物として産出されることが多く、例えば世界最大のインジウム鉱山であった日本の豊羽鉱山（北海道）...

Ta コンデンサ（携帯電話、デジタルカメラ、パソコン）

の補足説明には用語の存在する場所を示す北海道が与えられている。また、Ta コンデンサの例では括弧表現内の補足説明は Ta コンデンサの使用先を与えている。これらは場所や用途といった、用語が持つ属性に対して、値を与えていると考えることが出来る。

3. 例示

以下に示す二つの例では、どちらも括弧表現内の補足説明は用語に対して具体例を与えている。つまり用語によって定まる部分集合に対して、補足説明はその集合の要素または下位概念となる部分集合を列挙する関係を持っていることが分かる。本稿ではこの関係を例示と呼ぶことにする。

非化石燃料や新エネルギー技術（たとえば、太陽エネルギーやバイオマスエネルギー）を利用する社会...

オゾン層破壊と関連のある物質（硝酸、二酸化窒素、亜酸化窒素、硝酸塩素、メタン、水蒸気、エアロゾルなど）の濃度が測定されました。

4. 限定

以下の例では、括弧表現内の補足説明は、用語によって定まる集合の中でも特に文章中に合致する集合内の要素や下位概念となる部分集合を与える関係になっている。例示と似ているが、こちらは文脈によって一つに定まっており、列挙する関係を持っていない。本稿ではこのような関係を限定と呼ぶことにする。

噴射剤については特殊な用途を除いて炭化水素（液化天然ガス：LPG）

回転の状態が変わるとそのエネルギーの差を電波（ミリ波）として放出します。

5. その他

以下の例では、補足説明は用語に対して付加的な説明、非制限的な説明を与えている。どちらも前述したような同一、属性値、例示、限定のうち、どれにも当てはまらない。このような非制限的な説明を与える補足説明の関係をその他と本稿では呼ぶことにする。

表1に示す31鉱種（ただし、レアアース（希土類）は17鉱種を総括して1鉱種とする）をレアメタルと規定しています。

こうした緊張関係の中、権利の保護（＝新たな創造への投資のインセンティブ）と利用の促進（＝創造の果実の社会による享受や再創造へ

補足説明以外では次のような記述が見られた。

1. 定型表現

以下の例は、日常生活において広く用いられる括弧表現である。括弧表現内の記述が用語に対して持つ関係は、会社の種類という属性に対して値を与えているため属性値が適していると考えられる。しかし広く用いられる表現であるため特に区別して、定型表現と本稿では呼び、補足説明には含めない。

（独）石油天然ガス・金属鉱物資源機構希少金属備蓄部

2. その他

括弧表現内の記述が補足説明ではなく、定型表現でもない場合その他と呼ぶことにした。以下の例では、用語に相当する部分が曖昧なため下線は付けない。

朝日新聞 DIGITAL「（ネット選挙がわからん！）政党や候補者の戦術、どう変わるんじゃ」2013.04.23.

上記の補足説明が用語に対して持つ関係の分析に基づいて、調査対象の文書群の中に現れる括弧表現内の記述について人手で分類を行った。その結果を表3に示す。なお「対象外」は3.2節で述べたように、用語の前後に現れない場合である。

表 3: 現れ方の分類

補足説明	同一	168
	属性値	60
	例示	15
	限定	32
	その他	13
補足説明以外	定型表現	19
	その他	3
対象外		50
総数		360

この結果から、括弧表現内の補足説明の中でも同一が非常に多く存在することが分かる。これは同一と分類する基準が、用語が指している事物と同じであるとしたため、用語の頭文字の省略や言い換え表現、定義文を含めていることが原因だと考えられる。従って、同一に関しては更に細分化した分類を行う余地があると思われる。また括弧表現内の補足説明は括弧表現全体の80%となり、これは括弧表現内に補足説明が十分に存在することを示している。このことは括弧表現内から様々な補足説明が抽出できることの裏付けになっていると言える。

4 目標とする用語解説抽出システム

3節では括弧表現について調査と分析を行った。その中で補足説明と分類される記述を、本稿では用語解説と呼び、これを整理して提示するシステムを考える。

本節では、本稿の目標とするシステムの概要とその構成について述べる。4.1節では目標とするシステムのタスク設定を行い、4.2節では目標とするシステムの実装上の課題について述べる。

4.1 目標とするシステムとそのタスク設定

本稿で目標とするシステムは、利用者が各種文書を読み進めている際に、未知の用語に遭遇する状況を想定する。目標とするシステムの概観図を図1に示す。

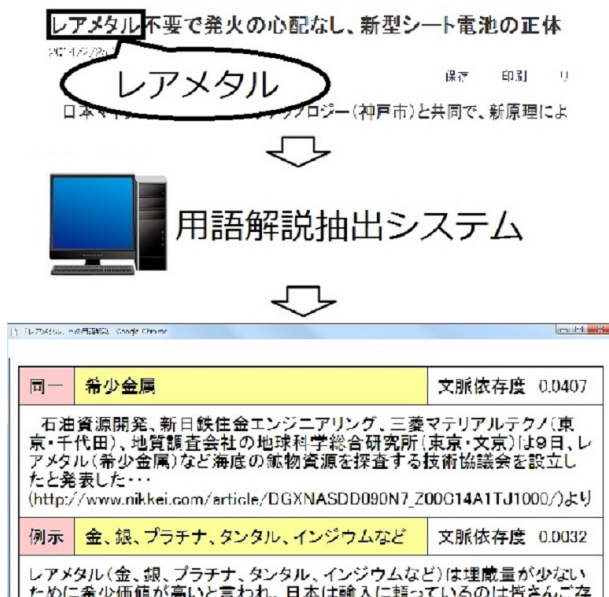


図1: 想定されるシステムの入出力

目標とするシステムへの入力には次の二つである。

1. 利用者が読み進めている文書
2. その中に現れる利用者が調べたい文字列

ここで文字列は、利用者の指示によって与えられるが、本稿で呼ぶ用語に相当することが期待されるものとする。これらの入力を受けた後、同システムはWeb上から文書群を取得し、そこに存在する用語解説それぞれに処理を行い、図1に例示するような結果を出力として返す。各結果に表示される情報を表4に示す。

4.2 実装上の課題

目標とするシステムを実装する上での課題について述べる。同システムは利用者から用語（以下、入力用

表4: 出力

他文書に現れる 当該用語解説	用語と用語解説の関係 用語解説が持つ文脈依存度
-------------------	----------------------------

語と呼ぶ)と、利用者が読み進めている文書(以下、入力文書と呼ぶ)を入力として受け取る。その後、同システムは入力用語と入力文書の情報に基づいてクエリを作成する。それを元にテキスト集合を検索し、文書群を取得する。得られた文書群から用語解説を抽出し、用語解説が持つ役割や文脈依存度で整理を行う。そして整理された結果を利用者に提示する。同システムの実現に必要な処理として以下のものが挙げられる。

1. 用語解説が現れる文書の効率的な獲得
2. 用語解説が持つ役割の自動分類
3. 用語解説の文脈依存度による順位付け

1については現在検討中だが、本稿で行う実験では検索エンジン Google を用いて Web 上に存在する文書群を獲得することとした。2については、3節にて用語解説が持つ役割の調査と分析を行った。しかし自動分類のためには、別途特徴表現などの調査が必要である。3は、同システムが持つ特徴的な処理であり、新規性となりうる部分である。そのため本稿では、3について5節にて順位付け手法の検討を行う。

4.3 CYCLONE との比較

用語への用語解説を集めて提示する点では、本稿で目標とする用語解説抽出システムは藤井ら [3] の CYCLONE によく似ていると言える。CYCLONE は Web から見出し語とその説明文章や関連語を収集し、事典コンテンツを自動構築し、更にコンテンツに対する多様なアクセスを可能にしている。本稿で目標とするシステムは、用語解説を抽出するに当たって、用語へ注釈を与える典型的な表現である括弧表現に注目することで、書き手による咀嚼がされた用語解説も抽出することを試みる。また用語解説が持つ役割をシステム側で分類しておくことで、利用者が求める用語解説を見つけ出しやすいようにしている。この点では同システムと CYCLONE が目指すことは同じであるため、今後参考にし検討する。また、CYCLONE では予め学習した分野モデルや大百科事典を基とした言語モデルなどを総合して、獲得した説明を分野に分類することで多義語の説明を区別する。同システムでは、用語解説が持つ文脈依存度を見ることで文脈依存性の問題に対応するとともに間接的に多義性の解消を試みる。これは用語が述べられる分野に特有の文脈が存在し、多数

の意味を持つ用語はその文脈によって意味が決まると考えたためである。

5 文脈依存度に応じた順位付け手法

目標とするシステムの一部の処理として、本節では文脈依存度に応じた順位付けを行う手法について述べる。

5.1 用語解説の文脈依存度

用語が現れるまでの文脈を明らかにし、用語解説の文脈依存度を判断するのであれば、係り受け解析や意味解析を行うべきだが、本稿では単純に文書内に現れる名詞群を文書の文脈を表現するものとして扱うこととする。

前述の通り、用語解説の中には文脈に依存するものも依存しないものも存在する。文脈依存度は候補の選別を行うための指標であり、解説候補間での相対的な大小が分かれば十分である。

用語解説の文脈依存度の算出手法について、我々は二つの手法を考えた。以下では、その二つの手法について述べる。ただし、文脈依存度の値は用語解説候補の順位付けに使用するので、システムの出力においては文脈依存度の具体的な値よりも候補間の順位付けが重要である。後の評価でもその観点で行っている。

5.2 文脈依存度算出手法

5.2.1 手法 1

用語解説が抽出された文書の文脈が、入力文書の文脈にどれだけ近いかを測る。具体的には、入力文書と、用語解説が抽出された文書の双方を名詞の出現頻度を値とするベクトルとし、それらの $\cos$ 類似度を取り、文脈依存度として順位付けを行う。これを手法 1 と呼び、ベースラインとする。

5.2.2 手法 2

手法 1 では、用語解説が抽出された文書の文脈が、入力文書の文脈に近いほど文脈依存度が高くなる。しかし、入力文書に類似する文書において偶然、一般的で他の文脈でも通用する用語解説が現れる状況では、その用語解説に意図しない高い文脈依存度を与えてしまう。これを回避するには用語解説がどのような文脈に現れるかを調べる必要がある。

用語 X について用語解説 Y 、つまり「 $X(Y)$ 」という表現が現れる文書群の文脈を考えると、典型的には次のような場合が考えられる。

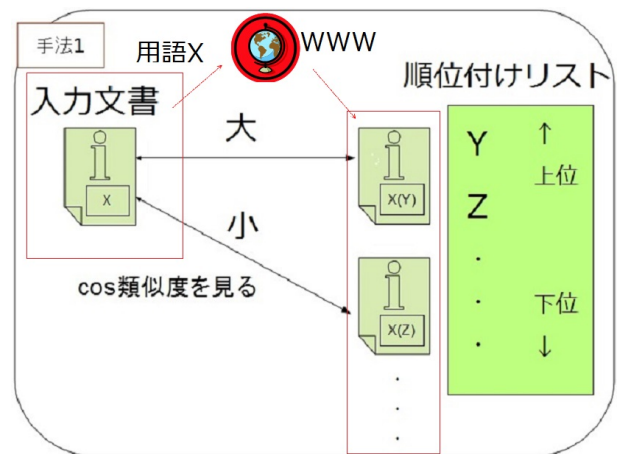


図 2: 手法 1

1. $X(Y)$ が現れる文書群の文脈は多様
2. $X(Y)$ が現れる文書群の文脈に偏りがあり、内容が入力文書に類似する

1 の時、用語解説 Y は文脈に依存しない用語解説と言える。様々な文書でそれぞれ違う文脈の下で現れるため、用語解説 Y は用語 X に対しての一般的な注釈付けと考えられる。2 の時、用語解説 Y は文脈に依存した用語解説と言える。特定の文脈に多く現れる用語解説 Y は、その文脈に依存した注釈付けと考えられるためである。

上記のことから、用語と用語解説の表現 $X(Y)$ が現れる文書群の文脈と入力文書の文脈との類似度を見ることで当該用語解説が持つ文脈依存度を測ることができると我々は考えた。すなわち Y が一般的な用語解説であった場合それは多様な文脈の文書群に現れるので、上記類似度が低くなることを期待する。具体的には、検索エンジンを用いて用語と用語解説の表現 $X(Y)$ について改めて検索し、その表現が現れる文書群を獲得した後、それらを一つの文書と考え、名詞の頻度を値とする文書ベクトルを作る。このベクトルと、入力文書に対応する文書ベクトルの間の $\cos$ 類似度を計算し、用語解説の文脈依存度とする。

