人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第3回)
SIG-AM-03-01

# 価値観に基づくユーザモデルを用いた情報推薦手法に関する検討

服部 俊一[1*]　高間 康史[1]
Shunichi Hattori[1], Yasufumi Takama[1]

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

**Abstract:** 本稿では，ユーザの価値観に基づくユーザモデルを用いた情報推薦手法を提案し，その特性について検討する．本稿において価値観とは，ユーザがどの属性を重視してアイテムへの評価を決定するかという属性毎の価値判断，いわば「こだわり」を表す要素とする．価値観に基づくユーザモデルを内容ベースフィルタリングと組み合わせ効用ベース推薦を実現する手法を提案し，その特性について考察する．

## 1 はじめに

本稿では価値観に基づくユーザモデルを用いた情報推薦システムを提案し，その特性について考察を行う．情報化技術の発展による情報量の増大に伴い，利用者にとって有用な情報を見つけ出す情報推薦システムが情報フィルタリングの一手法として注目されている．しかし，代表的な手法である協調フィルタリング [11] を用いた既存の情報推薦システムでは，新規に利用を始めたユーザや最近追加されたアイテムに対しての情報の少なさから，推薦の精度が低くなってしまうという cold-start 問題が指摘されている [14]．一方，書籍であれば著者やジャンルなど，アイテムの属性値に関する情報をモデリングし推薦に用いる手法は内容ベースフィルタリング [9] と呼ばれ，特定の属性値に対する嗜好をユーザモデルとして構築できれば適切な推薦が可能である．嗜好情報はユーザが明示的に与える場合と，ユーザの行動情報からシステムが推定する場合があり，ユーザの負担の観点からは後者が望ましいとされる．しかし，ショッピングサイトなどで取り扱われている多様なジャンルに属するアイテムを推薦対象とした場合，膨大な嗜好・行動情報を獲得しなければ多くの属性値に対して十分な情報を集めることは困難である．

一方，個人の嗜好や消費行動を推定するための概念として「価値観（Personal Values）」が挙げられる．価値観は個人の嗜好や消費行動に間接的な影響を与える要素であり，マーケティングや商品開発の分野では広く活用されている．価値観はアイテムの好き嫌いや良し悪しではなく，どの要素を重視するかという，アイテムの属性に対する価値判断を表す要素であることから，これを用いることでより少ない情報でユーザの嗜好や特性を推論することが可能になると考える．しかし，ユーザの価値観のモデル化およびそれに基づく情報推薦システムまだ確立されていない．

本稿では価値観と繋がりの深い要素としてユーザの「こだわり」に着目した情報推薦システムを提案する．前述のように，ユーザは自分の持つ価値観に基づき重視する属性について評価を行い，最終的にそのアイテムを受け入れるかどうかを決定すると考えられる．このような属性に対する価値判断はユーザの「こだわり」と表現することもできる．ユーザがアイテムのどの要素を重視しているかを推論することによってユーザの価値観をモデリングし，それに基づいてユーザの価値判断を反映させたアイテムを推薦する．本稿では，属性に対する評価がアイテムの評価に与える影響を測る指標である評価一致率 [3] に基づき，ユーザが重視する属性の集合としてユーザモデルを構築する．得られたユーザモデルを内容ベースフィルタリングと組み合わせ効用ベース推薦 [5] として用いることで，従来よりも少ない情報でユーザの価値判断に合致したアイテムを推薦するシステムの実現が期待できる．本稿ではユーザのこだわりを反映したユーザモデルを用いた情報推薦システムの概要およびシステム構成について述べ，その特性について考察する．

## 2 関連研究

### 2.1 情報推薦手法

情報推薦を行うためにはユーザやアイテムの特性をモデリングし，その結果に基づき推薦対象となるアイテムをフィルタリングする必要がある．既存の情報推薦

*連絡先：首都大学東京大学院
システムデザイン研究科情報通信システム学域
〒 191-0065 東京都日野市旭が丘 6-6
E-mail: shattori@krectmt3.sd.tmu.ac.jp

手法の多くは協調フィルタリング（Collaborative Filtering）と内容ベースフィルタリング（Content-Based Filtering）に分類することができる [12]．それぞれの手法の特徴について以下に述べる．

#### 2.1.1 協調フィルタリング

協調フィルタリングは多くのユーザの嗜好情報を過去の行動という形で記録し，そのユーザと嗜好の類似した他のユーザの嗜好情報を用いてユーザの嗜好を推測する手法である [11]．協調フィルタリングの利点は，アイテムの属性情報がなくても推薦が行えること，および処理が手軽であることであり，これらの理由からショッピングサイトなどで現在最も広く利用されている手法である．一般に，協調フィルタリングにより精度の高い推薦を行うためには，多数のユーザに情報推薦システムが利用され，多くのアイテムに関する行動情報が収集可能であることが必要となる．そのため，推薦システムを新たに利用し始めたユーザや新規に追加されたアイテムに対しては行動情報の少なさから類似する嗜好を持つユーザを発見できず，推薦の精度が低くなってしまうという欠点がある．これは cold-start 問題 [14] と呼ばれる．また，推薦対象として膨大なアイテムが存在し，その多くにおいてユーザとの関係が希薄である場合，ユーザに関する情報が十分確保されていても精度の高い推薦を行うことは困難である．これは sparsity 問題 [7] と呼ばれ，cold-start 問題と併せて協調フィルタリングを用いた情報推薦手法全般に共通する課題とされている．

協調フィルタリングにおいて，cold-start 問題および sparsity 問題に関する研究は広く行われている．代表的な手法として，嗜好パターンが類似するユーザをモデル化することで精度の向上を実現するモデルベース法が挙げられ，Breese らはクラスタモデルを用いて実装している [1]．モデルベース法以外のアプローチもいくつか行われており，Park らはユーザに加えて filterbot と呼ばれるロボットがアイテムへの評価を行うことで，ユーザ・アイテムの情報が少ない状態でも推薦に必要な情報を収集可能にする手法を提案している [10]．Lee らはユーザ同士の友人関係を類似度として利用することで，ユーザのアイテムの関係が希薄になる状態の改善を試みている [7]．また，Yildirim らはランダムウォークを用いてユーザ・アイテム間の類似度を求め，アイテムの推薦を行う手法を提案している [17]．これらはユーザの行動情報の少なさを友人関係などの間接的な情報で補うアプローチであるが，ユーザの嗜好や価値判断といった，ユーザの意思決定に直接関わる要因とは性質が異なると考える．

#### 2.1.2 内容ベースフィルタリング

内容ベースフィルタリングはアイテムの内容とユーザの嗜好情報を比較し，その関連度に基づいてフィルタリングを行う手法である [9]．アイテムの内容には評価のポイントとなる属性の値が用いられ，本稿ではこれを属性値と呼ぶ．例えば映画では監督や俳優の名前，ジャンル（アクションやコメディなど）が属性値となる．内容に基づくフィルタリングは協調フィルタリングと比べ，システムを使い始めたばかりのユーザでも特定の属性値に対する嗜好情報が得られれば精度の高い推薦が可能であるという利点があり，楽曲推薦などで活用されている [18]．しかし，ショッピングサイトなど多様なジャンル・アイテムを取り扱う際には，アイテムの属性値は膨大なパターンが存在するため，ユーザとアイテムの属性値との関係が希薄になってしまうケースも多いと考える．そのため，多様なジャンル・アイテムを推薦対象とした場合，多くの属性値に対して推論を行うのに十分な情報を集めることは困難であり，協調フィルタリング同様に cold-start 問題および sparsity 問題が課題となる．

内容ベースフィルタリングにおいてこれらの問題に取り組んでいる研究として，関らのコンテキストを考慮した飲食店推薦システムが挙げられる [15]．情報推薦のための行動情報取得はシステムの利用ログなどから自動的に行動情報を収集する暗黙的手法と，アイテムや属性値に対する好き嫌いをユーザに直接回答してもらう明示的手法の 2 つに分類することができる [6]．暗黙的手法はユーザの負担が低いというメリットがあるが，対象ユーザの行動情報を十分収集する必要がある．これに対し，関らは事前に蓄積されたアイテムのコンテキスト情報から，求めるコンテキストをユーザが明示的に指定することで新規ユーザでも適切な推薦結果が得られるとしている．しかし，アイテムに関しては事前に十分な量の行動情報を獲得する必要があることから，新規アイテムを適切な推薦対象として扱うことは難しいと考える．ユーザ・アイテム双方に対して cold-start 問題および sparsity 問題を解決するためには，少量の行動情報から推薦アイテムを獲得する手法が必要と考える．そこで本稿では，次節に述べる価値観に着目する．

### 2.2 価値観に基づく嗜好・消費行動の推論

価値観は消費者の嗜好や行動に強く影響を及ぼすと考えられており，マーケティングの分野では古くから利用されている．Rokeach は消費者の嗜好に関わる価値観を 18 の要素に分類した Rokeach Value Survey [13] と呼ばれる調査方法を提案し，多くの調査で利用されている．Vinson らは，保守的な価値観を持つ大学と革

新的な価値観を持つ大学，それぞれに所属する学生の間に有意な嗜好の差があることをアンケート調査により明らかにしている [16]．近年でも，Holbrook が消費・購買行動に影響を与える価値観を 8 つに分類する [4] など，消費者の嗜好と価値観は関連の深いテーマとして研究および調査が進められている．Web インテリジェンスの分野においても，価値観はユーザの嗜好と関連の深い要素として利用されている．宮尾らはアンケートによりユーザの価値観を調査し，ユーザが持つ価値観に最適化された機能を持つ SNS プロトタイプを構築している [8]．また，Hattori らは，ユーザの嗜好と価値観の関連をアンケートにより調査し，情報推薦への適用可能性について考察している [2]．

価値観を情報推薦システムに適用することを考えた場合，価値観はユーザがアイテムのどの要素を重視して評価を決定しているかという，アイテムの属性に対する価値判断，いわば「こだわり」を表す要素であると言える．例えば映画の場合，アイテムの属性としてストーリーや監督，出演俳優などが挙げられる．ストーリーに対して強いこだわりを持つユーザの場合，俳優や監督に対する良し悪しは映画の全体的な評価にあまり影響を与えず，主にストーリーを好むかどうかが評価に強く影響すると考えられる．このような属性に対する価値判断をモデリングすることができれば，より少ない情報でユーザの嗜好や特性を推論することが可能になると考える．

## 3　価値観に基づく情報推薦システム

本節では，価値観と繋がりの深い要素としてユーザの「こだわり」に着目した情報推薦システムの概要について述べる．価値観は物事の優先順位や重み付けを表すものであることから，本稿では情報推薦における価値観を，図 1 に示すように「どの属性を重視してアイテムの評価を決定するか」を判断するための基準として定義する．この属性に対する価値判断はユーザの「こだわり」と表現することができる．従来手法では著者の名前やアクションなどのジャンル名といった属性値に対する好みからモデリングを行うが，提案手法ではアイテムの属性に対するこだわりの強さをモデリングする．ユーザが強いこだわりを持つ属性ほど，その属性に対する評価は安定してアイテムの評価に影響しており，少数の評価情報から適切な推薦が可能になると考える．

提案手法が従来の内容ベースフィルタリングと異なる点は，以下に示す 3 点に分類できる．

**（1）属性の利用：** 著者の名前などの属性値ではなく,「著者」「ジャンル」といった属性をモデリングに用いる．内容ベースフィルタリングで多様なアイテムを取り扱う場合，大量に存在する属性値に対して十分な量の評価を収集することは困難であり，特に新規ユーザにおいて属性値との関係が希薄になってしまうと考えられる．属性数は属性値数よりはるかに少数であるため，新規ユーザに対しても各属性に対して推薦に必要な量の評価を収集することができると考える．

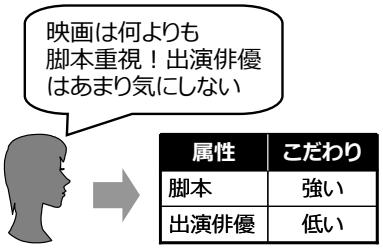


図 1: 属性に対するこだわりの例

**（2）評価に与える影響度を推論：** アイテムの内容に関する好き嫌いではなくアイテムの評価に与える影響度を推論する．内容ベースフィルタリングでは暗黙的に評価を収集することが多いが，あるアイテムを好むからといってそのアイテムの属性値全てに満足しているとは限らず，属性値レベルでは間違った推論を行ってしまう可能性がある．提案手法では，価値観に基づいてアイテムへの評価に強い影響を与える属性を推論することで，ユーザの価値判断に合致するアイテムをより短期間で取得可能になると考える．

**（3）ユーザへの負荷軽減：** 従来の内容ベースフィルタリングで明示的に情報収集を行う場合，大量の属性値が存在することから全ての値に対して評価を問うことはユーザにとって負担が大きい．(1) でも述べたとおり，属性数は属性値と比較してはるかに少数であるため，推薦システムとの対話により属性に対する評価を収集する場合においても，ユーザに大きな負担をかけることなく情報を収集することが可能になると考える．

属性に対するユーザのこだわりを情報推薦システムに用いることで，内容ベースフィルタリングにおけるモデリングへの貢献が期待できる．すなわち，上記 (2) で述べたようにユーザがこだわりを持つ属性に絞ってユーザの評価を収集することで，より少数の情報から嗜好情報を獲得可能となることが期待できる．また，(3) で述べたように属性値について全て評価を行うことはユーザに対する負荷の観点から困難であるが，属性値よりも少数の属性に対してはそれほど負荷をかけずに評価可能と考える．加えて，効用ベース推薦における貢献度の暗黙的取得も期待できる．効用ベース推薦では，

各属性のとる値がアイテムの評価に与える影響を効用関数としてモデル化する [5]．しかし効用関数を明示的に獲得することはユーザに高い負担を課すことに繋がり，暗黙的に算出する場合でも多くのシステム利用履歴が必要となる．この問題に対し，提案手法で用いるユーザモデルはユーザが重視する属性を表したものであるため，効用関数の代替として利用可能と考える．

## 3.1 推薦システム構成

図 2 に示すように，本システムは「評価抽出モジュール」「ユーザモデリングモジュール」「情報推薦モジュール」と呼ぶ 3 つのモジュールから構成される．それぞれのモジュールの概要について以下に述べる．

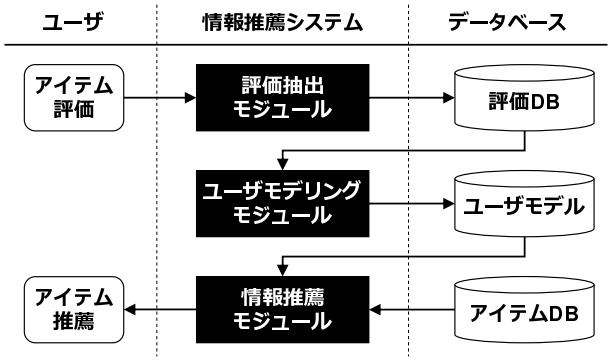


図 2: 情報推薦システム構成図

### 3.1.1 評価抽出モジュール

本モジュールでは，システム上でユーザが行った評価（アイテムの総合評価および各属性への評価）を取得し，評価 DB に保存する．評価 DB はユーザ毎に作成され，ユーザの評価としてアイテム及びアイテムの属性に対する評価極性（好評または不評）を保持する．

また，本システムにおいて評価・推薦対象となるアイテムは映画を対象とし，Wikipedia 日本語版[1]の情報を LOD (Linked Open Data) 形式で公開している DBpedia[2]を情報源としてアイテム DB を構築する．本アイテム DB では表 1 に示すように，アイテム毎に属性および属性値を保持する．属性は「ジャンル」「出演俳優」などのように評価や好みのポイントとなる観点をあらわし，表 1 に示すようにあらかじめ決められた 6 つの要素からなる．属性値は「ジャンル」であれば「アクション」や「サスペンス」,「出演俳優」であれば俳優の名前のように，各属性がとる値を表す．

[1]http://ja.wikipedia.org/
[2]http://ja.dbpedia.org/

表 1: アイテムが持つ属性・属性値

| 属性 | 属性値 |
|---|---|
| ジャンル | ジャンル名（サスペンス等） |
| 監督 | 製作総指揮, 監督名 |
| 出演俳優 | 俳優名 |
| 脚本 | 脚本家名 |
| 演出 | 演出家, 映像監督名 |
| 音楽 | 作曲家名 |

### 3.1.2 ユーザモデリングモジュール

本モジュールでは，アイテムの属性に対するユーザのこだわりを評価一致率 RMRate (Rating Matching Rate) [3] と呼ぶ指標を用いてモデリングする．ユーザのこだわりは，ある属性に対する評価がアイテム全体に対する評価に与える影響の度合いに表れると考える．そこで，アイテムに対する評価極性に加えて各属性に対する評価極性を抽出して属性毎にアイテムの評価に与える影響度を評価一致率として算出する．ユーザ $u$ がアイテム $i$ に対して行った評価 $e_{ui} \in E_u$ において，あるアイテム $i$ の極性 $p_{item}(u,i)$，および $i$ の属性 $j$ の極性 $p_{attr}(u,i,j)$ が一致するかどうかを調べ，一致する評価の回数（アイテムの個数）を $O(u,j)$，一致しない回数を $Q(u,j)$ とする．この時，ユーザ $u$ における属性 $j$ の評価一致率 $P(u,j)$ は式 (1) で算出される．これにより，ユーザのこだわりを表すユーザモデルは属性数を $m$ とすると $m$ 次元のベクトルとして表される．

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

作成するユーザモデルでは，あるユーザが行った評価からそれぞれの属性に対する評価一致率を計算し，属性ごとに保持する．例として，ユーザがある 2 種類のデジタルカメラに対して評価を行った結果を表 2 に示す．また，表 2 の評価に基づいて属性ごとに評価一致率を計算した例を表 3 に示す．表 3 の計算例に示す属性「操作性」「バッテリー」のような評価一致率が高い属性はユーザが強いこだわりをもっており，アイテムの評価に影響を与える「推薦時に重要度の高い属性」であると推論される．一方で，評価一致率はアイテムおよび属性での評価極性が一致するかどうかを示すものであることから，評価一致率が 0.5 前後，またはそれ以下の属性に対してユーザが持つこだわりは弱いと考えられる．そのため，表 3 の「デザイン」「画質」はアイテムの評価にそれほど影響を及ぼさず,「推薦時に重要度の低い属性」であると推論される．

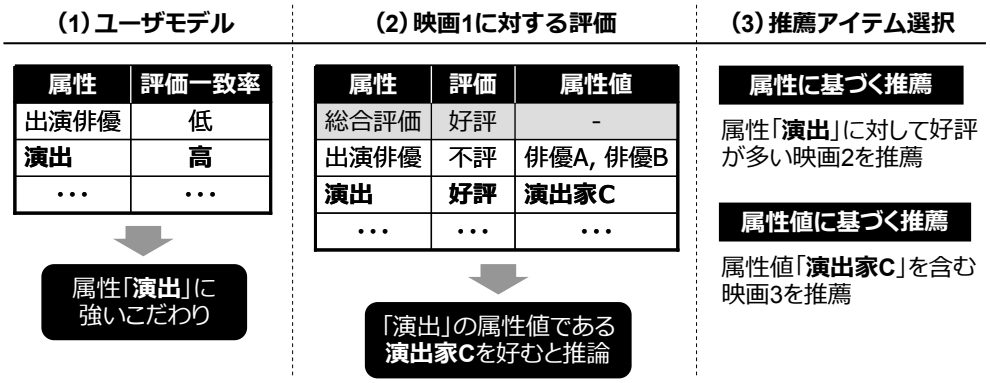


図 3: ユーザモデルを用いた推薦アイテムの選択例

表 2: アイテムへの評価例

**(1) アイテムAへの評価**

| 属性 | 極性 |
|---|---|
| 総合評価 | 好評 |
| 監督 | 不評 |
| 出演俳優 | 不評 |
| 脚本 | 好評 |
| 音楽 | 好評 |

**(2) アイテムBへの評価**

| 属性 | 極性 |
|---|---|
| 総合評価 | 不評 |
| 監督 | 好評 |
| 出演俳優 | 不評 |
| 脚本 | 不評 |
| 音楽 | 不評 |

表 3: 評価一致率の計算例

| 属性 | 一致 | 不一致 | 評価一致率 |
|---|---|---|---|
| 監督 | 0 | 2 | 0.00 |
| 出演俳優 | 1 | 1 | 0.50 |
| 脚本 | 2 | 0 | 1.00 |
| 音楽 | 2 | 0 | 1.00 |

#### 3.1.3 情報推薦モジュール

本モジュールでは，ユーザモデリングモジュールで作成されたユーザモデルを用いて推薦アイテムを選択し，ユーザへ提示する．提案する推薦手法は属性に基づく推薦と属性値に基づく推薦の 2 つに大別される．

属性に基づく推薦では，ユーザが強いこだわりを持つ属性に対して多くのユーザから「好評」と評価されているアイテムを評価 DB から検索し，推薦対象として選択する．例えば，図 3(1) に示すユーザモデルのように属性「演出」に高い評価一致率を持つユーザの場合，同じ属性である「演出」に対して多くのユーザが高い評価をしているアイテムを推薦対象として選択する．

一方，属性値に基づく推薦では，ユーザが好評と評価したアイテムの属性に対して強いこだわりを持っている場合その属性が持つ値を好むと推論し，その属性値を含む他のアイテムを推薦候補として選択する．例えば，図 3(1) に示す属性「演出」に強いこだわりを持つユーザが，図 3(2) のようにあるアイテムの属性「演出」に対して好評と評価した場合，そのアイテムの属性値である演出家 C が関わっている他の映画 3 を推薦対象として選択する．従来の内容ベースフィルタリングでは属性値に対する評価を暗黙的に収集することから，ユーザが好むと評価したアイテムの全属性が持つ値を好むと推論する．そのため，ユーザが好まない属性値も好むと推論されてしまい，間違った推薦結果をユーザに提示してしまう可能性がある．提案するユーザモデルにおいて高い評価一致率を持つ属性はアイテムの評価に与える影響が強いことから，ユーザが高い評価一致率を持つ属性に対して「好評」と評価した場合は，アイテムへの総合評価に関わらずその属性に含まれる値を好む可能性が高いと推論することができる．このような推論により総合評価が好評・不評どちらのアイテムからもユーザの嗜好情報を獲得することが可能となり，より少数の情報から適切な推薦アイテムの推定が可能になると考える．

## 4 おわりに

本稿では，ユーザがどの属性を重視してアイテムへの評価を決定するかという属性毎の価値判断，いわばユーザの「こだわり」に基づく情報推薦システムを提案し，従来手法との特性の違いについて考察した．また，現在開発を進めている推薦システムについてその概要およびシステム構成を示した．今後は提案したユーザモデルに基づく情報推薦システムの実装を進め，被験者実験により価値観に基づく情報推薦手法の有用性を実証する．

また，本稿では価値観のモデリングのためユーザがこだわりをもつ属性を推論する手法を提案したが，今

後はユーザのこだわりに加えてその評価傾向についても分析を行う．全ての属性をバランス良く評価するユーザや特定の属性に強いこだわりを持つユーザなど，評価傾向に基づきユーザをいくつかのタイプに分類することで，ユーザに対して新たな着眼点の提示や意外性のあるアイテムの推薦が可能になるのではないかと考える．

## 参考文献

[1] J. S. Breese, D. Heckerman, and C. Kadie, “Analysis of Predictive Algorithms for Collaborative Filtering,” Uncertainty in Artificial Intelligence 14, pp. 43-52, 1998.

[2] S. Hattori and Y. Takama, “Investigation about Applicability of Personal Values for Recommender System,” Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol. 16, No. 3, pp.404-411, 2012.

[3] S. Hattori and Y. Takama, “User and Item Modeling Methods Using Customer Reviews towards Recommender System Based on Personal Values,” 2012 International Workshop on Intelligent Web Interaction (IWI-2012), S3204, 2012.

[4] M. B. Holbrook, “Consumer value: a framework for analysis and research,” Routledge, 1999.

[5] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich, 田中克己（訳）, 角谷和俊（訳）, 情報推薦システム入門 -理論と実践-，共立出版，東京，pp.98-103, 2012.

[6] 神嶌 敏弘, 推薦システムのアルゴリズム (2), 人工知能学会誌 23 巻 1 号, pp.89-103, 2008.

[7] S. Lee, J. Yang and S. Y. Park, “Discovery of Hidden Similarity on Collaborative Filtering to Overcome Sparsity Problem,” Proceedings of discovery science: 7th international conference, DS 2004, (LNAI 3245), pp. 396-402, 2004.

[8] 宮尾 和樹, 原田 利宣, SNS サイトの分類とユーザの価値観に基づくプロトタイプの構築, デザイン学研究, 55 巻 1 号, pp.81-90, 2008.

[9] F. Pachet, P. Roy nad D. Cazaly, “A Combinatorial Approach to Content-based Music Selection,” Proceedings of IEEE Multimedia Computing and Systems International Conference 1999, pp. 457-462, 1999. IEEE Multimedia, vol. 7, pp. 44-51, 2000.