6 評価実験

本節では、5 節にて述べた用語解説の文脈依存度の計算手法について評価実験を行う。6.2 節で用語解説を抽出するための文書群の獲得について、6.3 節で出力例と出力に対する評価結果を述べる。

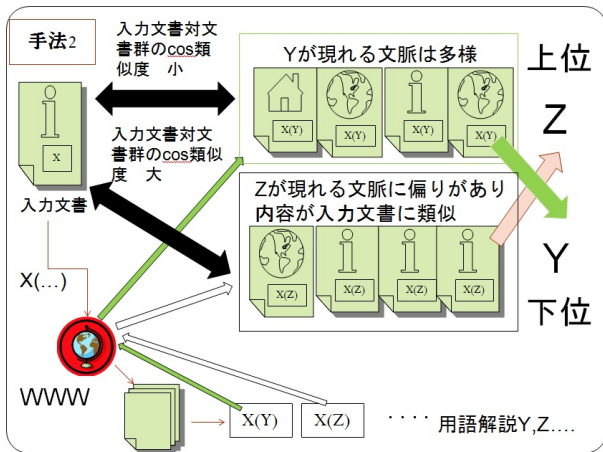


図 3: 手法 2

6.1 入力文書と対象とする用語

本実験で扱う任意に選択した 5 個の用語と、それが含まれる入力文書を表 5 に示す。オペやプレートは多

表 5: 入力文書と対象用語

対象用語	入力文書
オペ	http://www.bloomberg.co.jp/news/123-MHW5NK1A74E901.html
プレート	http://www.jishin.go.jp/main/yosokuchizu/kanto/p13_1_tokyo.htm
前駆体	http://www.kcn.ne.jp/~azuma/QA/IAQ/I026.htm
酵素	http://zeroone.searchnavi.jp/privacy2.html
微生物	http://www.kao.co.jp/pro/noro/how/p1.html

様な文脈で現れる用語である。前駆体、酵素、微生物は現れる文脈が限定されている。オペやプレートの場合に手法 2 で改善されることを期待する。

6.2 文書群の獲得と括弧表現内記述の抽出

ここでは、括弧表現が存在する文書群の獲得について述べる。本実験では、文書群を検索エンジン Google を用いて、Web 上から獲得する。想定されるシステムは、入力用語と入力文書を受け取る。しかし単純に入力用語が存在する文書を検索して得られる文書は非常に多く、その全てから抽出を行うのは現実的ではない。そこで抽出を行う文書の文脈をある程度制限することとした。まず、入力文書の単語頻度で降順のランキングを作成し、その中で上位 2 単語を、検索条件として用語に加えた。つまり、入力用語が X で、入力文書内の単語頻度のランキング上位 2 単語が A,B であった場合、検索エンジンに与えるクエリは「X A B」である。

上記の手順によって獲得した文書群から、括弧表現内記述を抽出する。ここでは、「X (Y)」と記述されている括弧表現のみを対象に抽出を行った。また補足説明以外もそのまま手法 1,2 を適用する。

6.3 実験結果

出力の具体例を表 6,7 に示す。

表 6: 手法 1

入力文章 (一部)	用語解説
期待される効果は、マネタリーベースの供給拡大／加速によるデフレ期待の後退＝リフレ期待の台頭。心配される副作用は、国債バブルの助長、財政ファイナンス懸念による「悪い金利上昇」など。当座預金付利が撤廃されると、オペに応じる金融機関が減り、基金残高の積み上げが困難になる。	2月1日－8日 こていきんりオペ 6月14日-22日 貸付利率 0.1 % 6月15日-22日 8日－2010年3月4日 固定金利方式 新型オペ 短期国債の購入 2010年3月31日をもって完了 オペレーション 公開市場操作

表 7: 手法 2

入力文章 (一部)	用語解説
期待される効果は、マネタリーベースの供給拡大／加速によるデフレ期待の後退＝リフレ期待の台頭。心配される副作用は、国債バブルの助長、財政ファイナンス懸念による「悪い金利上昇」など。当座預金付利が撤廃されると、オペに応じる金融機関が減り、基金残高の積み上げが困難になる。	公開市場操作 オペレーション 短期国債の購入 2月1日－8日 固定金利方式 8日－2010年3月4日 2010年3月31日をもって完了 新型オペ こていきんりオペ 6月15日-22日 6月14日-22日 貸付利率 0.1 %

出力について、人手で評価を行った。入力文書の文脈に適合した用語解説が上位に存在することが重要である。著者以外の評価者 A,B の 2 名が別々に、手法 1,2 それぞれの出力について用語解説候補が、入力文書の文脈に適合しており用語への補足説明となると判断したなら○を、入力文書の文脈に適合してないが用語への補足説明となると判断したならば△を、用語への補足説明にもならないと判断したならば×を、抽出された用語解説候補それぞれにつけてもらうこととした。ただし各手法が候補に付与した文脈依存度の高い順に左から右にむけて候補の評価を並べている。左側に○が偏り、右側に×が偏るときの手法の出力が望ましい。

表 8,9 で、各用語の用語解説候補に対する評価結果を示す。

表 8: 評価者 A

用語		←上位	下位→
前駆体	手法 1	×	○××××
	手法 2	×	×
オペ	手法 1	×	×
	手法 2	○△○×	×
プレート	手法 1	×	×
	手法 2	○△○×	×
酵素	手法 1	○△×	×
	手法 2	△×	×
微生物	手法 1	×	×
	手法 2	×	×

表 9: 評価者 B

用語		←上位	下位→
前駆体	手法 1	×	×
	手法 2	△×	×
オペ	手法 1	×	×
	手法 2	△×	×
プレート	手法 1	×	×
	手法 2	○×	×
酵素	手法 1	○△△×	×
	手法 2	△△△×	×
微生物	手法 1	×	×
	手法 2	×	×

7 考察

本節では 6 節で行った実験の結果について考察を行う。手法 1 に対する評価では、オペは下位に○がつくが、それ以外の用語について上位 5 件に一つは○が存在することが分かる。手法 2 ではオペも含めて、どの用語についても○が上位 5 件に一つは存在するようになっている。そのため手法 2 は、少なくとも今回選択した用語に対して、文脈依存度に応じた順位付け手法として安定した働きを見せている。

また△の推移を見ると、評価者 A でのオペ、評価者 B でのプレートといった多様な文脈で現れうる用語の場合は、手法 2 に期待するような文脈依存度の低い用語解説を下位に置く効果が現れている。これは 6.1 節で述べた期待通りだが、多様な文脈で現れうる用語に対してさらなる検証が必要である。

逆に手法 1 では、現れうる文脈が限定されるような用語「微生物」「前駆体」「酵素」について手法 2 より上位に○が来ていた。このことから、入力用語が現れる文脈についてトピック分析を行い、各手法と組み合わせなどの手法の拡張が考えられる。

8 おわりに

本稿では、用語解説が現れる中でも、書き手による注釈付けが行われる典型的な表現として括弧表現に注目し、括弧表現内の記述の現れ方を調査した。それを受けて、利用者の状況に応じた用語解説抽出システムの提案を行った。また、その実現に向けて提案システムに必要な処理を整理し、用語解説に存在する文脈依存度について実験を行った。

その結果、手法 2 が今回対象とした用語のどれに対しても、上位 5 件のうちの一つは入力文書の文脈に依存した用語解説を含んだ出力を得ることが出来た。対象とする用語を更に増やして評価実験を行い、順位付け手法を検討すること、想定されるシステムの他の処理について検討を行うことが今後の課題である。

参考文献

- [1] 土橋 惇一, 荒木 健治: WWW 検索エンジンを利用した用語解説文抽出システム, 情報処理北海道シンポジウム, pp.26-27, 2004
- [2] 土橋 惇一, 荒木 健治: WWW 上の定義文における表現特徴を利用した用語説明文抽出のためのテンプレートの自動生成について, 言語処理学会第 11 回年次大会発表論文集, pp.791-794, 2005
- [3] 藤井 敦, 伊藤 克亘, 石川 徹也: Web マイニングによる事典的コンテンツの構築と多様なアクセス手法, 電子情報通信学会技術研究報告, Vol.4, pp.31-36, 2004
- [4] 藤井 敦, 伊藤 克亘, 石川 徹也: WWW は百科事典として使えるか?, 情報処理学会研究報告 自然言語処理研究会報告, 2002-NL-149, pp.7-14, 2002
- [5] 藤井 敦, 石川 徹也: World Wide Web を用いた事典知識情報の抽出と組織化, 電子情報通信学会論文誌, Vol.J85-D-, No.2, pp.300-307, 2002
- [6] 中山 悟, 森田 和宏, 泓田 正雄, 青江 順一: 括弧表現の抽出・分類に関する研究, 言語処理学会第 16 回年次大会発表論文集, pp.379-382, 2010
- [7] 岡崎 直観, 石塚 満: 言い換え可能な括弧表現の抽出法, 言語処理学会第 13 回年次大会発表論文集, pp.911-14, 2007
- [8] 岡崎 直観, 石塚 満: 日本語新聞記事からの略語抽出: 人口知能学会全国大会論文集, Vol.21, pp.2G4-4, 2007

EC サイトにおける商品探索プロセスに着目した プレゼント探索支援システムの提案

Proposal of a System to Support People Who Look for a Present Using E-commerce Websites

盆子原 健太¹ 大塚 直也² 松下 光範^{1*}
Kenta Bonkohara¹ Naoya Otsuka² Mitsunori Matsushita¹

¹ 関西大学 総合情報学部 ² 関西大学大学院総合情報学研究科
¹ Faculty of Informatics, Kansai University ² Graduate School of Informatics, Kansai University

Abstract: The goal of our research is to support people who look for a present using E-commerce websites. When choosing presents, we need to consider his or her tastes and interests and explore many candidates with a trial and error manner to pick the appropriate one. This paper proposes a system that shows a connection between a variety of genres to facilitate such exploration. We conducted an experiment to confirm performances of the system. As a result, we found that exploration time and the number of explored items increased. The participants repeated a trial and error with a wider range of goods in comparison to a conventional system.

1 はじめに

近年、インターネット上で商品やサービスの販売を行う EC(Electronic Commerce: 電子商取引) サイトが急激に増加しており、時間や場所に関係なく商品が購入できるようになった。こうした EC サイトの普及に伴う探索可能な商品数の急増により、ユーザの選択の幅が大幅に増加している。これは、選択肢の増加をもたらす一方で、ユーザが欲しい商品を効率的に探索することが困難な状況を生み出す原因にもなっている。この問題を解決するために、EC サイトでは膨大な商品の中から、ユーザが欲しい商品へのアクセス支援を目的とした、商品推薦サービスが提供されている。

商品推薦サービスとは、ユーザの興味や嗜好に合った商品を推薦するものである [4]。例えば楽天市場¹では「この商品を買った人はこの商品にも興味があります」といった形式で商品推薦が行われている。

従来の商品推薦サービスは、主に探索者のユーザプロフィール(個人の情報や記録)に基づいており、自身のための購買の支援を企図している。他方で、商品購買では他者に贈るプレゼントを購入するような場面も想定される。従来の商品推薦サービスでは、相手のユーザプロフィールを作成することが困難であり、他者のための購買の支援は考慮していない。そのため、他者

に贈るプレゼントを購入する場合、商品推薦サービスに頼るのではなく、自身で相手の趣味や嗜好を考慮し、相手に喜んでもらえると思う商品を見つけ出す必要がある。また、相手の趣味や趣向が明確でない場合、それらを推測する必要も生じる。これらの原因により、他者へのプレゼントを目的とした購買は、自身のための購買に比べ商品探索が困難であると言える。

本研究では、このような状況を「大量の購買候補集合から、未知の制約条件を推定しつつ適合解を見つける意思決定」と捉える [5]。このような意思決定の場面では、試行錯誤を繰り返し、相手に喜んでもらえると思う商品を、自身が納得して見つけ出す必要がある。そこで、このような問題に対し、本研究では (1) ユーザがシステムとのインタラクションを通して相手の趣味や嗜好を考慮した試行錯誤を繰り返しながら商品探索を行う行為を対象とする、(2) プレゼントを贈る相手が喜ぶプレゼントを見つけることのみならず、それを見つけて出す探索プロセスに探索者自身が満足し納得を得ることも考慮する、という方針で臨む。

2 関連研究

2.1 探索的検索に関する研究

従来の参照型(Lookup)の情報検索モデルは、ユーザが自身の情報要求に基づいてクエリを作成し、シス

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

¹<http://www.rakuten.co.jp/>

テムがクエリとマッチングした文書を返すという単発のやり取りを扱っている [1]。

しかし、ユーザの情報要求は常に明確であるとは限らないため、何度も繰り返し検索を行わなければならない場合もある。このような情報検索モデルは、Exploratory Search と呼ばれている [1]。Exploratory Search を行うユーザの情報要求は曖昧であり、何度も繰り返し検索を行う中で、ユーザは検索対象に関する知識や理解を深めていく。他者に贈るプレゼントを購入する状況は、相手の趣味や嗜好を考慮しつつ試行錯誤を繰り返す Exploratory Search として捉えることができる。

2.2 商品購買の思考過程に関する研究

庄司らは、購買の思考の過程に問題解決型とコンセプト精緻化型の 2 種類があると述べている [3]。問題解決型の購買とは、顧客の欲しい商品に対するイメージや必要だと考える機能が明確であり、何を買うべきか最初から決まっているような購買を指す。これは、自分の要求を満たす商品を探す思考の過程である。一方、コンセプト精緻化型の購買とは、顧客の欲しい商品についての要求が曖昧で、何を買うかを考えながら決めるような購買を指す。これは、店員との対話などを通じて徐々にイメージを作り上げていく思考の過程である。庄司らは、後者のコンセプト精緻化型の購買を支援するためには、気付き (conception) の支援と、理解と納得 (conviction) の支援が有効であると述べている [3]。他者に贈るプレゼントの購入は、相手の趣味や嗜好を考慮しつつ、相手に喜んでもらえると思う商品を見つけ出す必要があることから、コンセプト精緻化型の購買であると捉えることができる。

3 デザイン指針

3.1 EC サイトの構造

EC サイトでは商品ジャンルを設定することで膨大な商品が整理されている。具体的な商品や道具の名称を表す下位ジャンル (e.g., G-Shock、炊飯器) は、上位ジャンル (e.g., ファッション、エンタメ・デジタル家電) の意味や内容を受け継ぎ、それを次の階層につなげていくようなツリー構造を成している。自身のための購買の場合、このようなツリー構造の枝分かれしたノードを、自身の条件 (趣味・価格・機能など) に合わせて絞り込みながら選択していくことで、欲しい商品を効率的に探し出すことが可能である。

しかし、他者に贈るプレゼントを購入する場合、相手の趣味や嗜好を考慮した探索を行う必要があるため、上記で述べたような条件を絞り込んで効率的に探索す

ることは困難であると考えられる。そこで、相手の趣味や嗜好といった曖昧な条件から、相手に喜んでもらえると思う最適な商品を見つけ出すために、試行錯誤を繰り返しながら自身が納得いく商品探索を行う必要がある。

図 1 に示すように、従来のツリー構造では、上位ジャンルから下位ジャンルへ階層を辿り商品を閲覧した結果、相手の趣味や嗜好に合っていないと考えた場合、上位ジャンルへ戻って他のジャンルの探索を行う必要がある。他者に贈るプレゼントを購入する場合、相手に喜んでもらえる商品を見つけ出すために、様々な商品を何度も閲覧し相手の趣味や嗜好を考慮した探索を行う必要がある。しかし、ユーザは上記で述べたツリー構造に依存した探索を行うため、柔軟な探索を行うことができない。

3.2 ツリー構造の横断的探索

既存の EC サイトのツリー構造において、下位ジャンルは 1 つの上位ジャンルしか持たないため、ユーザはそのジャンルから階層を辿り商品探索を行うことになる。

本研究では、ツリー構造をユーザが円滑に探索できるようにすることで、他者に贈るプレゼントを購入する商品探索の支援を行う。これを実現するために、下位ジャンルに複数の上位ジャンルを持たせることで、ユーザが多様な観点を持って商品探索を行えるようにする。

例えば、楽天市場では腕時計はファッションというジャンルの下位に存在している。しかし、腕時計はファッション以外に、ビジネス用品としても捉えることができる。また、ゴルフ用品は、車・スポーツというジャンルにあるが、腕時計と同様にビジネス用品としても捉えることができる。ビジネスという使用場面を想起させる情報 (本研究では、使用用途や使用場面を想起させる情報を場面情報と定義する) を上位ジャンルとして追加することで、商品探索のトリガになると期待される。

庄司らは、コンセプト精緻化型の購買をする顧客に対して、商品の使用場面や詳細を提示することが有効であると結論づけている [3]。そこで、ファッションジャンルにある腕時計と、車・スポーツジャンルにあるゴルフ用品にビジネスという場面情報を意図したジャンルを持たせることで、ユーザはビジネスという観点を持って商品探索を行うことができ、ツリー構造を横断的に探索することで、商品探索の試行錯誤を円滑に行うことができると考える (図 2 参照)。

3.3 提案するシステムのインタラクション

以下に提案システムを用いることで実現が期待されるインタラクションの例を示す。

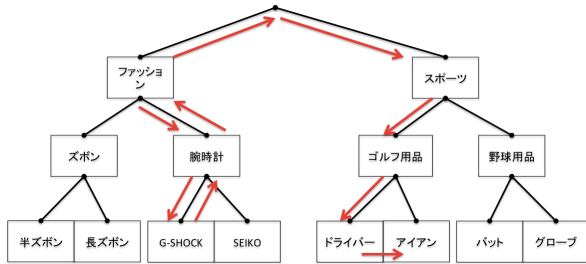


図 1: ツリー構造における試行錯誤

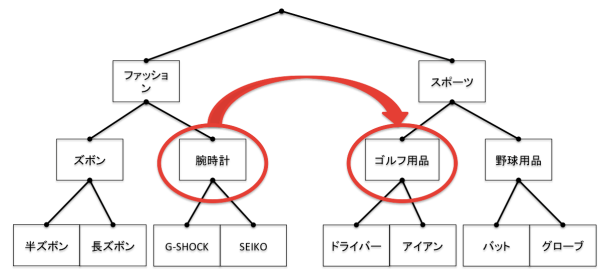


図 2: ツリー構造の横断

大学生 A は父に贈るプレゼントとして、おしゃれな腕時計を探索している。

(1) 腕時計はファッションに含まれるジャンルに含まれておりと考えていたが、ふと腕時計は「ビジネス用品」として考えることができると気が付き、ビジネス用品をプレゼントにしようと考えた。

(2) ビジネス用品を考えた際に、父が以前、取引先の企業とゴルフへ行っていたことを思い出し、ゴルフ用品の探索を行う。

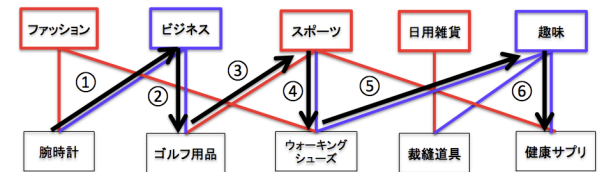
(3) しかし、ドライバーやアイアンはプレゼントには高すぎると考える。

(4) そこで、ゴルフ用品を諦めて、父に仕事の気分転換として少しは体を動かしてほしいと思い、スポーツ用品のウォーキングシューズへの探索に切り替えた。しかし、父は最近仕事が忙しくウォーキングする時間がないので、今回はウォーキングシューズはプレゼントとしてふさわしくないと考えた。

(5) そこで、仕事で忙しく運動不足で太り気味の父に、何か趣味はないかと考え探索を行う。

(6) しかし、父には趣味がないと考え、近年健康食品ブームであり、テレビで女性が趣味で健康食品を集める人が多数いることを思い出し、父にも健康サブリをあげて健康で長生きして欲しいという考えに至る。

上位ジャンル



下位ジャンル

図 3: 提案システムのインタラクション例

提案システムでは、楽天市場（以下、楽天と記す）で使用されているジャンルと商品を使用した。使用するデータは、楽天のジャンル一覧から上位ジャンル 6 種類と、下位ジャンル 37 種類とした。また独自に上位ジャンル 18 種類（e.g., ビジネス、父の日）を設定した。

図 4 に、提案システム起動時の初期画面を示す。赤色の円（ノード）は楽天の下位ジャンルを示している。図 5 に、その中から腕時計のノードを選択した時の遷移例を示す。図 4 で選択したノードが中央に移動し（図 5-A）、システム画面の下部に商品が 5 種類表示される。また、赤いノードの周囲にある図 5-B、図 5-C は、楽天の上位ジャンル及び独自で設定した上位ジャンルであり、それら 2 種類のジャンルが混在して表示される。また、図 5-D に示すように、商品探索過程で興味を持った商品をストックできる機能を設けている。システム画面の下部に表示された商品を選択すると、画面右側に商品がストックされ、見比べながら商品探索を行うことができる。

5 実験

5.1 実験の目的

提案システムに実装された多様なジャンル間の繋がりがユーザの商品探索行為に与える効果を観察することを目的とする実験を行った。この実験では、既存の EC サイトのようなツリー構造の従来型システムとの

4 実装

本章では、3 章で述べたデザイン指針に基づいて実装したシステムについて述べる。

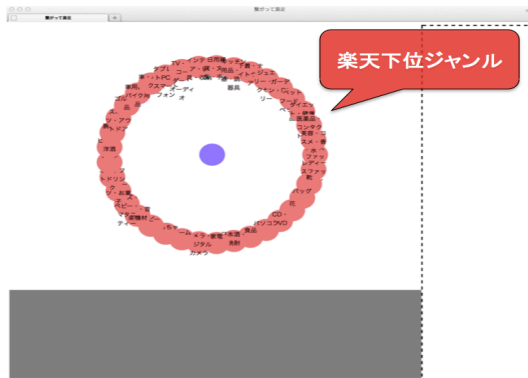


図 4: 提案システムの初期画面

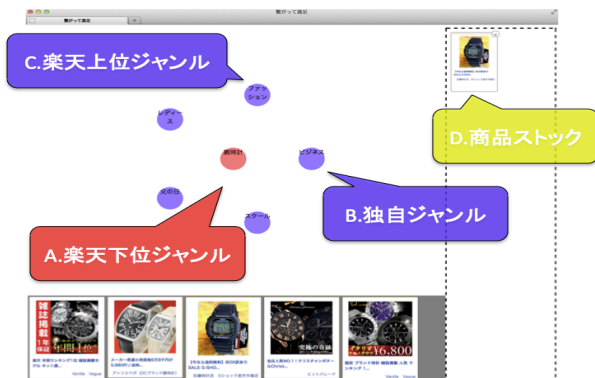


図 5: 提案システムのスクリーンショット

比較検証を行う。実験で用いる従来型システムは、実在する EC サイトではなく、機能（ツリー構造）に焦点を当てて、実験用に実装したものである（図 6 参照）。使用するデータは、楽天市場のジャンル一覧（以下、楽天ジャンル）から上位ジャンル 6 種類と、下位ジャンル 37 種類とした。従来型システムでは、画面左側にジャンルを表示するメニューを設置した。青色の文字で表示しているジャンルは楽天の上位ジャンルであり、その下に黒色の文字で表示しているジャンルはそのジャンルに属している下位ジャンルである。下位ジャンルを選択すると、画面上部に商品が表示される。図 6 は上位ジャンル「ファッション」の中にある下位ジャンル「腕時計」を選択した場合のシステムの画面である。

5.2 実験の概要

従来型システムを使った実験は、情報系学部の大学生 5 名（男性 3 名、女性 2 名）を参加者とした。また、提案システムを使った実験は、従来型システムを使った実験に参加していない情報系学部の大学生 5 名（男性 3 名、女性 2 名）を参加者とした。