[10] S. T. Park, D. Pennock, O. Madani, N. Good and D. DeCoste, “Naive filterbots for robust cold-start recommendations,” KDD ’06 Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 699-705, 2006.

[11] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “GroupLens: An Open Architecture for Collaborative Filtering of Netnews,” Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, pp.175-186, 1994.

[12] D. Riecken, “Personalized Views of Personalization,” Communications of the ACM, Vol. 43, No. 8, pp. 26-28, 2000.

[13] M. Rokeach, “The Nature of Human Values,” New York: The Free Press, 1973.

[14] A. I. Schein, A. Popescul, L. H. Ungar and D. M. Pennock, “Methods and metrics for cold-start recommendations,” 25th Annual ACM SIGIR Conference, pp. 253-260, 2002.

[15] 関匠吾, 中島伸介, 張建偉, アイテム利用時のユーザコンテキストを考慮した情報推薦システムの提案, 第 3 回データ工学と情報マネジメントに関するフォーラム（DEIM 2011）, B1-1, 2011.

[16] D. E. Vinson, J. E. Scott, and L. M. Lamont, “The role of personal values in marketing and consumer behavior,” The Journal of Marketing, Vol. 41, No. 2, pp. 44-50, 1977.

[17] H. Yildirim and M. S. Krishnamoorthy, “A random walk method for alleviating the sparsity problem in collaborative filtering,” RecSys ’08 Proceedings of the 2008 ACM conference on Recommender systems, pp. 131-138, 2008.

[18] 吉井和佳, 後藤真孝, 音楽推薦システム, 情報処理, 50 巻 8 号, pp. 751-755, 2009.

# 「動向に関する問い」を対象タスクとした コンテクスト検索の提案

## Proposal of Context Search Engine Focusing on Trend-related Queries

加藤 優[1]　桑折 章吾[2]　高間 康史[1,2*]

Yu Kato[1], Shogo Kori[2], Yasufumi Takama[1,2]

[1] 首都大学東京大学院システムデザイン研究科
[1]Graduate School of System Design, Tokyo Metropolitan University
[2] 首都大学東京システムデザイン学部
[2]Faculty of System Design, Tokyo Metropolitan University

**Abstract:** 本稿では，「動向に関する問い」を対象タスクとしたコンテクスト検索を提案する．既存の検索エンジンは汎用的に利用可能な反面低機能なため，情報要求をクエリに分解するのに要するユーザの負担は大きい．本稿で提案するコンテクスト検索は，タスクを限定することで高度な検索機能を提供する．動向に関する問いは広く一般に見られるものであり，提案手法は幅広いドメインに貢献することが期待できる．

## 1　はじめに

本稿では，「去年流行したアイテムは？」や「東日本大震災の影響を受けたアイテムは？」といった「動向に関する問い」に対して行う検索をコンテクスト検索と定義し，これに適した基本検索機能を提供する次世代検索エンジンについて提案する．

現在，検索エンジンを利用した情報収集・分析作業はドメイン・タスクを問わず広く一般に行われているが，既存検索エンジンが提供する機能と，ユーザの情報収集目的との乖離が大きいという問題がある．すなわち，既存検索エンジンが提供するのは，キーワードベースの検索要求指定，ページ単位での結果出力といった低機能にとどまったままであり，情報要求をキーワードに分解するのに要するユーザの負担が大きいと考える．

次世代検索エンジンの実現に向けて，自然言語文での問い合わせを受付け，ユーザの問いに直接回答するような検索エンジンの知的化のアプローチも考えられるが，本稿では検索エンジンが提供する基本検索機能を見直すことにより，ユーザの情報要求とのギャップを小さくするアプローチを採用する．基本検索機能として，「動向に関する問い」というタスクに着目する．近年，人気や流行といったアイテムの動向に関する問いは一般的なものと考える．

検索エンジンの知的化において，十分な性能を得るためには対象ドメインを限定する必要があると考えられるのに対し，本稿ではドメインに依存しないタスクを対象とすることにより，広く一般的に利用可能な検索エンジンの実現を目指す．現在の検索エンジンがユーザを限定せず，日常的に用いられる存在である以上，対象ドメインを限定しない本稿のアプローチは，次世代検索エンジン実現において重要な視点と考える．

本稿では，コンテクスト検索のコンセプトについて提案すると共に，現在構築中のプロトタイプシステムについて述べる．web で入手可能な動向情報は，検索エンジンでの検索数やヒット数などに表れる主観的動向情報と，官公庁を含めた様々な組織・機関が公開する価格や生産量のデータ，統計データなどの客観的動向情報に大別できる．本稿では，それらの動向情報を Web 上から抽出し，データベースを構築する．システムが提供する基本検索機能として，「指定アイテムに関する動向情報のピーク時期検索」，「指定期間に動向情報のピークを持つアイテム検索」を提案する．構築したプロトタイプシステムを用いて検索を行った事例を示す．

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

# 2 関連研究

## 2.1 次世代検索システムへの試み

Web が普及してから 20 年弱が経過し，Web 上には膨大な量の情報が蓄積されている．現在，最も用いられている情報検索手法は，検索エンジンを利用する方法である．しかし，既存の検索エンジンによるキーワードベースの検索は，ユーザが入力したキーワードを含むページを探すという低機能なものにとどまったままであり，情報要求をキーワードに分解する際のユーザの負担が大きいという問題がある．このようなユーザへの負担を軽減するために次世代検索システムの開発・研究がなされている[2][5].

亀井ら[2]は，Web 上に存在するソフトウェア開発に関する知見や情報を検索するための検索エンジン構築を提案している．多くのソフトウェアが開発されているが，それらの知識は必ずしも有効に蓄積・利用されていないために，似たようなソフトウェアが開発されていたり，同じようなミスでソフトウェア開発が滞ることがある．それらの問題を解決するため，巡回ロボットにより，Web 上に存在するソフトウェア資源を収集し，ソフトウェアメトリクスやパッケージ名，クラス名などの指定によりユーザに適切な情報を提供する検索エンジンを構築している．

小久保ら[5]は，新たな専門検索エンジンの構築手法として，「検索隠し味」を用いる方法を提案している．検索隠し味とは，機械学習の一種である決定木学習アルゴリズムを元に，Web ページ集合から抽出したブール式であり，ユーザの入力クエリに加えることで，汎用検索エンジンの検索結果をある特定ドメインに特化させることが可能となる．

これらを含めた多くの次世代検索システムの研究では，ドメインを狭い領域に限定することで検索性能の向上を図っている．自然言語によるクエリを受け付ける検索エンジンも次世代検索エンジンの一つとみなせるが[1]，この場合も性能向上のためにはドメインの限定が必要になると考える．これに対し，本稿で提案する検索システムでは，ドメインに依存しないタスクを対象とすることにより，広く一般的に利用可能な検索エンジンの実現を目指す．

## 2.2 動向情報に着目をした研究

動向情報とは，ある商品の価格や売上げの状況，ある会社の業績状況，内閣や政党の支持状況などの時系列データを基として，その変化を通時的にとらえつつ，それらを総合的にまとめ上げることで得られるものである[3]．これら動向情報は，様々なタスク・ドメインにおいて意思決定の材料として用いられており，世の中の社会活動に深く関わっている．近年，官公庁を含めた様々な組織・機関による情報公開が進み，Web 上には，多種多様で大規模な動向情報が蓄積されている．この流れは，今後も益々進んでいくことが予想される．このような背景から動向情報を利用した研究が多くなされている[4][6][7].

松下ら[6]は，動向情報テキストを視覚情報として要約することを目的として，テキストに含まれる情報を用いてグラフを描画する方法を提案している．テキスト中の明示的かつ定量的な数値情報に加えて，テキスト中で暗示されている情報を比較表現や背景知識によって抽出することで，より多くのプロットが可能となる．また，テキストに出現する「安定」や「緩やかな増加」などの定性表現を用いてグラフ概形を示唆するアノテーションをグラフに貼り付けることで，動向の理解を支援している．

山本ら[7]は，ユーザが指定した動向情報と多様な動向情報間の関連度を計算することで，関連する単語と，その動向情報を効率的に獲得する手法を提案している．山本らが提案するシステムを用いることにより，「ある会社の株価の変動と同期している株価をもつ会社を探したい」や，「ある製品の売上げの変動とともに使用されるようになった単語を知りたい」といった問いに答えることができる．

# 3 動向情報を対象とした コンテクスト検索システム

## 3.1 システム構成

提案するコンテクスト検索システムの構成を図 1 に示す．提案システムでは，Web 上から抽出した動向情報を事前に抽出し，データベースに格納しておく．データベース管理システム（DBMS）には MySQL を利用し，Web サーバの実装には Webrick を用いている．Web アプリケーションフレームワークには，Ruby on Rails を使用した．

動向情報は，検索エンジンの検索数やヒット数などの主観的動向情報と，アイテムの価格や生産量データ，統計データなどの客観的動向情報に分けられる．3.2 節，3.3 節に主観的動向情報および客観的動向情報の抽出手法をそれぞれ示す．

あるアイテムに関する動向を調査する際には，アイテムの人気や流行に応じて変動する動向情報において，その変動の最大値の検索が重要であると考える．そのため，本稿で紹介するプロトタイプシステムでは「指定アイテムに関する動向情報のピーク（最

大値）時期の検索」,「指定期間に動向情報の最大値を持つアイテムの検索」の 2 つを基本検索機能として実装している.

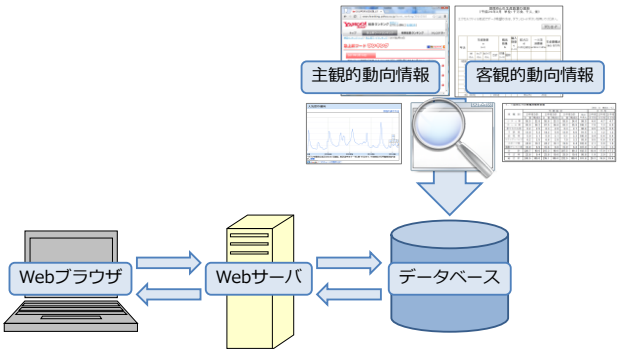


図 1　コンテクスト検索システムの構成

## 3.2 主観的動向情報抽出

主観的動向情報とは，検索エンジンの検索数やヒット数，ブログの記事数などのユーザの興味や関心に基づいて値が増減する情報であり，それらを Web から抽出する．現在，抽出対象としている主観的動向情報とその情報源を表 1 に示す.

表 1 抽出対象の主観的動向情報

| 分類 | 情報源 | URL |
|---|---|---|
| 検索数（指数） | Google Trends | http://www.google.com/trends/ |
| ヒット数 | Yahoo!検索（ウェブ検索） | http://www.yahoo.co.jp/ |
| ブログ記事数 | Yahoo!検索（ブログ検索） | http://search.yahoo.co.jp/blog |
| 急上昇ワードランキング | Yahoo!検索ランキング | http://searchranking.yahoo.co.jp/ |
| きざしランキング | kizasi.jp | http://kizasi.jp/ |
| HOT ワード | ついっぷるトレンド | http://tr.twipple.jp/ |

検索数は，検索数の推移を調査することができるサービスである Google Trends[1]から取得している．Google Trends で取得できる値は，各単語が Google で検索された回数を 1 週間単位で集計し，検索された総回数に対する相対値を 0～100 の指数で表したものである．ヒット数・ブログ記事数は，Yahoo!検索サービスにおいてウェブ検索・ブログ検索を利用して検索した際の検索結果件数を取得している．急上昇ワードランキングは，Yahoo!JAPAN が運営する Yahoo!検索ランキング[2]，きざしランキングは kizasi.jp[3]，HOT ワードは，Twitter 話題ランキングサイトのついっぷるトレンド[4]がそれぞれ提供しているランキング結果を Web スクレイピングによって Web ページから抽出している.

## 3.3 客観的動向情報抽出

客観的動向情報とは，販売量や売上高のデータ，統計データなどの定量的な測定が可能な情報であり，これらも Web から抽出する．主観的動向情報と異なる点として，これらのデータは集約されておらず，各企業・団体などでそれぞれ公開されている点，その公開形式も様々である点が挙げられる．一般的な公開形式として，Web ページに HTML で直接記載されている他，CSV・PDF・Excel などが用いられる.

HTML から情報を抽出する場合には，Ruby のライブラリである nokogiri を用いて HTML 解析を行う．多くの Web ページ内では，数値情報は表形式となって表されているため，HTML の<table></table>タグで囲まれた箇所から，各セルを意味する<td></td>タグ内の情報を抽出する.

Excel・CSV 形式の場合は，ファイルをダウンロードし，数値などの重要な情報が記載されているセルから情報を抽出する.