いずれの実験においても、課題として「身近な人のた

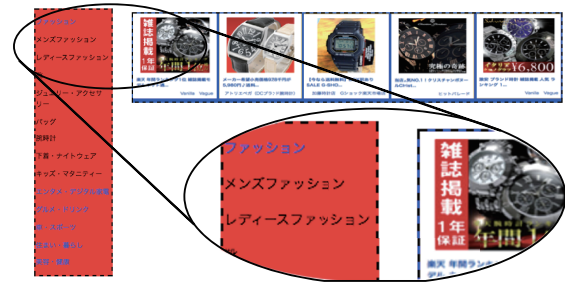


図 6: 従来型システムのスクリーンショット

めにプレゼントを購入する」というタスクを参加者に課した。実験の条件は、(1) プレゼントを贈る相手は途中で変更しても良い、(2) 実験の制限時間は設定せず、参加者が納得のいくプレゼントを選択するまで、とした。まず初めに、システムの操作説明を行い、課題中にシステムの操作で不明な点が合った場合、手を上げて知らせるよう参加者に伝えた。実験は発話思考法 [2] で行い、システムを操作しながら、思ったことや考えたことを実験者に対して発言するように指示をした。

システムの動作環境は、web ブラウザ (Mozilla Firefox ver.26.0²) とした。参加者のシステムの操作過程や課題時における発言の記録には、QuickTime Player のデスクトップキャプチャ機能を利用した。実験者は、課題中の参加者の振る舞いに関して逐次メモを取り、後のインタビューに備えた。

5.3 従来型システムを用いた実験の結果

従来型システムを用いた実験参加者の平均探索時間は 3 分 57 秒であり、平均探索回数（楽天市場の上位・下位ジャンルを選択した回数）は 12.6 回であった。図 7 は従来型システムの探索過程の一例をグラフに示したものである。グラフの縦軸は楽天のジャンルを示している。黄色のラインは楽天の上位ジャンルであり、各下部にはそのジャンルの下位ジャンルを示している。横軸は探索におけるジャンルの選択回数であり、何回目にどのジャンルを選択したかを示している。

5.4 従来型システムの探索傾向

従来型システムを用いた実験において、参加者の探索過程の特徴を観察したところ、以下の 2 点が確認された。

- (1) 上位ジャンルから下位ジャンルへと階層を辿る傾向がある

²<http://www.mozilla.jp/firefox/>

(2) 少数のジャンルにみを閲覧し商品決定する傾向にある

(1) に関しては、ツリー構造に従って、上位ジャンルから下位ジャンルへの探索を繰り返すことで商品探索を行っていた。このような探索は、プレゼントを考えながら探していくようなコンセプト精緻化型の購買行為にあてはまる。(2) に関しては、最初から想定していた商品以外の探索には至らなかった。参加者が相手の趣味や嗜好について、ある程度予測をしてジャンルや商品を絞り込んで探索を行っていたことが確認された。このような探索は、問題解決型の購買行為にあてはまる。

5.5 提案システムを用いた実験の結果

提案システムを用いた実験参加者の平均探索時間は6分42秒であり、平均探索回数は26.8回であった。図8は提案型システムの赤いノードと青いノードを利用した探索過程をグラフに示したものである。赤いノードを選択した時と青いノードの楽天ジャンルを選択した時に座標を打ち、それぞれの座標を線で結んでいる。

提案システムで実際に利用されたジャンルの割合は、楽天市場で用いられている上位ジャンルが52.9%、独自で設定した上位ジャンルが47.0%であった。

5.6 提案システムの探索傾向

提案システムを用いた実験において、参加者の探索過程の特徴を観察したところ、以下の2点が確認された。

- ツリー構造を横断的に探索する傾向がある
- 興味のあるジャンルを選択し、相手の趣味や嗜好を考慮した探索を行う傾向にある

図8に示す様に、多様なジャンル間の繋がりを提示したことで、実験参加者がツリー構造における下位ジャンルから他の上位ジャンルが持つ下位ジャンルへ横断的に探索を行い、商品を見つけ出す様子が確認された。ツリー構造の横断的な探索を実現したことで、ユーザが試行錯誤を繰り返しながら柔軟な探索を行うことが可能になり、その結果、探索回数が従来型システムに比べ増加したと考える。図8に示す参加者の探索過程の様子と発話内容を、表1に示す。

6 考察

提案システムを用いた実験の参加者は、ツリー構造を横断的に探索し、商品探索の試行錯誤を円滑に繰り返すことができた。

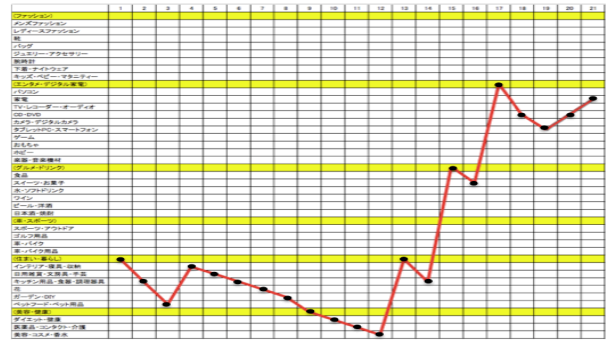


図 7: 従来型システムの実験結果の例

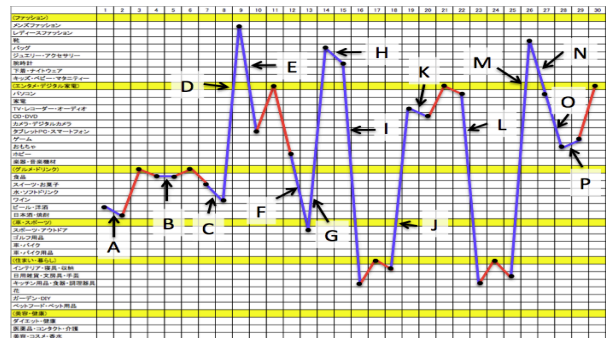


図 8: 提案システムの実験結果の例

返すことができた。このことから、多様なジャンル間の繋がりは、他者にプレゼントを購入するような試行錯誤を必要とする商品探索に有効であると考えらる。

今後の展望として、より多様なジャンル間の繋がりを見出すために、(1) 商品タイトルや商品レビューといった商品に付随する情報を分析する、(2) 実験における参加者の探索過程や発話内容から、どのような繋がりを意識して探索を行ったのか、どのような繋がりがあれば、より円滑な商品探索を行うことができたのかを分析する。しかし、多様なジャンル間の繋がりが膨大になると、ユーザの選択肢の幅が増加するため、効率的に商品探索を行うことが困難になると考える。そのため、多様なジャンル間の繋がりを見出す際に、適切なジャンル数を設定する必要があると考える。

7 結論

本研究ではECサイトを利用したプレゼント選定における商品探索の試行錯誤を円滑にすることを目的とし、多様なジャンル間の繋がりによるツリー構造の横断的な探索を試みるシステムを提案した。今回は多様なジャンル間の繋がりの有効性を観察するため自身でジャンルを設定した。提案システムが提示する多様なジャンル間の繋がりがユーザに与える影響を観察するため

表 1: 探索過程におけるジャンルの推移

	探索前	独自ジャンル	探索後	発話内容
A	ビール・洋酒	ギフト	日本酒・焼酎	父に誕生日プレゼントを貰おう そういえば、お酒が好きだったからお酒で探そう
B	食品	料理	食品	ギフトって考え方もあるのか それなら食品を見てお酒のおつまみを考えてみよう
C	スイーツ・お菓子	ギフト	ワイン	カニいいな、あ、そういえばお父さんアーモンド好きだったな
D	ワイン	父の日	メンズファッション	お父さんって趣味あんまりないからな。あ、iPad あるやん お父さん YouTube めっちゃ見てるし、これ良いかもしれない
E	メンズファッション	ビジネス	タブレット PC・スマートフォン	
F	ホビー	趣味	スポーツ・アウトドア	
G	スポーツ・アウトドア	スクール	バッグ	腕時計より、iPad かな
H	バッグ	ビジネス	腕時計	
I	腕時計	父の日	キッチン用品・食器・調理器具	
J	インテリア・寝具・収納	インドア	TV・レコーダー・オーディオ	鑑賞、音楽好きじゃないけど、映画なら見るかもしれないな
K	TV・レコーダー・オーディオ	鑑賞	CD・DVD	キッチン用品で何かおもしろいものがあれば喜んでくれるかも 便利なキッチン用品は、どちらかと言えばお母さんが喜んでくれるものだな
L	パソコン	新生活	キッチン用品	日用雑貨、どうしよう、靴
M	日用雑貨・文房具・手芸	ビジネス	靴	パソコン見たっけ
N	靴	ビジネス	パソコン	あ、このコピー用紙、家に欲しい、オタクって新たなジャンル
O	パソコン	オタク	おもちゃ	ホビー、ゲームか、あ、そういえばお父さん麻雀やってるな
P	おもちゃ	オタク	ゲーム	だいたい見たかな、やっぱりお父さんにプレゼントするなら iPad で決まり

に、従来型システムを実装し比較を行った。実験結果から、提案システムは従来型システムに比べ、探索回数と探索時間が増加した。探索過程を観察すると、相手の趣味や嗜好といった曖昧な条件のもと、様々な思考を巡らせながらツリー構造を横断的に渡り歩くことで、ユーザの興味の幅が広がり様々なジャンルを閲覧しつつ探索的にプレゼントを見つけ出す様子が確認できた。ユーザが多様な観点を持って商品探索の試行錯誤を円滑に行う 1 つの手段として、多様なジャンル間の繋がりを有いたツリー構造の横断的な探索が有効であると言える。

今後は、システムとのインタラクションだけでなく、システムのアルゴリズムにも着眼し、より商品探索の試行錯誤が円滑に行うことができるインタフェースへと改良する。

謝辞

本論文の執筆にあたり関西大学大学院総合情報学研究科の白水菜々重氏、空中海人氏に多大なる支援を頂いた。本研究は科学研究費補助金基盤研究 (C)(課題番号: 24650040) の助成を受けた。また、システムデザインの際に、楽天株式会社から提供された楽天データセットを利用した。記して謝意を表す。

参考文献

- [1] Marchionini, G.: Exploratory Search: from Finding to Understanding, *Communication of the*

ACM, Vol. 49, No.4, pp. 41-46 (2006)

- [2] Ericsson, K. and Simon, H.: *Protocol Analysis: Verbal Reports as Data*, MIT Press (1993).
- [3] 庄司裕子: 気づきからコンセプト精緻化へ、そしてチャンス発見へ、人工知能学会誌, Vol. 18, No. 3, pp. 269-274 (2003).
- [4] 土方嘉徳: 情報推薦・情報フィルタリングのためのユーザプロファイリング技術, 人工知能学会誌, Vol. 19, No.3, pp. 365-372 (2004).
- [5] 宮崎和也, 空中海人, 松下光範: 贈り物選択を対象にした探索的な情報検索行為の支援, ファジィシステムシンポジウム講演論文集, Vol. 29, pp. 613-618 (2013).

探索的データ分析支援に向けた TETDM インタフェースの改良に関する基礎検討

Basic study on an Improvement for TETDM Towards the Goal of Supporting Exploratory Date Analysis

井須 弘恵¹
Hiroe Isu¹

大塚 直也²
Naoya Otsuka²

松下 光範^{1*}
Mitsunori Matsushita¹

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

Abstract: Knowledge discovery on text mining requires a trial-and-error process so that a user's informational requirements are unclear when they start his or her exploration. Our purpose is to support a user's information seeking behaviour on text mining. In this paper, we observe how a user behave on TETDM : Total Environment for Text Date Mining. According to the experiment, we found that 5 usability problems and 1 problem for TETDM. From obtained results, we sort the system requirements and propose that a design criteria to facilitate a user's information seeking.