PDF の場合には，PDF ファイルの全文をテキストファイルに変換可能なツールである xdoc2txt[5]を用いてテキストファイルに変換し，不要な情報を除去して情報を抽出する.

前述の通り，客観的動向情報は多くの Web サイトに分散して存在するため，網羅的な収集は困難である．現状では，野菜や即席めんなどの価格や生産量などに関する情報を中心に 31 種類の客観的動向情報を収集しているが，今後も拡充していく予定である.

# 4　提案システムを用いた検索事例

プロトタイプシステムを用いて，想定する検索タスクについて，検索を行った事例を紹介する．プロトタイプシステムでは，主観的動向情報として Yahoo!検索や Google Trends など 3.2 節に示した 6 つの情報源から取得した動向情報を，客観的動向情報

[1] http://www.google.com/trends/

[2] http://searchranking.yahoo.co.jp/
[3] http://kizasi.jp/
[4] http://tr.twipple.jp/
[5] http://www31.ocn.ne.jp/~h_ishida/xdoc2txt.html

として 3.3 節に示した抽出方法によって，統計局や産業振興協会など 7 つの Web サイトから取得した 31 種類の動向情報をそれぞれデータベースに格納している．3.1 節で述べたように，プロトタイプシステムでは基本検索機能として「アイテムから探す」と「期間から探す」の 2 つを提供しており，ユーザは自身の「動向に関する問い」を，これらの基本検索，および既存検索エンジンへのクエリに分解して調べることを想定している．

提案システムの入力画面を図 2 に示す．上部のラジオボタンによって「アイテムから探す」，「期間から探す」を選択可能である．「アイテムから探す」を選択した場合は，検索ボックス内に検索したいアイテム名を入力することで，指定したアイテムに関する主観的・客観的動向情報の最大値およびその時期，情報を公開している Web サイトの URL，動向情報の変化を表したグラフが出力される（図 3）．「期間から探す」を選択した場合は，セレクトボックスに検索したい期間を月単位で指定することで，指定した期間内に動向情報の最大値を持つアイテム名，動向情報の最大値，URL，グラフを出力する（図 4）．

4.1 節に基本検索機能を用いた検索事例を，4.2 節にプロトタイプシステムと既存検索エンジンを併用した検索事例を示す．


図 2 提案システムの入力画面

## 4.1 基本検索機能を用いた検索事例

以下に，基本検索機能である「アイテムから探す」を選択した場合の検索事例と「期間から探す」を選択した場合の検索事例を示す．

● 「アイテムから探す」を利用した検索

ユーザが「野菜」に関する動向情報について調査したいと考えた場合を想定する．この場合，ユーザは入力フォーム上部の「アイテムから探す」を選択した上で，検索ボックスに「野菜名」（例えば，にんじん）を入力し，検索を実行する．プロトタイプシステムによる検索結果を図 3 に示す．システムによる出力から，ユーザは「にんじんの価格」の最大値が 2006 年 8 月の 517 円であることを知ることができる．


図 3 「アイテムから探す」を選択した場合の出力画面

● 「期間から探す」を利用した検索

ユーザが「過去に流行したアイテム」について関心を持ち，該当するアイテムを調査したいと考えた場合を想定する．この場合，ユーザは入力フォーム上部の「期間から探す」を選択した際に表示されるセレクトボックスに検索対象の期間（例えば，2011 年 3 月～2011 年 9 月）を指定し，検索を実行する．プロトタイプシステムによる検索結果から，ユーザは，対象期間に「自転車の販売量」などが最大値を迎えたことを知ることができる．

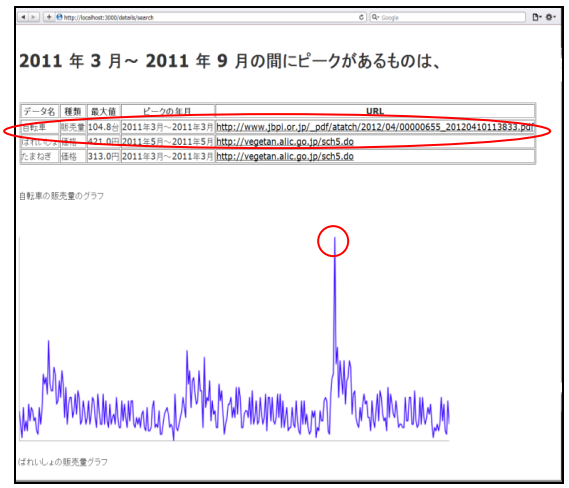


図 4 「期間から探す」を選択した場合の出力画面

## 4.2 提案システムと既存検索エンジンを併用した検索事例

提案システムを利用した実際の動向情報調査では，

4 節冒頭で述べたプロトタイプシステムの基本検索機能の他，既存検索エンジンの併用を想定している．本節では，基本検索機能と既存検索エンジンの両方を用いて「東日本大震災の影響を受けたアイテムの調査」と「過去において同時期流行したアイテムの調査」という動向に関する問いに答える検索事例を示す．

● 東日本大震災の影響を受けたアイテムの調査

ユーザが「東日本大震災がアイテムに与えた影響」に関心を持ち，様々なアイテムに関する動向情報の震災後における変化について調査したいと考えた場合を想定する．この場合，ユーザは「期間から探す」を選択し，クエリとして「2011 年 1 月～2011 年 12 月」を指定し，検索を実行する（図 5）．

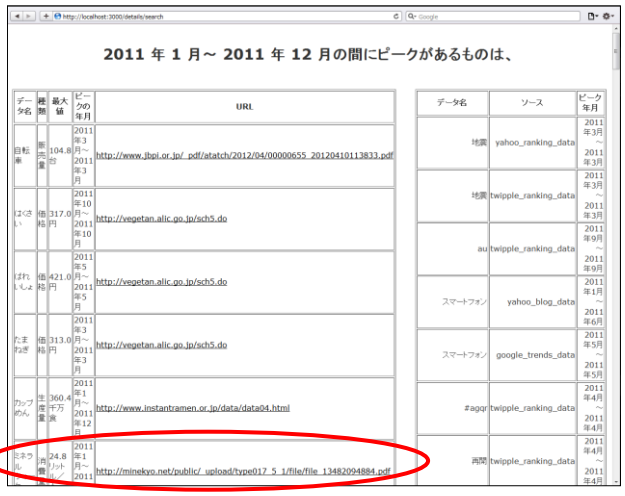


図 5 プロトタイプシステムの検索結果
（クエリ：2011/01～2011/12）

このとき，ユーザは検索結果から「ミネラルウォーターの消費量」が対象期間に最大値を迎えていることに興味を持ったとする．この場合には，続いてプロトタイプシステムの「アイテムから探す」から「ミネラルウォーター」をクエリに検索を実行し，さらに詳しい情報を得ることができる（図 6）．図より，「消費量」だけでなく，「検索数」や「ブログ記事数」などの主観的動向情報においても同期間に最大値を迎えていることが読み取れる．そこで既存検索エンジンを用いて，「ミネラルウォーター 2011」で検索した結果（図 7）から，ユーザはミネラルウォーターの消費量や検索数，ブログ記事数などの動向情報が大きく値を伸ばし，最大値を迎えたのは，東日本大震災の影響を受けたためではないかと推測することができる．この様に，提案する基本検索機能を用いることで，関心のあるアイテムを絞り込み，既存検索エンジンで効率良い情報収集が可能となる．

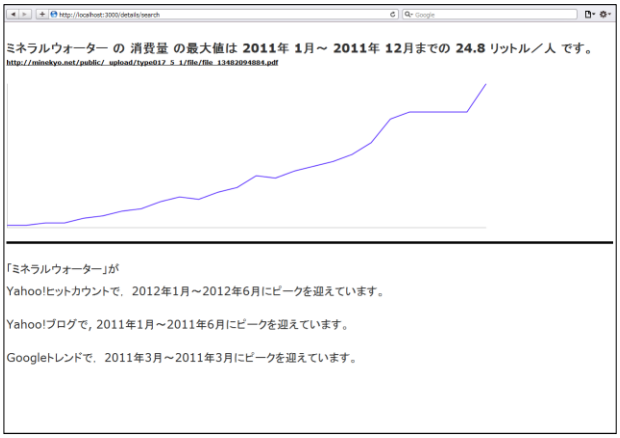


図 6 プロトタイプシステムの検索結果
（クエリ：ミネラルウォーター）


図 7 既存検索エンジンの検索結果
（クエリ：ミネラルウォーター 2011）

● 過去において同時期流行したアイテムの調査

ユーザが「過去において同時期に流行したアイテム」について調査したいと考えた場合の検索の流れを図 8 に示す．この検索には，状況に応じて，いくつかの異なる方法が考えられる．

一つは，ユーザが調査したい対象アイテムを想定している場合である．この場合には，プロトタイプシステムの「アイテムから探す」を用いて，対象アイテムの動向ピーク期間を調べたあとで，「期間から探す」によって，基準となる対象アイテムが動向の最大値を迎えた期間に，同じく動向の最大値を迎えているアイテム群を検索可能である．さらにその際

に，既存検索エンジンによる検索を併用し，実際にどのように話題となったのかを確認することで，流行の根拠を知ることができると考えられる．

ユーザが調査したい期間を予め想定している場合には別の方法が考えられる．その場合には，プロトタイプシステムの「期間から探す」を実行し，得られた結果から，興味を抱いたアイテムについて，「アイテムから探す」の実行や，既存検索エンジンでの検索により，調査を進めていくことが可能である．

また，どちらの方法であっても，既存検索エンジンを用いた検索中に，新しく関心の湧いたアイテムを発見した場合には，そのアイテムをプロトタイプシステムの「アイテムから探す」を用いて検索し，そのアイテムの動向情報を得ることも想定している．


図 8 同時期流行アイテム検索の流れ

## 5 おわりに

本稿では，「動向に関する問い」を対象タスクとして行う検索をコンテクスト検索と定義し，これに適した基本検索機能を提供する検索エンジンのプロトタイプシステムを構築した．また，動向に関する問いの例として，「東日本大震災の影響を受けたアイテムの調査」や「過去において同時期流行したアイテムの調査」というタスクに対して調査を行う事例を想定し，ユーザの情報要求が基本検索機能および既存検索エンジンへのクエリの組み合わせに分解される様子を示した．本稿で提案するコンテクスト検索は，タスクを限定することで高度な検索機能を提供しつつ，幅広いドメインへの適用が期待できるものであり，次世代検索エンジン実現に適した性質を備えていると考える．開発中のプロトタイプシステムでは，指定したアイテムに関する動向情報のピーク時期の検索，指定した期間に動向情報の最大値を持つアイテムの検索という2つの基本検索機能を提供するが，今後より充実させていく予定である．また，検索対象となる情報も，現状では主観的動向情報が6つの情報源から6種類，客観的動向情報が7つのwebサイトから31種類と小規模であるが，今後，収集する動向情報の量を増やすことで，さらに多くの問いに対して答えることが可能となる．構築したシステムを公開し，運用を通じて必要な基本検索機能の検討やユーザインタフェースの改良を行うことも重要であると考える．

## 参考文献

[1] A. Ferreira, J. Atkinson: Intelligent Search Agents Using Web-Driven Natural-Language Explanatory Dialogs, IEEE Computer, Vol. 38, No. 10, pp. 44-52 (2005)

[2] 亀井 俊之, 門田 暁人, 松本 健一: WWW を対象としたソフトウェア検索エンジンの構築, 電子情報通信学会技術研究報告 ソフトウェアサイエンス 102(617), pp.59-64 (2003)

[3] 加藤 恒昭, 松下 光範, 平尾 努: 動向情報の要約と可視化に関するワークショップの提案, 情報処理学会研究報告．自然言語処理研究会報告 2004(108), pp.89-94 (2004)

[4] 加藤 優, 高間 康史: Web コンテクスト情報に基づく同時期流行アイテム検索手法の提案, ファジィシステムシンポジウム講演論文集 28, pp.115-118 (2012)

[5] 小久保 卓, 小山 聡, 山田 晃弘, 北村 泰彦, 石田 亨: 情報処理学会論文誌 43(6), pp.1804-1813 (2002)

[6] 松下 光範, 加藤 恒昭: 数値情報の補填とグラフ概形の示唆による複数文書からの統計グラフ生成, 日本知能情報ファジィ学会誌 知能と情報 18(5), pp.721-734 (2006)

[7] 山本 健一, 谷岡 広樹, 殿井 加代子: 動向情報の検索による情報編纂, 第 21 回人工知能学会全国大会 (JSAI2007), 3H9-3 (2007)

# 語の上位下位概念と文書の主題との関連度を用いた例示部分特定手法

# Identifying Example Segments based on Relations between Hyponymy and Themes in a Document

大野 和久[1*]　田村 直之[1]　松本 征二[1]　新堀 英二[1]
Kazuhisa Oono[1]　Naoyuki Tamura[1]　Seiji Matsumoto[1]　Eiji Shinbori

[1] 大日本印刷株式会社 honto ビジネス本部
[1] Dai Nippon Printing Co., Ltd. honto Business Operations

**Abstract:** Recognizing logical paragraphs and relations of points in documents helps us to comprehend the documents. The logical paragraphs contain various segments such as "elaboration", "contrast" and "example". The authors often write their insistence using abstract terms. Therefore, they express their insistence in concrete cases with the examples, and accelerate to comprehend the documents. In this paper, we identify the example segments based on relations between hyponymy and themes in a document. We consider that sentences which contain concrete terms are divided into two types. The first type expresses themes, and the second type expresses examples. We calculate a rate of theme terms in a sentence, and capture whether the sentence expresses the themes or not. Thus, we find out that the sentence is likely to express the example if the sentence is not likely to express the theme. In our experimental evaluation, we confirmed that our proposed method scored better recall and F-measure than the baseline method.