1 はじめに

近年、構造化されていないテキストデータから新たな情報や知識を発見するための分析手法として、テキストマイニングが注目されている。テキストマイニングとは、文書集合から新しい情報や知識を発見することのできる「構造化されていないテキストデータからの情報抽出に関する技術」である [1]。

テキストマイニングによる知識発見は、あらかじめ分析のゴールが明確に定まっているようなゴール指向のタスクではなく、有益な情報や知識を発見する探索的な情報アクセス [2] を必要とするタスクと捉えることができる。Hearst によると、テキストマイニングのゴールは、データから新たな情報や知識を発見かつ誘導し、何かしらのパターンが存在するかを調べることであり [3]。テキストマイニングは、単なる情報検索の技術、発話解析、語義曖昧性の解決、辞書作成などの自然言語処理技術やテキスト要約などの技術にあるのではなく、それらを利用した「探索的データ解析」にある。

テキストマイニングの技術は既に多くの研究成果が報告されているが、実際に世の中で使われる技術は一部に限られている。この問題を解決するために、Total Environment for Text Data Mining(以下 TETDM と略す)が提案されている [5]。TETDM とは、世の中に分散しているテキストマイニングの技術を同一環境上で柔軟に組み合わせて分析を行うことができる統合

環境である。現在 TETDM の統合環境は開発途中であり、そのインタフェースはテキストマイニングのような試行錯誤を伴う探索的な情報アクセスに必ずしも適したものではないことが指摘されている [4]。

そこで本研究では、テキストマイニングにおける知識発見のためのユーザの探索行為を支援するために、TETDM のインタフェースが満たすべき要件を整理することを目的とする。そのために、実際に TETDM を用いて、ユーザの情報探索行動を観察する。得られた結果から TETDM が抱える問題点を整理する。また、TETDM インタフェース改良の先行研究として提案されている大塚らのインタフェース (図 3 参照) の有用性を評価するために、ユーザ観察を行う [4]。それら 2 つの実験の結果から、ユーザの探索行為の円滑化のために満たすべき要件を整理し、TETDM のインタフェースの改良指針を提案する。

2 関連研究

2.1 TETDM の概要

TETDM は複数のウィンドウで構成されており、それぞれの画面で異なった分析および結果を表示することが可能である。TETDM ではテキストマイニングのプロセスを、自然言語処理、データマイニング、情報可視化の 3 つのプロセスとして捉えている。TETDM に入力されたテキストは、形態素解析などの前処理の後、各モジュールによって処理が行われる。モジュール

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

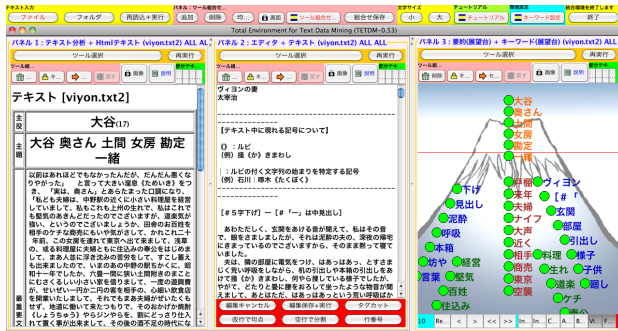


図 1: TETDM の概観

とは、テキストマイニングを行う機能ごとにまとめられた要素であり、これを TETDM に追加していくことにより、扱える技術を拡張することができる。ユーザは、自らの興味や目的に応じてツールを選択してそれらを自由に組み合わせたり、処理結果を比較することによって、多角的なテキスト分析を行うことができる。

TETDM では、マイニングツールと可視化ツールの 2 つの技術が統合環境内のモジュールとして実装されており、それらをユーザが切り替えることで処理内容と処理結果の表示手法を変更することができる。そのため、TETDM を用いて分析を行う場合、ユーザはマイニング処理ツールと可視化ツールをそれぞれ 1 つずつ選択し、組み合わせて結果を比較しながら分析を進めていく。ユーザは様々な分析手法の中から適当な手法を選択して分析を行い、得られた結果を解釈・考慮して次の分析手法を試みる、といった探索プロセスを繰り返す。

現在のバージョン 0.54 では、利用可能なマイニング処理ツールと可視化ツールはいずれも 28 種類ずつであるが、組み合わせのパターンによって 45 種類の分析が可能である。マイニング処理ツールと可視化ツールは単体では機能せず、それぞれ対となるモジュールを必要とする。図 2 は、ユーザがマイニング処理ツールを選択する場面において、「光と影」が選択された状態である。左の列がマイニング処理ツールのリスト、右の列が可視化ツールのリストである。オレンジ色で示されているものが、推奨されたツール同士の組み合わせである。マイニング処理ツールである「光と影」に対しては、「キーワード選択」、「スコア分布」、「テキスト(カラー)」の 3 種類の異なった可視化結果が推奨されている。

2.2 試行錯誤を支援するインタフェース

前節でも述べたように、テキストマイニングの本質は探索的なアプローチにあり、分析を進めていく中で、自分の求める情報要求を精緻化・明確化していく。しかし大塚らは、現在の TETDM インタフェースは必ず

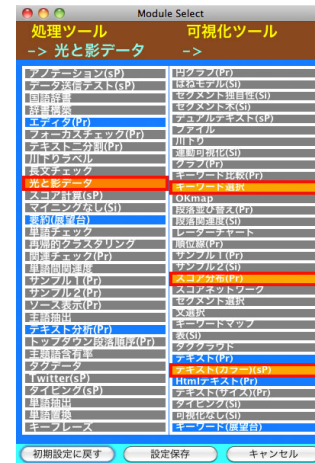


図 2: 「光と影」を選択した状態

しもこのような探索的な分析に適していないことを指摘し [4]、ユーザの試行錯誤の円滑化を目的とした新しいインタフェースを提案している (図 3 参照)。大塚らは、モジュール切り替えのためのインタフェースは、直感的な操作をできることが望ましいとした。また、複数のテキストマイニングツールを用いて多面的に分析を行うことを想定して、分析内容が多岐に渡り、作業が煩雑になる可能性があるとして指摘した。これを解決するために、ユーザは自身の分析プロセスを常に把握し、現在行っている分析に引き続いてどのような分析ができるのかを把握できることが必要であると考えた。このような要求に応えるため、提案インタフェースでは、各モジュールをノード、それらを処理の順に繋ぐ線をリンクとするグラフ表現が採用されている (図 3 左参照)。提案手法では、各モジュールを表すノードをマウスで直接操作することによって、ツールの選択・切り替えを行えるようになっている。このようにオブジェクトを直接操作することにより、直感的なモジュール切り替えを実現でき、ユーザは試行錯誤を妨げられることなく分析を進めていくことができると考えられる。

このような先行研究を受けて、本研究では、TETDM インタフェースが満たすべき要件を明らかにするために、TETDM を実際に用いた実験を行い、ユーザの情報探索行動を観察する。また、大塚らのインタフェースの有用性を測り、TETDM インタフェースの満たすべき要件を明らかにするために実験を行う。

3 従来インタフェースを用いた観察

上述したように、本研究の目的は、TETDM を用いてユーザの情報探索を観察し、TETDM インタフェースが探索的なデータ分析を行うユーザを支援するにあたって考慮すべき点を発見することである。探索的なデータ

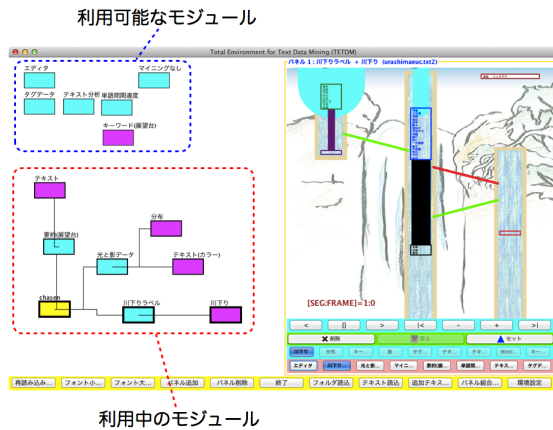


図 3: 大塚らによる提案インタフェース [4]

分析では、ユーザによってその分析方法が異なるため、同じデータを分析対象にした場合でも、すべての分析者が同じ結論にたどり着くことは期待できない。そのため、その行為を支援するシステムの評価においては、タスクの達成度や正答率といった指標を用いることができないことから、本実験では、ユーザが実際に TETDM を使用し、ユーザがどのように分析を行うかについて観察を行う。

3.1 実験の概要

実験では、TETDM (ver. 0.53) を用いてユーザ観察を行った。実験参加者は情報学を専攻する大学院生 2 名 (男性 2 名)、4 年生大学の情報学部に通う 1 名 (男性 1 名)、3 年生 1 名 (女性 1 名) と社会人 1 名 (女性 1 名) の計 5 名であった。事前説明では、TETDM インタフェースの改良を目的としていることを実験参加者に説明し、ユーザの立場から問題点や改善点を指摘することを求めた。実験の課題は、TETDM を用いて小説『ヴィヨンの妻 (太宰治著)』から登場人物を発見し、発見した人物の特徴 (性格や年齢、職業など) について説明することとした。実験は、時間制限を設けず、参加者の考えがまとまった時点で終了とした。これは、テキストマイニングがゴールがある課題ではなく、一定の結論に至ったと自覚することがゴールとなるような課題と捉えられるためである。実験を始める前に、参加者に対して、TETDM の使用経験と、課題とする小説の読書経験について確認を行ったところ、参加者全員が過去に TETDM を使用したことがなく、課題の書籍が未読であることが確認された。参加者は実験中にメモを取ることが許され、実験中の様子は VTR で記録された。また、実験終了後に、分析を進めるにあたって障害となった点を聞き取るために、(1) どのよ

うに分析を進めたか、(2) 分析を進める上で障害と感じた点はあるか、の 2 つの質問を用意した。

3.2 実験の結果

以下ではユーザ観察を行った順に、参加者を A, B, C, D, E と記す。実験参加者が TETDM を操作する様子を観察したところ、TETDM インタフェースの 4 つの問題が明らかになった。

ツールの組み合わせに関する問題

TETDM では、マイニング処理ツールと可視化ツールをそれぞれ一つずつ選択し、組み合わせることで分析結果を得ることができる。事後インタビューにおいて、質問 (1) に対しては、全員から、様々なツールの組み合わせを試行したのちに、使用する分析手法を決定し、分析を開始したという回答が得られた。しかし、ユーザ観察の結果、全員が推奨されていないツールを選択し、正しい処理結果が得られなかったことが観察された。また、ツール選択の際、処理の順序が無視されていた。本来であれば、まずマイニング処理ツールを選択してから可視化ツールを選択する必要があるが、可視化ツールを最初に選択する様子が全員に観察された。これについて事後インタビューで参加者 C は、ツールの組み合わせがよく理解できなかったと述べた。また、参加者が使用するツールの組み合わせのペアは、45 種類ある中で 2、3 種類程度と限られたものであるということが分かった (表 3.2 参照)。

処理の進行状況の提示に関する問題

事後インタビューにおいて、5 名中 4 名が、自分が選択したツールの処理が正しく行われているかどうか分からなかった回答した。参加者 A は、「処理に時間がかかるものだと、使えないのか使えるのかが分からないので、処理中であることを表すものが欲しい」と述べた。また、ユーザ観察の結果、5 名中 3 名が、処理結果が表示される前に別の分析へ移行する様子が観察された。

ツール名称に関する問題

事後インタビューにおいて、5 名中 4 名がマイニング処理ツールと可視化ツールの名称から分析内容や可視化結果を想像できないため、分析を進める上で障害となったと回答した。TETDM においてユーザは、可視化ツールであれば、「川下り」や「タグクラウド」、「キーワード (展望台)」、「OKmap」、「セグメント独自性」などといった名称から選択する。しかし、テキスト分析に馴染みが薄いユーザにとって、「キーワード (展望台) (図

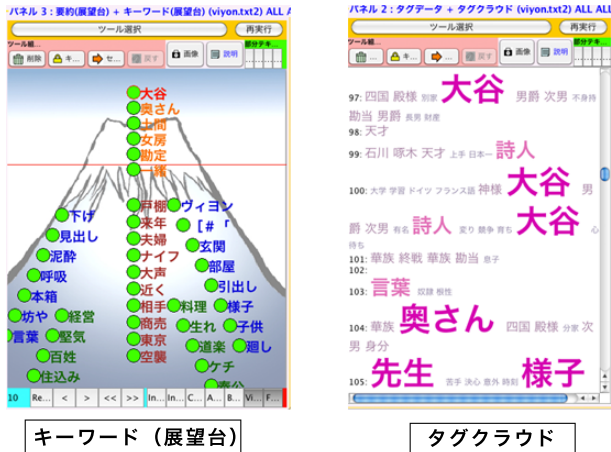


図 4: TETDM での可視化結果

表 1: 実験参加者が使用したツールの組み合わせ

実験参加者	使用したツール
A	光と影データ+キーワード選択 テキスト分析+ HTML テキスト (Pr)
B	テキスト分析 (Pr) + HTML テキスト (Pr) 光と影データ+キーワード選択 要約 (展望台) + キーワード (展望台)
C	タグデータ+タグクラウド 単語抽出+テキストカラー (Sp)
D	関連チェック (Pr) + キーワード選択 テキスト分析+ HTML テキスト 要約 (展望台) + テキスト (Pr)
E	タグデータ+タグクラウド 長文+ HTML テキスト (Pr) 要約 (展望台) + テキスト (Pr)

4 左参照)」や「タグクラウド (図 4 右参照)」などのツールの名称から可視化結果を想像することは困難であると想像される。

ユーザの情報要求の解決に関する問題

事後インタビューにおいて、5 名中 3 名が、自分の情報要求に対する適切な分析手法が分からず、ツールの選択に躊躇したと回答した。参加者 B は、「たくさんツールはあるけど、使うツールの組み合わせが決まってくるから、予測変換機能のようなものがあれば良いと思う」と述べた。

4 提案インタフェースを用いた観察

本章では、大塚らの提案するインタフェースの有用性を評価することを目的に実施した実験について述べる。大塚らが提案するインタフェースは、オブジェクトを直接操作することで、ツールの選択・切り替えが行うことができる。直感的なモジュールの切り替えに

より、ユーザは試行錯誤を妨げられることなく分析を進めていくことができると考えられる。さらに、処理の流れがリンクされるのでユーザは分析内容を把握できると推測される。そこで、本実験では、提案インタフェースを使用し、ツール選択に関して改善点があるか、分析内容把握できているかの 2 つの観点について、ユーザ観察を通して考察する。

4.1 実験の概要

実験参加者は情報学を専攻する大学院生 1 名 (男性 1 名)、4 年生大学の情報学部に通う 2 年生 2 名 (女性 2 名)、4 年生 (女性 2 名) の計 5 名であった。実験の参加条件、実験で対象とした小説、課題は TETDM を用いた実験 (3.1 節参照) と全て同様である (3.1 参照)。実験終了後のインタビューでは、提案インタフェースの有用性を測るために (1) ツールの切り替えは円滑に行えたか、(2) 分析内容を把握できたか、(3) 分析を進める上で障害と感じた点はあるか、の 3 つの質問を用意した。

4.2 実験の結果

以下ではユーザ観察を行った順に、参加者 A, B, C, D, E と記す。実験参加者がインタフェースを操作する様子を観察したところ、推奨されていないツールの組み合わせを選択する様子は全員に観察されなかった。加えて、(1) の質問に対して全員が、ツールの組み合わせを迷うことなく、切り替えもスムーズに行えたと回答していることから、ツール組み合わせに関する問題は改善されたと考えられる。しかし、(2) の質問に対して、5 名中 4 名が分析内容を把握できなかった、自分が行っている分析内容をあまり意識しなかったと回答していることから、分析内容の把握については改善されていないことが示唆される。さらに、(3) の質問に対して得られた指摘や意見から 4 つの問題が明らかになった。以下に、それぞれについて説明する。

ユーザとシステム間のインタラクションの問題

事後インタビューにおいて、5 名中 4 名が分析の過程で気になった特定のキーワードや調べたい事柄が、元のテキストデータ内のどの位置にあるか知りたかったと述べた。例えば、タグクラウド (図 4 右参照) では、頻出度の高い単語 (e.g., 大谷、奥さん) が拡大して表示されるが、それらの単語は原文とリンクされておらず、本文中のどの位置に出現しているか把握することができない。ユーザ観察からも、5 名中 4 名に可視化結果をクリックする様子が見られた。

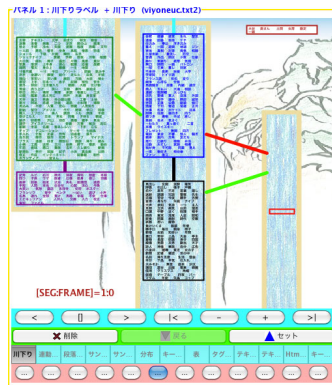
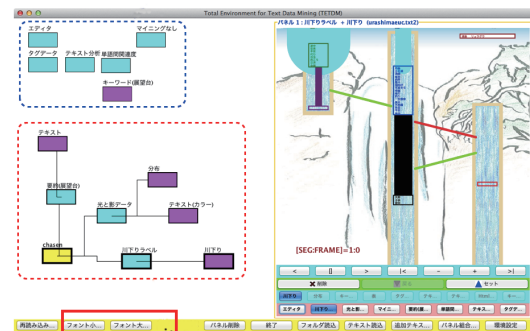


図 5: 川下りの処理結果



文字サイズ変更ツール

図 6: 文字サイズ変更ツール

ユーザの情報要求の解決に関する問題

事後インタビューにおいて、5 名中 4 名から調べたいことに対してどの手法が適しているのか分からず、様々な分析手法を何度も繰り返したという意見が得られた。参加者 E は「自分の調べたいことに対してどのツールが適しているのか分からない。」と述べている。

ツール間の関係性に関する問題

質問 (2) に対する意見として、5 名中 4 名が自分の行っている分析内容をあまり意識せず、可視化結果を中心に分析を進めていったと回答した。ことことから、どのような分析手法で可視化を導くかの選択を決定するマイニング処理の部分が、ユーザに意識されなかったことが明らかになった。参加者 C は、質問 (2) に対して、「どういった過程で可視化結果を導いたかの説明がされていないから、結果を信用することができない」と述べた。

メニューの表示方法に関する問題

この問題は、大塚らがインタフェースを構築する際に従来のインタフェースとは異なるデザインを採用したため、起きた問題である。事後インタビューにおいて、全員が提案システムの結果を表示するウィンドウが小さいため、細かい文字が読めなかったという指摘を行った。特に一番指摘の多かった可視化ツールが「川下り(図 5 参照)」であり、ユーザ観察からも可視化結果をクリックするなどの様子が全員に見られている。参加者 A は、「クリックすれば画面が文字が大きくなる」と思い何度もクリックしてしまった」と述べている。フォントサイズは文字サイズ変更ツール(図 6 参照)の位置で変更可能だが、全員がそのボタンの存在に気付かなかった。

5 デザイン指針

本研究では、TETDM と提案インタフェースを用いた 2 つのユーザ観察(3 章および 4 章参照)を通して得られた問題点をインタフェースの問題点と、TETDM 機能の問題として大別し、それぞれについて解決策の検討を行う。

5.1 TETDM インタフェースの問題

本研究では、3 章と 4 章で明らかになった問題点を TETDM インタフェースの問題を以下の 5 つにまとめる。

1. 処理の進行状況の提示に関する問題
2. メニューの表示方法に関する問題
3. ユーザの情報要求の解決に関する問題
4. ツールの名称に関する問題
5. ツール間の関係性に関する問題

以上の 5 つの問題点から探索的な分析を行うにあたって TETDM のインタフェースが改善すべき点について検討する。

1 つ目の問題に対する解決策として、プログレスバーを用意することで、ユーザに処理状況のフィードバックを与えることを提案する。ユーザ観察から分かるように、システムから操作に対する反応がない場合、多くの人が繰り返し同じ操作を試みたり、他の操作に移ったりする。そこで、処理に時間がかかるツールが選択された場合は、その進捗状況を知らせるためのプログレスバーを表示するべきであると考えます。

2 つ目の問題に対する解決策として、パネルないしツールバーを用意し、副次的なツールやコンテンツをまとめることを提案する。ユーザ観察からも分かるように、ユーザは細部ツールへ注意を向けていないこと

が分かる (4.2 節参照)。そこで、文字サイズの変更やチュートリアル参照といった特定のツールに付随するツールは、パネルを用意してまとめるべきであると考え。

3 目目の問題に対する解決策として、2 点の案が考えられる。1 点目は、選択肢のグループ化である。現在のインタフェースにおいて、選択可能なツールを示すリストでは、ツール間の関連性を考慮せずにツール名が並べられている状態である (図 2 参照)。その中から、ユーザが自身の要求に対して適切なツールを即座に選択することは困難である。この問題を解消する方法として、ツールをグループ化することで、ユーザの選択を支援することが望ましい。2 点目は、ツール推薦機能の付与である。ユーザが行いたい分析に適したツールをインタフェースが推薦することによって、ユーザの試行錯誤は円滑に進められると考えられる。

また、ツール推薦機能を追加することにより、5 目目の問題のツール間の関係性に関する問題も解決されると想定している。推薦を行う際に、例えば、タグクラウドであれば、単語の頻度により大きさが変わるといように、テキストデータから分析を導き出すまでの過程が説明されれば、テキストマイニングに知識がない初心者でも使いやすいインタフェースになると考えられる。

4 目目の問題に対する解決策として、分析内容や可視化結果を容易に判断できるようなアイコンをツールリストを示すことを提案する。探索的データ分析では、可視化結果によって洞察を得ながら分析を進めていくため、ユーザが得られる結果を一目で理解できることが望ましい。

5.2 TETDM 機能の問題

提案インタフェースを用いた実験 (4.2 節参照) でのユーザ観察から、「タグクラウド」(図 4 右参照) や「キーワード (展望台)」(図 4 左参照) などの可視化結果で得られたキーワードから原文にリンクしたいという意見や、特定のキーワードを検索をしたいという意見が得られたことから、TETDM とユーザ間のインタラクションが円滑に行えていないことが予想される。このような問題を解決するために、クエリ機能や検索窓を設けるなどユーザとシステム間の対話を妨げないデザインがなされるべきであると考え。他にも、従来インタフェースを用いた実験 (3 章参照) での事後インタビューで、「主語抽出ができていないと思う」、「キーワード選択が可能なが、それが反映されているように思えない」という意見も挙げられている。このように、TETDM 機能は他にも多くあるため、個々の問題については新たに考察する必要があると考える。

6 おわりに

本稿では、テキストマイニングによる知識発見のためのユーザの探索行為の支援を目的と、ユーザの情報探索行動を観察することで、TETDM が抱える問題点を整理した。また、先行研究として提案されているインタフェースを用いてユーザ観察を行い、その有用性を評価した。2 つの実験の結果からユーザの探索行為の円滑化のために TETDM のインタフェースが満たすべき要件について考察を行い、要件を整理した。また、それらを基に TETDM インタフェースの改良指針を提案した。今後は、提案した改良指針に基づいて実装を行い、テキストマイニングを行いたいユーザの探索行為の円滑化について検討していく。

7 謝辞

本研究は科学研究費補助金基盤研究 (C)(課題番号: 22300048) の助成を受けた。記して謝意を表す。

参考文献

- [1] Rajiman, M. and Besancon, R.: Text Mining : Natural Language techniques and Text Mining applications, *Proc. 7 th IFIP 2.6 Working Conference on Database Semantics*, pp. 7–10 (1997).
- [2] Marchionini, G. : Exploratory Search: From Finding to Understanding, *Communication of the ACM*, Vol. 49. No. 4, pp. 41–46 (2006).
- [3] Hearst, M. A.: Untangling text data mining, *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 3–10 (1999).
- [4] Otsuka, N. and Matsushita, M.: Graphical Interface that Supports Users' Trial-and-Error Process of Text Mining, *Proc. JSAI2013 International Organized Session: Special Session on Intelligent Data Analysis and Applications*, 1A3-10S-3a-2 (2013).
- [5] 砂山渡, 高間 康史, Bollegala, D., 西原 陽子, 徳永 秀和, 串間 宗夫, 松下 光範: Total Environment for Text Data Mining: テキストデータマイニングのための統合環境, *人工知能学会論文誌*, Vol. 26, No. 4, pp. 483–493 (2011).

統合環境 TETDM を用いた知識創発支援の枠組み

Framework of Knowledge Creation Support with TETDM

砂山 渡^{*1}
Wataru Sunayama

¹ 広島市立大学大学院 情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

Abstract: In TETDM challenge, we develop and distribute an integrated environment to flexibly combine multiple text mining techniques. Text mining techniques include numerous tasks such as salient sentence extraction, keyword extraction, topic extraction, textual coherence evaluation, multi-document summarization, and text clustering. Although tools that individually perform one or more of the above-mentioned tasks exist, it is difficult to integrate and activate multiple tools for a particular task. Therefore, TETDM is suggested as a system that can integrate numerous tools and can contribute to our creative work. In this paper, framework of knowledge creation support with TETDM is described. Knowledge creation consists of “trials and errors” and “logical interpretations of the facts”.