## 1　はじめに

文書の読解には，文書を意味段落ごとに区切り，各意味段落での要点どうしを関係づけることが有効と考えられている [1][2]．その理由は，要点の関係性の認識について，認知的負荷が軽減され，情報の整理や文章構造を理解しやすくなるためである．

文書内容に含まれる意味段落のうち，読解を支援する要素の一つとして，例示がある．例示は，一般化された著者の主張を具体化する表現である．著者の主張は抽象的な表現で記述される場合があり，その表現だけでは，読者は著者の主張を理解しにくい場合がある．そこで，読者は，例示を用いた具体的表現を併せて読むことにより，既に持っている知識を想起し，その想起されたイメージと著者の主張を結びつけることができ，著者の主張を理解しやすくなると考えられている [3]．そのため，文書の読解には，著者の主張だけでなく，その主張を説明するための例示も把握することが有効となる．

読解を支援するための文書構造解析技術として，談話構造解析がある．談話構造解析では，意味段落や文章間の関係性を明らかにする．このことにより，話題の推移をとらえたり，各段落および文章の役割を認識することが可能となる．

談話構造解析の既存研究では，例示特定のために，手がかり語を用いて特定する手法や，語の抽象度の遷移を用いて特定する手法がある．手がかり語を用いる手法では，文章中に "例えば" や "〜の「ように」" が含まれる場合，その文章を例示として特定する [4]．また，語の抽象度の遷移を用いる手法では，複数の文章のまとまりがあり，ある文章 $A$ で述べられた事象や状態の具体例が，文章 $A$ に続く文章 $B$ で提示される場合，文章 $B$ を例示として特定する [5]．

ただし，これらの既存手法では，手がかり語が存在しない場合や，例示部分の周りに抽象表現が存在せず，例示部分が独立して出現する場合は，例示部分の特定が困難である．

そこで本研究では，これらの既存手法で特定できなかった例示部分の特定を可能にすることを狙い，語の上位下位概念と文書の主題との関連度を用いて例示部分を特定する手法を提案する．提案手法では，具体的な表現を含む文章が例示になりやすいことに着目した上で，その具体的な表現と主題との関係性を考慮することにより，例示特定についての再現率向上を行うと

*連絡先：大日本印刷株式会社 honto ビジネス本部
〒 162-8001 東京都新宿区市谷加賀町 1-1-1
E-mail: Oono-K6@mail.dnp.co.jp

同時に，精度向上も行う．

ここで本研究では，文書は，章もしくは節全体を示し，文章は，句点で区切られた一つの文を示す．

## 2 提案手法で特定する例のパターン

本研究では，例示を交えながら著者の主張を論述する文書を対象とする．例えば，評論文，論説文，随筆といった文書である．

これらの文書では，一般的な表現と，具体的な表現を繰り返すことにより，著者の主張を論述している [6]．ここで，一般的な表現は著者の主張を表し，具体的な表現は，読者に対して著者の主張を納得させるための証拠を表す．この具体的な表現が，主張に対する例示となる．このように，文章内容は，二つの表現に分けられる．

具体的な表現である例示は，手がかり語が存在する場合と存在しない場合に分けられ，合わせて 5 種類のパターンがある．

手がかり語が存在する場合については，2 種類のパターンがある．

1. 例示を示す接続詞
   “例えば” といった接続詞を用いて，例示を述べる場合である．

2. 具体例の列挙を示す語
   具体例を挙げる際に，“〜の「ように」”，“「ある」書籍では”，“〜「など」” といった語を用いる場合である．

一方，手がかり語が存在しない場合では，3 種類のパターンがある．

3. 主題と異なる分野での事象を記述
   文書の主題とは異なる分野での事象を用いて，例示を述べる場合である．例えば，ジャーナリズム論の文書の中で，物理学の理論や美術家の行動を用いて例を述べる場合である．なお，比喩法として用いられる隠喩についても，この場合に含まれると考える．

4. 上位語から下位語へ遷移
   複数の文章があるときに，まず抽象的な表現を述べ，次に，抽象的表現の下位語にあたる語を用いて，例示を述べる場合である．例えば，抽象的表現として，「マスメディアは様々な情報を発信している.」といった文章があり，その文章の後に，「新聞は紙を媒体とし，日々のニュースを発信している.」といった文章がある場合を考える．このとき，マスメディアという上位語から，新聞という下位語へ話題が遷移している．そこで，新聞に関する文章は，マスメディアに関する文章の例示として述べていることになる．

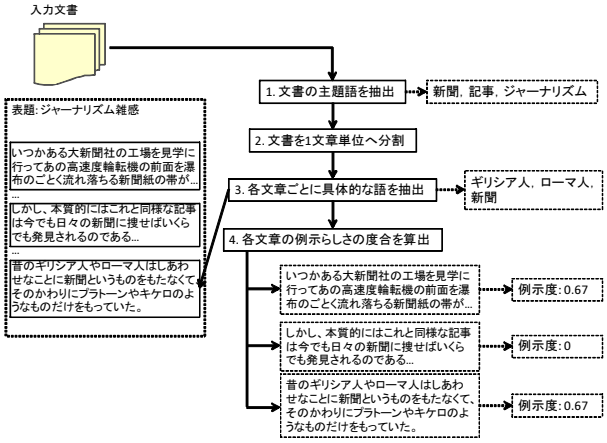


図 1: 提案手法の流れ

5. 著者の体験を記述
   著者の体験を用いて，例示を述べる場合である．例えば，ジャーナリズム論について述べる際に，著者が記者と対談した経験を引合いに出し，その経験を一般化した主張を述べる場合である．

本研究では，この 5 種類のうち，「3. 主題と異なる分野での事象を記述する場合」と，「4. 上位語から下位語へ遷移する場合」の 2 種類について例示部分を特定する手法を提案する．

## 3 上位下位概念と主題との関連度を用いた例示部分特定

### 3.1 基本的な考え方

本研究では，まず，主題を表す主題語を文書から抽出した上で，各文章中に含まれる具体的な語の総数に対して，主題語が占める割合を算出する．そして，その総数全体の割合から主題語が占める割合を引くことにより，各文章の例示らしさの度合いを算出し，例示部分を特定する．ここで，主題語とは，主題を表す単語を示す．

この理由として，まず，具体的表現を述べる文章には，具体的事象や事物を示す語が含まれやすく，それらの語が含まれている文章ほど，例示になり得ると考える．具体的な語とは，明確な実体や，個々の事物に即している語であり，例えば，“ローマ人” や “大日本印刷” という語が該当する．

このとき，具体的な語が文章に含まれていても，その語が主題を表していれば，例示として記述されてい

るとは考えにくい．例えば，ジャーナリズム論の文書中に “ローマ人” が出現する場合は例示として考えられるが，ローマ人の歴史に関する文書中に “ローマ人” が出現する場合は，文書の主題に沿っていると考えられ，例示として用いられているとは考えにくい．このように，具体的な語を含む文章は，主題に沿う文章か，もしくは，例示の文章のいずれかに分けられると考え，文章が例示を表しているかどうかを特定するためには，具体的な語が主題に沿っているかどうかを特定する必要があると考える．そこで，文章に含まれる具体的な語の出現傾向が主題に沿っていれば，その文章は例示である可能性が低いととらえ，主題に沿っていなければ，その文章は例示である可能性が高いととらえることにより，例示部分を特定することができると考える．

提案手法の流れについて，図 1 に示す．図 1 では，寺田寅彦による「ジャーナリズム雑感」[1]を例として用いている．

なお，例示であるかどうかの判定については，文書を句点 (“。”，“．”) で区切り，区切られた一つの文章ごとに，例示特定の判定を行う．

### 3.2 利用する上位下位語の条件

文章から具体的な語を抽出するためには，その語が明確な実態や個々の事物を示している語であるという情報が必要である．そのため，形態素解析結果の名詞だけを利用するといった方法では，その名詞が具体的な語であるかどうかを判断することができない．

そこで本研究では，文章に対して形態素解析処理を行い，形態素が名詞-一般および名詞-固有名詞である場合，その形態素を上位下位概念の情報を含む DB と照合し，その DB に含まれていれば，利用する語の候補とする．そして，各候補の語に対して，DB における語の深さを算出し，ある一定の深さよりも深い場所に位置する語だけを用いる．

語の深さを用いる理由は，DB に含まれる語をそのまま用いると，上位概念の語を利用する場合が発生するためである．例えば，文章から “人気”，“方法” といった語が得られ，これらの語が DB 内に存在する場合，これらの語も抽出対象となる．しかし，具体的な語としては，これらの語は不適と考えられる．

本研究では，形態素解析器として，MeCab[2] を用い，上位下位概念 DB として，日本語 WordNet[7] を利用した．日本語 WordNet では，語の意味概念ごとに上位概念，下位概念が定められており，抽象的な表現である上位概念から具体的な表現である下位概念まで，木構造によって構成されている．

[1]http://www.aozora.gr.jp/cards/000042/files/2492_10275.html
[2]https://code.google.com/p/mecab/

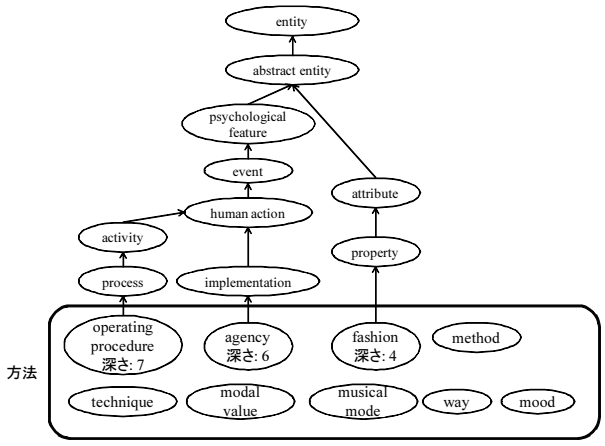


図 2: 深さによる具体的な語の絞込み

語の深さを用いて，利用する上位下位語を絞り込む処理について，図 2 に示す．まず，日本語 WordNet において当該語の上位概念を辿る．このとき，日本語 WordNet では，表記が同じでも，複数の意味を持つ場合，その意味ごとに *Synset* と呼ばれる概念が割り当てられている．そして，その概念ごとに語の上位概念が存在する．例えば，“方法” は，9 個の概念に属しており，その概念ごとに上位概念が存在している．

そこで，各概念ごとに上位概念からさらに上位概念へ辿って行き，上位概念が存在しない概念まで辿る．このときのそれぞれの深さを算出し，最小の深さを当該語の深さとする．この深さについて，深さの値が大きいほど具体的な語であり，小さいほど抽象的な語であると考える．最小の深さを適用する理由は，複数の概念のうち，一つでも深さの値が小さければ，その語は抽象的な意味で用いられる場合があり，すべての概念において深さの値が大きくないと，具体的な語とはいえないと考えるためである．

図 2 の例では，“operating procedure” という概念では，深さが 7 となり，一方，“fashion” という概念では，深さが 4 となった．“方法” では，最小の深さが 4 であったため，“方法” の深さを 4 とした．

そして，この深さに対して，あらかじめ閾値を設定しておき，閾値を超えていれば，具体的な語として用いる．

### 3.3 主題語の抽出

文書の主題とは，文書の中で著者が特に主張したい内容を表している表現である．先行研究では，表題に含まれる語を主題語としてとらえたり [8]，文書内の出現頻度が高い語を主題語としてとらえる考え [9] がある．そこで，本研究においては，以下の二つの条件のいず