1 はじめに

人工知能学会全国大会の近未来チャレンジテーマとして進められている TETDM (Total Environment for Text Data Mining: テトディーエム)¹は、複数のテキストマイニング技術を柔軟に組み合わせて使える統合環境を構築し、社会的創造的活動を支援できる環境としての提供を目指している [砂山 11, 砂山 13, 砂山 14]. 近未来チャレンジでは、5 年以内の目標達成を必要条件としており、現在 4 年目 (2014 年の全国大会で 5 年目に入る) の TETDM は、統合環境の正式公開を 2015 年 4 月と目標を定めて進めている (図 1).

これまでに、複数のモジュールを統合的に扱う環境の枠組み作りを行い、2011 年 12 月の試用 α 版の公開を皮切りに、2014 年 3 月現在でバージョン 0.55 までを公開している. 統合環境ならびに必要な情報の公開は、TETDM サイト上で行っており、継続的にアクセスとダウンロードがなされている.

本稿では、この統合環境 TETDM を用いた知識創発支援の枠組みについて述べるとともに、現在までの実装の状況と今後の見通しについて述べる. 本研究での知識創発は、データとツールを用いた創造的思考と考えられ、さまざまな可能性を列挙する発散的思考のために、多様な観点からデータを分析して知見を集める場面と、集められた知見の集合を、収束的思考により一つの汎用的な知識となるようにまとめる場面からなる.

(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

¹TETDM サイト: <http://tetdm.jp/>

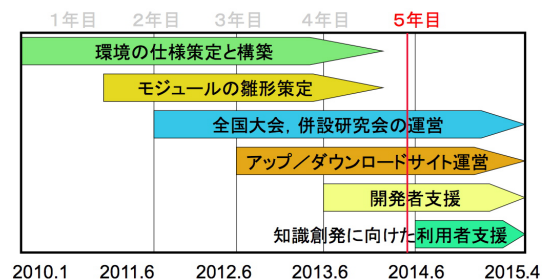


図 1: 近未来チャレンジ TETDM のスケジュール

2 TETDM 統合環境の構成

本章では、TETDM 統合環境の構成と、利用者支援のために新たに実装した機能について述べる. TETDM は、あるテキストを入力したときに、その分析処理を行う処理モジュールと、処理結果を可視化して出力する可視化モジュールを複数備えた環境となっている (図 2). またこれらのモジュール群は、単独の開発者によって提供されるものではなく、任意の開発者がモジュールを作成して利用することができる.

TETDM は、独立した複数のパネル内に、処理モジュールと可視化モジュールを 1 つずつペアとしてセットすることで動作する. 各パネルにセットされる異なる開発者によって作成されたモジュールが、それぞれ独立に動作するだけではなく、それらを協調的に連動させることができる点が TETDM の特徴となっている.

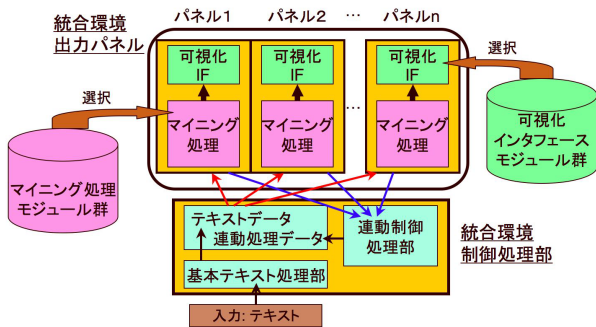


図 2: TETDM 統合環境構成図

以下で各部の説明を述べる²。

2.1 入力：テキスト

TETDMに入力されたテキストは「セグメント」「文」「単語」の3つに分割して扱われる。「単語」へ区切る際は、形態素解析器を用いて単語に分割する。この際、指定した品詞の単語だけを、キーワードとして取り扱う。「文」に区切る方法は、テキスト中の句点記号（「。」や「.」）をもとに分割する。「セグメント」に区切る方法は、特定の文字列をテキスト中に挿入し、その文字列をもとに分割する。その後、単語の出現情報や頻度情報の計算、キーワードやセグメント間の関連度計算を行った結果をデータ構造に格納し、このテキストデータをもとに各モジュールが処理を行う。

2.2 マイニング処理モジュール

マイニング処理モジュールは、統合環境内のテキストデータをもとに、テキストの理解や分析に役立つ情報をテキストから抽出する。またマイニングという言葉にこだわることなく、テキストに何らかの処理を施すモジュールも対象とする。現在までに30以上の処理モジュールが作成、公開されている。マイニング処理モジュールの処理結果は、可視化インタフェースモジュールに渡されて出力される。

2.3 出力：可視化インタフェースモジュール

可視化インタフェースモジュールは、マイニング処理モジュールによる出力結果を可視化する³。可視化インタフェースモジュールでは、入力として受け取れるデータ型（boolean, int, double, String 型とその一次

²構成の詳細は [砂山 14] を参照していただきたい。

³マイニング処理モジュールの結果によらず、統合環境のテキストデータを可視化するモジュールであっても良い。

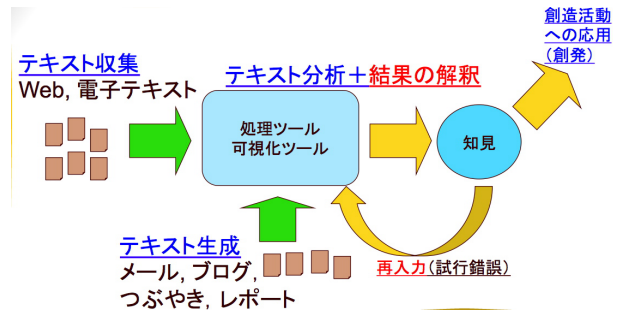


図 3: テキストマイニングのプロセスと知識創発

元配列、二次元配列 (String 型以外) の11種類) を定め⁴、そのデータの意味を表す整数型変数との組合せにより、マイニング処理モジュールからデータを受け取ることができる。現在までに30以上の可視化モジュールが作成、公開されている。

3 TETDM を用いた知識創発の枠組みと実装

本節では、統合環境 TETDM を用いた知識創発支援の枠組みについて述べる。

3.1 テキストマイニングのプロセスと知識創発

図3に、テキストマイニングのプロセスと知識創発との関係を示す。テキストマイニングのユーザは、テキストをツールに入力してその分析結果を取得する。ユーザは、得られた結果に対して解釈を与え、意味を理解することによって知見（一つの考え）を得る。また得られた知見について、その知見の信頼性や妥当性の確認しながら、質の高い知見を得るための試行錯誤を行う。そして集められた知見を総合的に解釈することで、実際に活用可能なレベルの知識へと変換させる（創発）。これらの用語は以下のようにまとめられる。

知識創発：知見を積み重ねることで汎用可能な一般化された新たな知識を発見すること

試行錯誤：得た知見について根拠や理屈を確認するため、または新たな知見を得るために、再入力を行うこと（発散的思考）

結果の解釈：テキストの意味や性質を理解して知見（一つの考え）を得ること（収束的思考）

⁴TETDM は Java で実装されている。

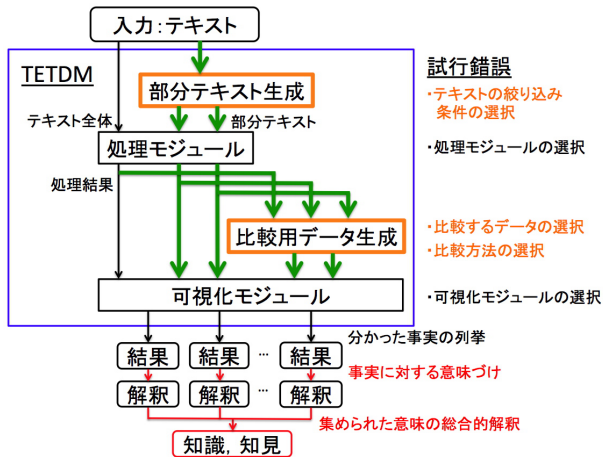


図 4: TETDM を用いた知識創発の枠組み

3.2 TETDM を用いた知識創発の枠組み

前節の定義に基づいて、知識創発を実現するためには、多くの知見を得るための試行錯誤を繰り返す必要があり、この試行錯誤のバリエーションを用意することで、得られる結果の幅が広がると想定される。

図4に、TETDMを用いた知識創発の枠組みを示す。2章で述べた基本的な構成のTETDMにおいては、使用する処理モジュールと可視化モジュールを選択すること（ならびに各モジュール内で用意された機能を用いること）で試行錯誤を行っていた（図4右の「処理モジュールの選択」「可視化モジュールの選択」）。新しい枠組みにおいては、テキストデータを絞り込む「部分テキスト生成」、ならびに全体と部分、あるいは部分同士を比較可能にする「比較用データ生成」の機能をTETDMにもたせることで、試行錯誤の幅を広げる。前者は、データ量が多い場合には必須となり、データ量が少ない場合でも、まずはデータの一部に着目して、そこからわかる事実を見つけることの重要性は高い。後者は、一部のデータの特徴を捉えるためには、基準となるデータ全体の傾向との比較や、他の部分データとの傾向を比較する必要がある。

また、結果の解釈については、通常のツールは結果を出力するのみとなっており、その結果の解釈を積極的に支援するものにはなっていない。出力された結果に意味づけを行うこと、また意味付けされた結果を集めて整理する環境を用意することが、知識創発につながると考えられる。

これらの「試行錯誤」と「結果の解釈」は、創造的思考における、発散的思考と収束的思考 [Guilford 59] に対応すると考えており、既存研究から得られている知見も適宜採り入れていきたいと考えている。現時点での枠組みの詳細について以下で述べる。

3.3 試行錯誤

知見を得るための試行錯誤として、以下の3つを想定している。

1. 想定内外のツールの選択と組合せ
2. 部分テキストの生成
3. 比較用データの生成

1. はTETDMにおける使用するツールの選択における試行錯誤、2. はテキストデータの絞り込み方法における試行錯誤、3. は絞り込んだテキストを解釈するための試行錯誤となる。この試行錯誤の過程は、創造的思考における発散的思考に位置すると考えられる。

3.3.1 想定内外のツールの選択と組合せ

TETDMにおいては、使用する各パネルにおいて、処理ツールと可視化ツールを1つずつセットして利用する。そのため、各パネルにおける処理ツールと可視化ツールの組合せ、並びに、複数のパネルで用いるツールの組合せには自由度がある。

ここで、さまざまなツールを用いることで、多様な分析結果を獲得できると考えられ、用いるツールの組合せによって、想定されるツールの組合せから得られる知識、ならびに想定外のツールの組合せから得られる知識とに分けられる。

想定されるツールの組合せとは、「ユーザが予め指定するツールの組合せによる分析」または「ツールの作成者が予め指定するツールの組合せによる分析」を指す。すなわち、ユーザまたはツールの作成者があらかじめ利用をイメージしたツールの組合せによってツールを利用して、結果を獲得する場合となる。この場合、利用方法の見通しが立っているため、意味のある結果を得やすい反面、得られる結果が予想の範囲内にとどまることも多いと考えられる。

想定外のツールの組合せとは、「ユーザが想定していなかったツールの組合せによる分析」または「ツールの作成者が想定していなかったツールの組合せによる分析」を指す。TETDMにおいては、データ型コンバータ（可視化モジュールが受け取りを指定するデータ型のデータ以外でも、受け取れるデータ型に自動的に変換してデータの受け渡しを可能にする機能、表1）が実装されており、ツールの組合せの自由度は高くなっている。そのため、あるツールが開発されたときには存在しなかったツールと組み合わせる利用することも考えられ、そのような自由な組合せにおいては、得られる結果を誰も予想していない場合がある。

また複数のパネルを利用する場合、各パネルは基本的に独立になっているため、各パネルに任意のツール

表 1: データ型コンバートのルール

変換前, 後	boolean	int	double	String
boolean	-	0 か 1 に	0.0 か 1.0 に	文字列に
int	0 以外 true	-	(double) に	文字列に
double	0 以外 true	(int) に	-	文字列に
String	"" 以外 true	文字数	文字数	-

をセットして利用できる。そのため、複数のパネルに対するツールの組み合わせ方には限りがなく、むしろ想定外の組合せの方が多くなる。

これらの想定外のツールが組み合わせて利用される場合、意味のない結果が出る可能性もあるが、斬新さや面白さを備えた結果が得られる可能性も秘めており、多様な知見を得る上では一定の効果があると考えられる。

3.3.2 部分テキストの生成

入力したテキスト全体に関する知見を得るだけではなく、特定の条件を満たす一部のテキストに対する知見を獲得したい場合がある。特にデータ量が多くなるに連れ、一部のデータについての特徴を集めることが必要と考えられる。相関ルール分析等の一般的なデータマイニングにおいても、条件部の条件を満たす一部のデータについてのルールを出力した上で、人間が解釈を行っている。

そのため、着目すべき部分テキストを選択する試行錯誤には意味があり、部分テキストの選択肢を提供する支援には意味がある。TETDM においては、形式的には「一部の文あるいはセグメント (段落 or テキスト) を選択」することがテキストの絞り込み (部分テキストの生成) となる。

この一部の文あるいはセグメントを選択する基準として、例えば以下のものが挙げられる。

- 特定の単語 (話題) を含む : ex) 「研究」を含む文
- ある指標の評価値が高い : ex) 句読点が多い文
- 出現位置情報 : ex) 前半の文
- ユーザが指定する文

3.3.3 比較用データの生成

前節で生成した部分テキストに関する特徴を得るためには、単に部分テキストに対する処理結果を得るだけではなく、他の部分テキストに同様の処理を行った場合の結果や、テキスト全体に対する処理結果と比較することは有効と考えられる。すなわち、特徴を理解

表 2: TETDM モジュールのツールタイプ

タイプ名	モジュールの定義
シンプル	データのやりとりが不要で単独で完結する
プリミティブ	一種類のデータ型のデータを 1 つだけ渡す, または受け取る
セミプリミティブ	複数のデータを渡す/受け取るが, 指定モジュール以外との組合せでも動作する
特殊	データに特定の意味があり指定モジュール以外との組合せでは動作しない

したい部分テキスト A に対して、他の条件で絞り込んだ部分テキスト B, あるいはテキスト全体との比較を可能にする。

TETDM においては、複数のパネルを並べることができるため、例えば 3 つのパネルを並べて、それぞれで、部分テキスト A の処理結果、部分テキスト B の処理結果、テキスト全体の処理結果を並べることにしても一定の効果があると考えられる。

しかしその場合、並べられた結果の違いを目で比較する必要があり手間がかかったり、違いを見落とす可能性が生じる。そこでより積極的に比較を促すために、処理結果同士を比較した結果を生成して確認できるようにすることには意味がある。

TETDM では、処理ツールと可視化ツールが分離しているため、各部分または全体テキストに対する処理ツールの処理結果を集めて比較演算を行い、その結果を可視化ツールに渡すことで、既存の可視化ツールに手を加えることなく、比較データを生成して確認することが可能となる。ただし、現時点で利用することができる可視化ツールは、表 2 に示す TETDM の 4 種類のツールタイプのうち、データを 1 つだけ受け渡す「プリミティブ」タイプとしており、前述のデータ型コンバートと組み合わせることで、任意の「プリミティブ」タイプの処理ツールと可視化ツールを、比較用データの生成に用いることが可能となる。

比較演算としては、数値データが得られる場合であれば、減算, max, min などの演算を、文字列データが得られる場合であれば、一致比較 (両方に出現, 片方にのみ出現) などの演算を施すことが考えられる。

ここでさまざまな比較方法を用意することで試行錯誤を促し、部分テキストからの知見獲得を助ける。

3.4 結果の解釈

多くのマイニングソフトウェアでは、結果を出力した後は、積極的なユーザ支援を行っていない。高次元のデータを可視化ツール上で理解しやすく表現する方法によって、高次元データを一度に俯瞰してさまざまな知見の獲得につなげることを意図した研究 [小林 13]

はあるが、実際に知見を獲得する過程には踏み込んでいない。一般的な場面では、出力の中から着目する結果を選び、また得られた結果を解釈する際には、ユーザの頭の中、あるいはユーザが用いるテキストエディタなどの上で整理されることが多い。

そこでTETDMにおいては、以下の3つの点についてソフトウェア内で積極的に支援する方法を検討、実装していきたいと考えている。

1. 結果へのアノテーション付け（着眼）
2. 事実への意味付け（解釈）
3. 解釈の一般化（創発）

1. はユーザの着眼支援として、可視化結果の一部に着目を促す仕掛けの導入、2. は結果の解釈支援として、着目した結果に対する意味付けの支援、3. は解釈結果からの創発支援として、集めた解釈を一般化する支援を行う。この結果の解釈の過程は、創造的思考における収束的思考に位置すると考えられる。

3.4.1 結果へのアノテーション付け（着眼）

TETDMにおける可視化ツールの出力結果上に、結果として分かる事実に対して、解釈を与えるべきと考えられる箇所に「★」などの印を付けるアノテーション付けを行う機能を設ける。これにより、思考の材料の収集を支援する。

着眼と解釈の手順として、ひとつひとつの着眼と解釈をセットとして繰り返す方法も考えられるが、各着眼点を順次解釈していると、相対的に重要度の低い着眼点に時間がかけられることになったり、またその解釈を行っている間に、他の重要な結果を見落とす可能性も生じる。

そのため、出力結果を全体的に眺めた上で、解釈に値する結果、または解釈を与えるべきと考えられる出力に対して、一通り目印を付け、その後、次節で述べる解釈につなげることが望ましいと考えられる。

TETDMにおいて、具体的なアノテーション付けを実装する方法としては、1) 任意の座標にアノテーション付けを可能にする、2) 可視化ツール内の指定オブジェクトにアノテーション付けを可能にする、という2通りの方法を検討している。

3.4.2 事実への意味付け（解釈）

前節で着眼した結果に対して、その結果に対する意味付けにより解釈を与える支援を行う。着眼点を見ながら、直接その解釈を与えるフォームを用意することによって、直感的に思考できる環境を用意する。すな

わち、着目点が表す「結果」とその「解釈」をペアで入力できるフォームを用意する。

解釈の結果は文字情報として与えられるため、解釈時に分かっていたことでも、時間が経つと意味が分からなく可能性が生じる。そのため、解釈結果をTETDMの出力結果ならびに着眼点と関連づけて記録することで、解釈結果をもとに、その解釈を得るための出力を再現できるようにする。すなわち、各着目点の出力に用いた（部分）テキスト情報と使用したツールの情報は記憶しておき後で再現可能にする。

3.4.3 解釈の一般化（創発）

前節までの手順により得られた、解釈結果を総合的にまとめる支援を行う。複数の解釈結果を端的かつ簡潔に表す支援を行うことで、汎用的な知識の創発を支援する。すなわち、これまでに入力された「結果」と「解釈」のペアを分類、整理して複数の解釈を統合した解釈（知識）を生成する。

TETDM上に、解釈結果の集合を整理できるインタフェースを実装する。複数の「結果」と「解釈」のペアを自由に移動、配置できるようにした上で、関係がある複数のペアを組み合わせることで新たな「結果」と「解釈」を生成してもらうことで、複数の「結果」と「解釈」のペアを統合して、ペアの数を減らす。これを繰り返すことで、最終的に1つの「結果」と「解釈」のペアを作ることを目指す。

4 TETDMを用いた知識創発に向けた予備実験

本章では、前章で述べた知識創発の枠組みに対して、「部分テキストの生成」と「比較用データの生成」について簡易的な実装を行い、行った評価実験について述べる。

部分テキストの生成については、ユーザがフォームから入力した単語を含む段落の集合を部分テキストとする機能を実装した。

比較用データの生成については、実数値配列のデータが生成される場合に限定して実装を行い、実数値列を折れ線グラフとして表示する可視化モジュールを実装した。比較の方法は、二つの実数値列 A, B に対して、 $A - B$, $B - A$, $\max\{A, B\}$, $\min\{A, B\}$ の4つの比較方法のいずれかを選択し、比較用データを生成できる機能を実装した。

12名の大学生、大学院生に対して、26段落(3574字)からなる作家とアナウンサーによる対談テキストを用いて、テキストの分析を行ってもらう実験を行った。テキストを絞り込むための、部分的な話題を表す単語は、

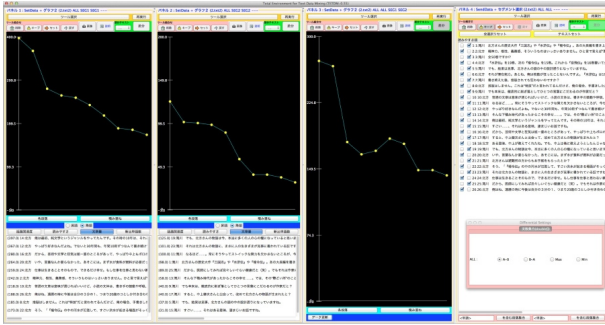


図 5: 知識創発に向けた実験で用いた TETDM の画面

予め 26 段落の 4 割から 7 割程度の段落に出現する単語 10 個を被験者に提示し、使用してもらった。図 5 に、被験者が用いた TETDM の画面を示す。被験者は右端のパネルで、単語を入力することができ、その単語を含む段落からなる部分テキストを 2 種類まで作成することができる。生成する比較データの種類の選択は、右端のパネル上にあるポップアップウィンドウ上で選択できる。左の 3 つのパネルにおいては、各段落の光と影システム [西原 09] による文章の話題関連度、各段落が含む全文字数に対する漢字の割合、段落の文字数、段落で初めて出現する単語数、の 4 つのデータのいずれかを出力する処理モジュールと、与えられたデータを折れ線グラフで表示できる可視化モジュールがセットされている。左の 2 つのパネルは、作成された 2 種類の部分テキストに関するデータを出力し、右から 2 番目のパネルは、左の 2 つの比較データを作成して表示する。

被験者にはまず、文章を読んでもらった後、文章全体についてのデータ（グラフ）をもとに、その文章全体について言える事柄を挙げて、予め用意したエクセルファイルのフォーマットに記入してもらった。その後、指定した個々の話題（単語）についてのデータ、また 2 つの話題の比較データをもとに、話題間の特徴を比較してもらい、再び文章全体について言える事柄を挙げて、エクセルファイルに記入してもらった。

それぞれの回答は上限を 5 個としており、前者で合計 44、後者で合計 48 の事柄が挙げられた。前者の中には、部分的な話題に言及した事柄は 4 つ（2 名）しかなかったのに対して、後者の中には、22(10 名)が部分的な話題に言及するとともに、そのほとんどが複数の話題を比較した内容を挙げていた。このことから、実装した内容が、部分的に扱われている複数の話題に関する新たな事実の列挙、ならびにそれらの解釈を支援する効果があることが確認された。

すなわち、同じ内容を問う課題に対しても、話題によって文章の絞り込みを行う機能（部分テキストの生成）、ならびにその解釈を支援する機能（比較用デー

タの生成）によって、最初とは異なる観点からの考えを得ることができた。そのため本枠組みによって、試行錯誤の幅を広げ、より汎用的な知識の創発につなげられる可能性がある。

5 結論

本稿では、統合環境 TETDM を用いた知識創発支援の枠組みについて述べた。知識創発の枠組みにおいては、試行錯誤と、結果の解釈の 2 つを主な観点として挙げ、そのそれぞれに対する実現の方法について述べた。試行錯誤における部分テキストの生成と比較用データの生成に関して簡易的な実装と評価実験を行い、試行錯誤の選択肢を広げることが、幅広い解釈につなげられる可能性を示唆した。

本稿の実験では、解答をエクセルファイルに記入してもらった形で行ったが、今後は可視化結果への着眼から知識創発に至る論理的解釈を支援する枠組みを実装し、その上で評価を行っていきたいと考えている。

参考文献

- [Guilford 59] Guilford, J. P. : Three faces of intellect, American Psychologist, Vol.14, No.8, pp.469 – 479 (1959)
- [小林 13] 小林弘明, 三末和男, 田中二郎: 色による特徴表現を用いた高次元データの可視化, 情報処理学会研究会報告, Vol.2013-HCI-152, No.23, pp.1 – 8 (2013).
- [西原 09] 西原陽子, 佐藤圭太, 砂山渡: 光と影を用いたテキストのテーマ関連度の可視化, 人工知能学会論文誌, Vol.24, No.6, pp.480 – 488 (2009).
- [砂山 11] 砂山渡, 高間康史, Danushka Bollegala, 西原陽子, 徳永秀和, 串間宗夫, 松下光範: Total Environment for Text Data Mining, 人工知能学会論文誌, Vol.26, No.4, pp.483 – 493 (2011).
- [砂山 13] 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).
- [砂山 14] 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).