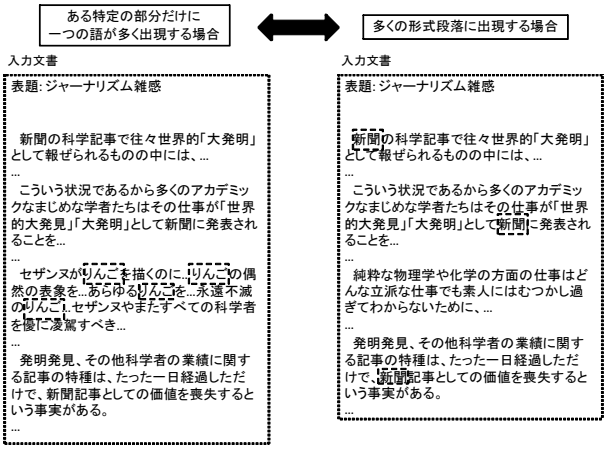


図 3: 出現段落数を用いた主題語抽出

れかに該当する語を主題語の候補として抽出し，各条件での候補を合わせた結果を主題語とする．

一つ目の条件は，表題に含まれる語である．まず，MeCab によって表題を形態素解析する．そして，3.2 節と同様に，形態素が名詞-一般および名詞-固有名詞である場合に限り，その形態素を日本語 WordNet と照合し，深さが閾値を超える語を，表題に含まれる主題語とする．

二つ目の条件は，出現する形式段落の数が多い語である．この理由は，著者の主張を示す主題語は，文書中で繰り返して記述されることが一般的だからであり，語の網羅性が主題語抽出のための指標として有効であると考えるためである．なお，このときの語は，3.2 節で述べた処理により，具体的な語だけに絞られている．

ある一箇所において特定の語が頻出する場合と，文書中に繰り返し出現する場合の違いについて，図 3 に示す．“りんご” という語は，ある形式段落において頻出するが，特定の形式段落にだけ出現するため，主題を表しているとは考えにくい．一方，“新聞” という語は，多くの形式段落において繰り返し出現しており，文書の主題を表していると考えられる．

文書内から全部で $m$ 個の具体的な語 $w_1, w_2, \cdots, w_m$ が得られるとき，$i$ 番目 $(i = 1, 2, \cdots, m)$ の語の主題らしさを表す度合 $s_i$ を算出する式を，(1) 式に示す．

$$s_i = \frac{p_i}{N} t \tag{1}$$

ここで，$p_i$ は語 $w_i$ が含まれる形式段落の数，$N$ は文書内の形式段落の総数，$t$ は表題に含まれる語への重みを示す．

(1) 式では，文書内の形式段落の総数に対する，語が含まれる形式段落の数の割合を算出している．このとき，表題に含まれる語は，著者の主張を特に表していると考え，重み $t$ を与えている．主題語として決定するために，$s_i$ が，あらかじめ設定した閾値を超えていれば，その語を主題語として抽出した．ただし，抽出する主題語の最大数は 3 とした．

なお，合わせた結果が主題語の最大数を超えていれば，表題に含まれる主題語が，より著者の主張を表していると考え，表題に含まれる主題語を優先して用いることとした．

## 3.4 主題語の上位語の抽出

3.3 節で抽出した主題語に加え，主題語の上位概念も，主題を表す語として用いる．この理由として，ある主題について記述する場合，その主題語の上位概念も，文書の主題に含まれると考えられるためである．例えば，“新聞” を主題語とする文書の中で，“マスメディア” という語が現れる場合を考える．この場合，“新聞” の上位概念となる “マスメディア” は，“新聞” よりも抽象的な表現となる．そのため，文書内容が抽象的な表現と具体的な表現に大別されることを考えると，この文書では，“マスメディア” は抽象的な表現である著者の主張を表していると考えられ，例示のような具体的な表現として扱われるとは考えにくい．そこで，抽出した主題語だけでなく，主題語の上位概念も主題として用いることとする．

本研究では，日本語 WordNet において，主題語が属する概念の各上位概念に属する語を，主題語の上位概念として用いた．このとき，主題語と直接結びつく上位概念の語だけを用いた．

## 3.5 例示度の算出

本研究では，文章中に含まれる主題語の頻度と，主題語以外の具体的な語の頻度とを比較し，例示らしさの度合を算出することにより，例示部分を特定する．ここで，例示らしさの度合を例示度とする．

3.1 節で述べたように，具体的な表現を含む文章は，主題に沿う文章か，例示の文章の二つに分けられると考える．そこで，文章中に出現する具体的な語の総数のうち，まず，主題語とその上位語の割合を求め，主題に沿う文章であるかどうかの度合いを算出する．そして，総数全体の割合から主題の割合を引くことにより，その文章の例示度を算出する．文書内から全部で $n$ 個の文章 $u_1, u_2, \cdots, u_n$ が得られるとき，$j$ 番目 $(j = 1, 2, \cdots, n)$ の文章の例示度 $e_j$ を算出する式を (2) 式に示す．

$$e_j = 1 - \frac{f_j + g_j}{M} \tag{2}$$

ここで，$f_j$ は文章 $u_j$ に含まれる主題語の出現総数，$g_j$ は文章 $u_j$ に含まれる主題語の上位語の出現総数，$M$ は文章 $u_j$ に含まれる具体的な語の出現総数を示す．

(2) 式により，例示度が高いほど，例示を表現している文章であると考え，例示度が低いほど，主題を表現している文章であると考える．

# 4 評価実験

## 4.1 実験方法

提案手法による例示特定結果と，既存手法による例示特定結果において，それぞれの再現率，精度，F 値を比較し，提案手法の評価を行う．

例示特定結果の評価として，文章単位での例示特定結果に加え，部分単位での例示特定結果についても評価を行う．

その理由として，提案手法および既存手法では，各文章が例示であるかどうかを判断するが，実際の例示は，一つの文章だけでなく，複数文章のまとまりによって一つの例示を表現している記述があるためである．そこで，実験対象の各文書内容を人手によって，例示を表現している複数文章のまとまりと，主題を表現している複数文章のまとまりに分けた．例示を表現している複数文章のまとまりを例示部分とする．

なお，例示部分を特定できたかどうかの判断としては，人手によって判定した例示部分に含まれる文章のうち，少なくとも一つの文章を特定できていれば，例示部分を特定できているとみなす．

提案手法については，例示度の閾値を $0.5, 0.6, 0.7, 0.8$ の 4 種類に設定し，各場合において，算出した例示度が閾値を超えていれば，例示とみなすようにした．

既存手法としては，手がかり語による例示特定を行った．このときの手がかり語としては，既存研究 [4][5] を参考に，“例えば”，“たとえば”，“例”，および，連体詞の “ある” を用いた．連体詞の “ある” は，“ある書籍” といった記述を指す．

表 1: 実験対象文書

| 文書番号 | 作品名 | 著者 |
|---|---|---|
| 1 | ジャーナリズム雑感 | 寺田寅彦 |
| 2 | 科学的新聞記者 | 桐生悠々 |
| 3 | 漫画と科学 | 寺田寅彦 |
| 4 | 教育映画について | 寺田寅彦 |
| 5 | 流言蜚語 | 寺田寅彦 |
| 6 | 形態について | 豊島与志雄 |
| 7 | 芸術と社会 | 津田左右吉 |

実験対象の文書として，青空文庫[3]より 7 作品を用いた．具体的な作品名および著者を表 1 に示す．これらの作品を用いた理由は，これらの作品が評論や随筆のジャンルに該当し，例示を交えながら著者の主張を記述しているためである．

これらの作品に対して，人手によって，各文書内の各文章が例示文章であるかどうかを判定した．判定基準としては，2 章で述べた，5 種類の例のパターンをもとに判定した．

各作品の例示文章の割合を表 2 に示し，例示部分の割合を表 3 に示す．例示文章の割合の平均は 0.41，例示部分の割合の平均は 0.54 となり，文書内容の半数が例示であることを示している．この結果は，2 章で述べた，抽象的な表現と具体的な表現が繰り返されて記述されるといった論述形式を表している．

3.2 節における，具体的な語を絞るための深さの閾値は，表 1 の各作品の文書内容に含まれる具体的な語に対して深さを算出し，その算出結果から，4 とした．また，3.3 節の，主題語の抽出における，表題に含まれる語への重みは，$t = 2$ とし，主題語決定のための閾値は，0.4 とした．

表 2: 例示文章の割合

| 文書番号 | 全文章数 | 例 | 例以外 | 例の割合 |
|---|---|---|---|---|
| 1 | 148 | 59 | 89 | 0.4 |
| 2 | 69 | 9 | 60 | 0.13 |
| 3 | 80 | 20 | 60 | 0.25 |
| 4 | 75 | 29 | 46 | 0.39 |
| 5 | 46 | 23 | 23 | 0.5 |
| 6 | 55 | 32 | 23 | 0.59 |
| 7 | 44 | 26 | 18 | 0.59 |

表 3: 例示部分の割合

| 文書番号 | 全部分数 | 例 | 例以外 | 例の割合 |
|---|---|---|---|---|
| 1 | 56 | 30 | 26 | 0.54 |
| 2 | 21 | 9 | 12 | 0.43 |
| 3 | 28 | 15 | 13 | 0.54 |
| 4 | 26 | 15 | 11 | 0.58 |
| 5 | 11 | 6 | 5 | 0.55 |
| 6 | 20 | 12 | 8 | 0.6 |
| 7 | 7 | 4 | 3 | 0.57 |

[3] http://www.aozora.gr.jp/

## 4.2 実験結果

文章単位での例示特定結果について，提案手法および既存手法の再現率，精度，F 値を表 4 に示す．また，部分単位での例示特定結果について，表 5 に示す．提案手法の結果については，7 作品に対する例示特定結果の平均値を示している．

これらの結果から，精度においては，既存手法の方が上回っているが，再現率においては提案手法の方が，いずれの閾値においても上回っており，既存手法では特定できていなかった例示を特定できていることがわかる．

例示度の閾値を変化したことによる結果の違いについては，例示文章単位では閾値が 0.6 での F 値が最も高くなり，例示部分単位では閾値が 0.5 での F 値が最も高くなった．

例示文章単位では，誤って例示特定する場合が多く見受けられ，その誤り数が精度に影響を与えている．このときの精度低下の原因として，主題語が省略されていることを考える．主張を表すすべての文章には，必ずしも主題語が含まれているとは限らず，省略される場合がある．その場合，主題語以外の具体的な語が一つでも出現していれば，主題に沿う文章であるにもかかわらず，例示度を高く算出した．結果として，主題語が省略されている文章が頻出し，誤って例示特定した文章が多くなり，精度に影響を与えたと考える．

例示度の算出結果について，具体的な結果例を図 4 に示す．文章 1 は，例示として正しく特定することができた文章である．文章 1 が現れる文書の主題語は “記事”，“新聞”，“ジャーナリズム” であり，主題語の上位語には “マスメディア”，“ニュース” といった語が含まれる．そして，この文章には “セザンヌ”，“りんご”，“カンバス” が具体的な語として出現する．これらの語は主題に含まれない語であるため，例示度は 1 と算出される．

文章 2 は，例示度を考慮することにより，著者の主張が含まれることを正しく算出することができた文章である．文章 2 は，文章 1 と同じ文書内に現れるが，具体的な語として，“セザンヌ”，“新聞”，“科学”，“記者” が出現する．このうち，“新聞” が主題語となり，3 回出現しているため，例示度は 0.57 と算出される．文章 1 に比べ，文章 2 は著者の主張も含まれていると考えられるため，例示度を用いて，その区別を行うことができたと考える．

ただし，誤った例示特定を行う場合もあった．文章 3 は，具体的な語としては適さない語を利用していることにより，具体的な語の割合が高くなり，誤って例示と判断した場合である．この文章では，“類型”，“自身” といった語を利用しており，それらの語の頻度が高くなることにより，誤って例示と判断している．3.2 節

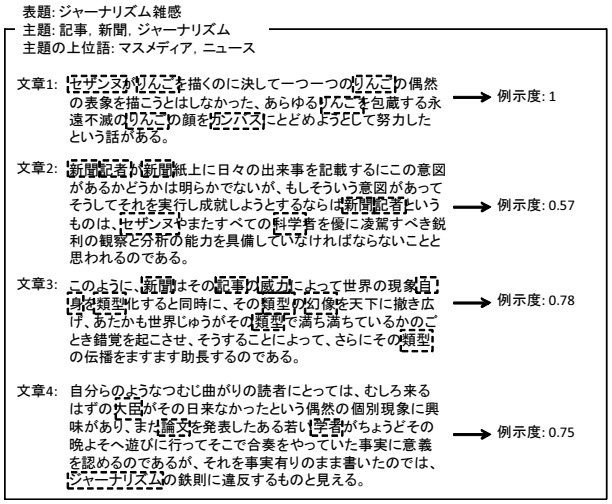


図 4: 例示度算出結果の具体例

による処理をもとに，利用する具体的な語の種類を絞り込んでいるが，深さだけでは抽象的な表現の語を除去しきれていないと考える．

また，文章 4 は，主題語に比べて具体的な語が頻出することにより，誤って例示と判断した場合である．この文章には，“論文”，“学者” といった具体的な語が出現する一方で，“ジャーナリズム” といった主題語も出現する．人手による判断では，この文章を著者の主張と判断したが，主題語以外の語の頻度に基づく提案手法では，この文章は例示である可能性が高いと判断する結果となった．出現頻度による重みだけでは，具体的な表現と一般的な表現を区別しきれていないと考える．

表 4: 文章単位の再現率，精度，F 値比較

| | 再現率 | 精度 | F 値 |
|---|---|---|---|
| 既存手法 | 0.16 | 0.86 | 0.25 |
| 提案手法 (閾値: 0.5) | 0.94 | 0.5 | 0.63 |
| 提案手法 (閾値: 0.6) | 0.89 | 0.55 | 0.66 |
| 提案手法 (閾値: 0.7) | 0.77 | 0.57 | 0.64 |
| 提案手法 (閾値: 0.8) | 0.7 | 0.57 | 0.62 |

表 5: 部分単位の再現率，精度，F 値比較

| | 再現率 | 精度 | F 値 |
|---|---|---|---|
| 既存手法 | 0.25 | 0.96 | 0.38 |
| 提案手法 (閾値: 0.5) | 0.95 | 0.71 | 0.8 |
| 提案手法 (閾値: 0.6) | 0.89 | 0.72 | 0.79 |
| 提案手法 (閾値: 0.7) | 0.76 | 0.72 | 0.73 |
| 提案手法 (閾値: 0.8) | 0.66 | 0.72 | 0.68 |

## 5 おわりに

本研究では，文書中の例示部分を特定する手法として，語の上位下位概念と文書の主題との関連度を用いて例示部分を特定する手法を提案した．この手法では，具体的な表現が含まれる文章は例示になりやすいことに着目し，既存手法では取得できていなかった，文章中に手がかり語が存在しない場合や，例示部分が独立して出現し，語の抽象度の遷移を取得できない場合において，例示特定についての再現率向上を行った．また，具体的な表現が含まれるからといって必ずしも例示にはならない場合があることも考慮し，具体的な語と主題との関係性を用いて例示度を算出することにより，精度低下の防止も行った．

評価実験の結果，手がかり語を用いた例示特定と比較して，再現率およびF値について，提案手法が上回っていることを確認した．

今後の課題として，精度向上のための2点の課題をあげる．1点目は，具体的な語としては適さない語を除去することである．

2点目は，人手による判断としては著者の主張として考えられる文章について，主題語に比べて具体的な語が頻出する場合でも，主題に沿う文章であると判別できるようにすることである．

## 参考文献

[1] Akio Suzuki. Differences in Reading Strategies Employed by Students Constructing Graphic Organizers and Students Producing Summaries in EFL Reading. *JALT Journal*, Vol. 28, pp. 177–196, 2006.

[2] 岡孝明, 武田英明. 技術論文のチャート化による論文理解の支援についての分析. 電子情報通信学会技術研究報告. AI, 人工知能と知識処理, Vol. 99, No. 447, pp. 27–34, nov 1999.

[3] 塚本真紀. 具体例の生成が文章理解による学習の転移に及ぼす影響. 尾道大学芸術文化学部紀要, Vol. 4, pp. 30–36, 2005.

[4] Daniel Marcu. The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Computational Linguistics*, Vol. 26, pp. 395–448, 2000.

[5] 梅澤俊之, 原田実. センタリング理論と対象知識に基づく談話構造解析システム DIA. 自然言語処理, Vol. 18, No. 1, pp. 31–56, jan 2011.

[6] 出口汪. 図解「出口式」論理力ノート : カリスマ講師が教える仕事で成功する思考法. PHP 研究所, 2006.

[7] Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. Japanese SemCor: A Sense-tagged Corpus of Japanese. *In Proceedings of The 6th International Conference of the Global WordNet Association (GWC-2012)*, 2012.

[8] 野本忠司, 松本裕治. テキスト構造を利用した主題の推定について. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 96, No. 65, pp. 47–54, jul 1996.

[9] Gerard Salton and Christopher Buckley. Term-weighting Approaches in Automatic Text Retrieval. In *INFORMATION PROCESSING AND MANAGEMENT*, pp. 513–523, 1988.

# 疑似ラベルを用いた潜在的ディリクレ配分法の提案

## Pseudo Labled Latent Dirichlet Allocation

鈴木聡子 [1*]　小林一郎 [2]
Satoko Suzuki[1]　Ichiro Kobayashi[2]

[1] お茶の水女子大学　理学部　情報科学科
[1] Department of Information Science, Faculty of Science, Ochanomizu University
[2] お茶の水女子大学大学院人間文化創成科学研究科理学専攻
[2] Advanced Science, Graduate School of Humanities and Science, Ochanomizu University

**Abstract:** In recent years, topic models have been widely used for many applications such as document summarization, document clustering etc. Labeled latent Dirichlet allocation (LLDA) was proposed based on latent Dirichlet allocation (LDA), and it regards the tags, i.e., labels, put on documents by humans as the ones expressing the contents of the documents, and uses them as supervised information to estimate latent topics of the documents. Moreover, it is reported that LLDA exceeds the ability of LDA in terms of topic estimation. However, normal documents usually do not have such tags with them, so, the use of LLDA is considerably limited.In this study, therefore, we make pseudo labels from the documents to be estimated their latent topics instead of tags put on documents by humans, and aim to make LLDA available for all documents.

## 1　はじめに

近年，文書の潜在情報であるトピックを考慮したトピックモデルが文書要約や文書分類に利用されている．潜在ディリクレ配分法 (LDA)[1] に基づいて提案された Labeled LDA(L-LDA)[2] は，人によって文書に予め付けられているタグを，その文書の意味内容を表すものと捉え，潜在トピック抽出における教師信号として利用することを考えたモデルであり，複数のタグ付き文書に対しての LDA を上回る性能を示すと知られている．しかし実際は，世の中のほとんどの文書にはタグが付与されておらず，L-LDA の使用される範囲は限られている．そこで本研究では，文書集合からタグの代わりとなる疑似ラベルを作成し，全ての文書に対して L-LDA が有用になることを目的とする．

## 2　Labeled LDA

L-LDA は，LDA におけるトピック分布を推定する過程で，文書に付与されたタグの情報を考慮したモデルとなっている．図 1 に L-LDA のグラフィカルモデルを示す．L-LDA と LDA との違いは，ラベル（文書に与えられているタグ）の情報が，$\theta$ を推定する際に影響を与えているという点である．

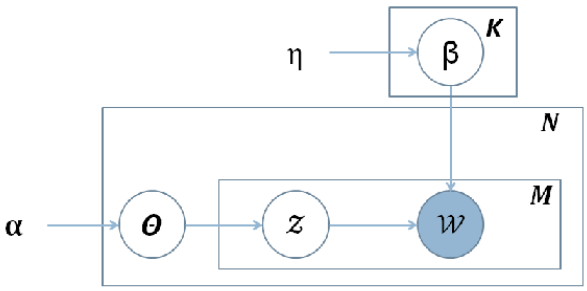


図 1: L-LDA のグラフィカルモデル

まず，文書ごとに付与されているタグの情報から，文書ラベル $\Lambda^{(d)}$ を生成する．

$$\Lambda^{(d)} = (l_1, \cdots, l_K) \quad l_k \in \{0, 1\} \tag{1}$$

$K$ は文書群に含まれる重複の無いラベルの個数であり，文書ごとにラベルの有無の情報を 1 または 0 の 2 値で与える．次に文書におけるラベルのベクトルを定義する．

$$\lambda^{(d)} = \{k | \Lambda_k^{(d)} = 1\} \tag{2}$$

$\lambda^{(d)}$ は，文書 $d$ に付与されているラベル番号である．

*連絡先：お茶の水女子大学理学部情報科学科小林研究室
〒 112-0012 東京都文京区大塚 2-1-1
E-mail: g0920519@is.ocha.ac.jp

そして，文書ごとに射影行列を生成する．

$$L_{ij}^{(d)} = \begin{cases} 1 & if\ \lambda_i^{(d)} = j \\ 0 & otherwise. \end{cases} \quad (3)$$

生成した射影行列と設定したハイパーパラメータ $\alpha$ から，文書ごとに新しいパラメータ $\alpha^{(d)}$ を生成する．ラベルの情報により制限された $\alpha^{(d)}$ から，トピック分布 $\theta$ を求める．他の過程は，LDA と同様である．

# 3 疑似ラベル生成

本研究では，文書の 2 つの表層的な情報からラベルの代わりとなる疑似ラベルを生成する．

## 3.1 単語の共起情報に基づく疑似ラベル生成

ここでは，Newman らの研究 [3] よる，文書の潜在的意味の一貫性は単語の共起関係と関連があるということから，共起関係の強い単語より疑似ラベルを生成することを考える．

まず，疑似ラベルを構成する候補として，各文書から TF-IDF の値が高い単語を抽出する．そして，抽出した単語の出現頻度を全ての文書において求める．ここで文書頻度が 1 である単語は，その文書特有の単語であると考え，抽出したリストから消去する．残った単語を疑似ラベルを生成する単語の候補とする．

次に，抽出した単語の共起関係を自己相互情報量 (PMI) を用いて求める．ここで，2 つの単語が同じ文書において疑似ラベルの候補に抽出されることを，共起していると捉えることとする．PMI が閾値以上の単語同士をグループ化し，グループごとを疑似ラベルとして設定する．また，共起情報によるクラスタリングで作られたラベルの他に，PMI の値は低いが出現頻度は高い単語も，ラベルとして採用する．

疑似ラベルを構成する単語が抽出されている文書に，同じラベルを与える．ただし，複数の単語で構成される疑似ラベルについては，構成する単語が 2 つ以上抽出されている場合にラベルを与える．

## 3.2 文書の類似度に基づく疑似ラベル生成

文書 $d_i$ におけるベクトル中の重みを式 (4) とし，生成した文書ベクトルをもとに文書分類を行う．

$$w_{ij} = (\log x_{ij} + 1.0) \log(N/n_j) \quad (4)$$

$x_{ij}$ は文書 $d_i$ における語 $t_j$ の出現回数であり，$N$ は全文書数，$n_j$ は語 $t_j$ の出現する文書数である．その結果，類似する文書に同じラベルを与える．ここでは，Leader-Follower 法と Crouch 法の 2 つの方法 [5][6] において疑似ラベルを生成する．この 2 つの方法は，分類の重複を許すアルゴリズムとなっており，1 文書に対し複数のラベルを生成することが可能である．

### Leader-Follower 法

ここで用いるのは，Leader-Follower 法である．本研究で用いたアルゴリズムの概要を以下で説明する．

1. 文書をクラスタに併合するための閾値を設定する．
2. 1 つめの文書を読み，クラスタとして設定する．
3. 1 文書ずつ読む．全ての文書が読み終わったら処理を終了する．
4. 読み込んだ 1 文書と，その時点で存在する全てのクラスタとの類似度を計算する．
5. 閾値以上の類似度をもつクラスタにその文書を併合し，クラスタの語の重みの値を更新する．その文書との類似度が，どのクラスタにおいても閾値を超えない場合は，新しいクラスタとして生成する．
6. 手順 3 に戻る．

ここでクラスタ $C_h$ 中の語 $t_j$ の重みを式 (5) と定義する．

$$\hat{w_{hj}} = \log \sum_{d_i \in C_h} x_{ij} + 1.0 \quad (5)$$

また，類似度にはコサイン類似度を用いる．これによって生成されたクラスタについて，同じクラスタに含まれている文書に同じ疑似ラベルを与える．クラスタを構成する文書数が 1 の場合には，疑似ラベルを与えないこととする．

### Crouch 法

この手法は Leader-Follower 法を拡張したものであり，クラスタの設定とクラスタへの文書の割り当てを 2 段階の処理によって行うことが特徴である．
Crouch 法では，Leader-Follower 法と同様に設定したクラスタと，全文書の類似度を計算し，閾値以上の値を持つ文書に同じ疑似ラベルを与える．クラスタと文書の類似度は式 (6) によって求める．

$$s(d_i, C_k) = f \frac{\sum_{j=1}^{M} min(w_{ij}, \hat{w}_{kj})}{min(\sum_{j=1}^{M} w_{ij}, \sum_{j=1}^{M} \hat{w}_{kj})} \quad (6)$$

ここで，$d_i$，$C_k$，$w_{ij}$，$\hat{w}_{kj}$ については，上述した変数であり，$M$ は全語彙数とする．

# 4 実験

タグ付けされていない文書集合に疑似ラベルを付与し，文書分類の課題を通じて各手法と LDA との比較を行う．

## 4.1 実験仕様

使用するデータは，20 Newsgroups[1]の 20 カテゴリの内，10 個のカテゴリを選び，その中からそれぞれ 100 文書を選んだ，合計 1000 文書を用いる．この合計 1000 文書から成る文書集合を 2 セット用意した．(2 つの文書集合を setA，setB と区別する．)

選んだカテゴリを表 1 に示す．

表 1: 選択カテゴリ

| setA | setB |
|---|---|
| alt.atheism | alt.atheism |
| comp.graphics | comp.graphics |
| com.sys.mac.hardware | comp.ibm.pc.hardware |
| rec.sport.baseball | misc.forsale |
| sci.med | rec.autos |
| sci.crypt | rec.motorcycles |
| sci.electronics | sci.electronics |
| sci.space | sci.space |
| talk.politics.guns | talk.politics.guns |
| soc.religion.christian | talk.politics.misc |

２つの文書集合における 10 個のカテゴリは，カテゴリ内で内容が偏らないように選んだ．また，setA と setB でカテゴリが共通する場合は，異なる文書が選ばれている．それらの文書集合は，ストップワードを除いた後，ステミング処理を施す．提案手法の実験は，単語の共起情報により疑似ラベルを生成した場合と，文書の類似度から疑似ラベルを生成した場合の 2 つのパターンにおいて行う．

単語の共起情報により疑似ラベルを生成した場合（パターン 1 とする）では，抽出する単語数は各文書ごとに TF-IDF の値が上位 30 単語とした．また，1 単語から構成されるラベル数は，出現回数が上位 5 位までの単語を選んだ．ここで，予備実験より PMI の閾値はラベルが複数できる範囲，setA では [4.5,6.2]，setB では [4.8,6.2] において実験を行うものとした．

文書の類似度から疑似ラベルを生成する場合（パターン 2 とする）では，Leader-Follower 法と Crouch 法の 2 つの方法で疑似ラベルを生成した．類似度の閾値は，[0.1,0.9]（パターン 2a とする）と 0.1 以下（パターン 2b とする）において実験を行った．全ての実験で，L-LDA に与えるハイパーパラメータの値は，$\alpha$=0.1，$\eta$=0.1 とした．比較対象である LDA では，まずトピック数を設定するための予備実験を行った．その結果，それぞれの文書集合における最適トピック数を setA では 16，setB では 28 と設定した．LDA において与えるパラメータの値は，提案手法と同じく $\alpha$=0.1，$\eta$=0.1 とした．

文書のトピック分布 $\theta$ から，各文書のトピックで構成されるベクトルを作り，k-means 法により，20Newsgroups の対象とした 10 カテゴリのグループに文書を分類した際の精度を見ることで提案手法の評価を行う．

[1]http://qwone.com/ jason/20Newsgroups/

## 4.2 評価手法

評価手法には，文献 [4] で用いられている評価手法を採用し，式（7）に示される相互情報量を利用した．

$$MI(L,A) = \sum_{l_i \in L, \alpha_j \in A} P(l_i, \alpha_j) \cdot \log_2 \frac{P(l_i, \alpha_j)}{P(l_i)P(\alpha_j)} \tag{7}$$

$L = \{l_1, l_2, \cdots, l_k\}$ は，k-means 法により分類された文書ラベルの集合であり，$A = \{\alpha_1, \alpha_2, \cdots, \alpha_k\}$ は分類された文書の正解ラベルである．また，$P(l_i)$ は分類により $l_j$ にラベル付けされる確率，$P(\alpha_j)$ は正解データにおいて $\alpha_j$ である確率，$P(l_i, \alpha_j)$ はこれら 2 つが同時に起こる確率である．

ここで，相互情報量を [0,1] の値で得るために式（8）により正規化を行う．

$$\widehat{MI} = \frac{MI(L,A)}{MI(A,A)} \tag{8}$$

## 4.3 実験結果

k-means 法を用いた分類をそれぞれの手法の各閾値において 10 回ずつ行い，評価値 $\widehat{MI}$ の平均を求めた．setA での実験結果を図 2〜4，setB での実験結果を図 5〜7 に示す．全てのグラフに関して，横軸は閾値，縦軸は評価値 $\widehat{MI}$ を示す．なお，比較のために LDA の $\widehat{MI}$ もグラフに示す．

LDA を含めた全手法に関して，setA の方が setB を用いた実験よりも良い $\widehat{MI}$ を得ている．図 2，5 から分かるように，単語の共起情報を用いた手法では，どちらの文書集合においても LDA よりも $\widehat{MI}$ は低くなった．また，2 つのグラフに類似性は見られない．文書の類似度を用いた手法では，一部の閾値において LDA を上回る結果を得ている．図 3，6 より，[0.1,0.9] では，両手法ともに $\widehat{MI}$ は閾値によって減少増加し，どちら

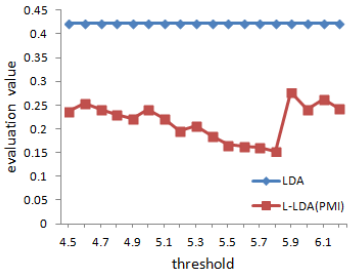


図 2: $\widehat{MI}$ パターン 1 (setA)

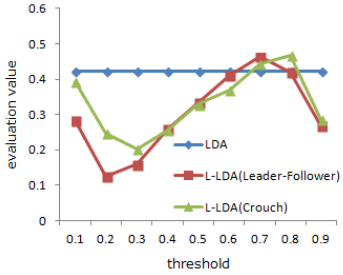


図 3: $\widehat{MI}$ パターン 2a(setA)

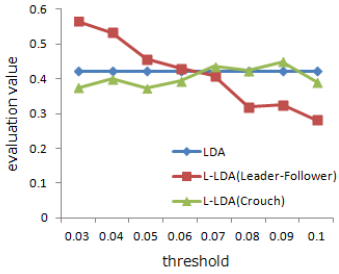


図 4: $\widehat{MI}$ パターン 2b(setA)

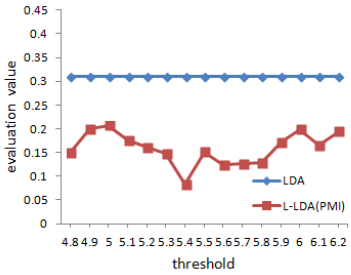


図 5: $\widehat{MI}$ パターン 1 (setB)

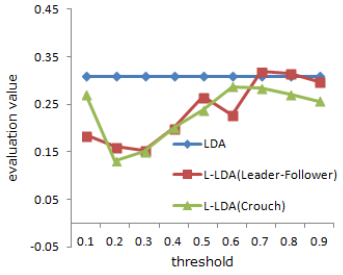


図 6: $\widehat{MI}$ パターン 2a(setB)

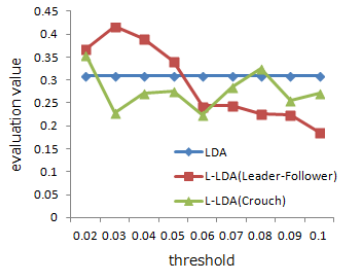


図 7: $\widehat{MI}$ パターン 2b(setB)

の文書集合においても，一部で LDA を上回る結果を得てはいるが，その値にあまり差はない．また，図 4，7 より，[0.03,0.1] では，Leader-Follower 法は左肩上がりのグラフとなっている．一方で，Crouch 法では LDA での $\widehat{MI}$ の付近で増加減少を繰り返すグラフとなっている．

次に，それぞれの手法において生成された疑似ラベルの数を表 2〜4 に示す． 表 2 のラベル数では，複数の単語によって構成されるラベルの数と，括弧の中に 1 つの単語で構成されるラベル数も加えたラベル数を示している．どちらの文書集合においても，閾値 5.2 でラベル数が増大し，5.6 でさらに急増する．そして，5.9 で減少した後，6.3 以上になるとラベルは生成されなくなった．パターン 2 では，Leader-Follower 法と Crouch 法でのクラスタを生成する過程が同じであるため，生成されるラベル数は同じである．どちら文書集合においても，閾値が大きくなるにつれてラベル数が増加し，0.2 で最大となった後に減少している．

表 2: パターン 1

| Threshold | Number of labels | |
|---|---|---|
| | setA | setB |
| 4.5 | 2(9) | - |
| 4.6 | 2(9) | - |
| 4.7 | 3(10) | - |
| 4.8 | 3(10) | 2(8) |
| 4.9 | 5(12) | 2(8) |
| 5.0 | 6(13) | 1(7) |
| 5.1 | 5(12) | 1(7) |
| 5.2 | 22(29) | 22(28) |
| 5.3 | 20(27) | 27(33) |
| 5.4 | 20(27) | 27(33) |
| 5.5 | 16(23) | 24(30) |
| 5.6 | 204(211) | 193(199) |
| 5.7 | 204(211) | 193(199) |
| 5.8 | 204(211) | 193(199) |
| 5.9 | 125(132) | 124(130) |
| 6.0 | 125(132) | 124(130) |
| 6.1 | 125(132) | 124(130) |
| 6.2 | 125(132) | 124(130) |

また，それぞれの文書集合における手法別の生成したラベル数と評価値 $\widehat{MI}$ の関係を散布図で表したものを図 8，9 に示す．グラフの横軸はラベル数，縦軸は評価値 $\widehat{MI}$ を表している．また，LDA は丸，単語の共起情報を用いた手法は菱形，文書の類似度を用いた手法のうち Leader-Follower 法を用いたものは正方形，Crouch 法を用いたものは三角形で表している．各手法ごとに着目してみると，菱形の点は，横軸 0〜50，100〜150，200 付近にまとまって存在し，ラベル数と評価値の相関関係は見られない．正方形では，例外も存在するが，基本的にラベル数が LDA での最適トピック数に近いほど，結果が良く，増加するほど悪くなっている，三角形では，正方形の点と比べて，安定した評価値を得ているが，最適トピックスとの差が大きくなるほど，そのバラツキも大きくなる．

表 3: パターン 2a

| Threshold | Number of labels | |
|---|---|---|
| | setA | setB |
| 0.1 | 185 | 212 |
| 0.2 | 228 | 220 |
| 0.3 | 202 | 179 |
| 0.4 | 155 | 140 |
| 0.5 | 102 | 93 |
| 0.6 | 59 | 56 |
| 0.7 | 28 | 25 |
| 0.8 | 11 | 9 |
| 0.9 | 2 | 6 |

表 4: パターン 2b

| Threshold | Number of labels | |
|---|---|---|
| | setA | setB |
| 0.02 | - | 6 |
| 0.03 | 19 | 21 |
| 0.04 | 42 | 42 |
| 0.05 | 69 | 80 |
| 0.06 | 101 | 119 |
| 0.07 | 118 | 144 |
| 0.08 | 151 | 181 |
| 0.09 | 169 | 194 |
| 0.1 | 185 | 212 |

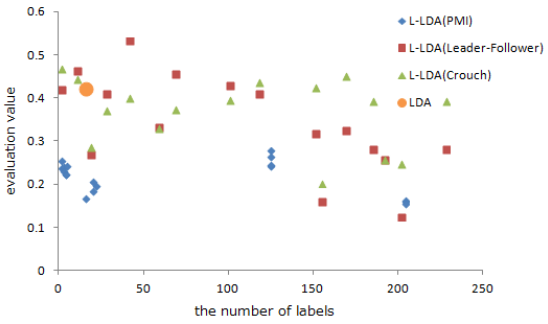


図 8: ラベル数と $\widehat{MI}$（setA）

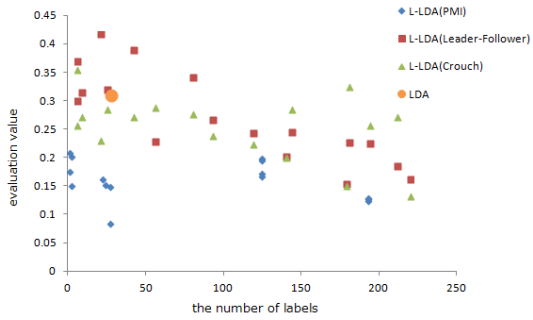


図 9: ラベル数と $\widehat{MI}$（setB）

## 5 考察

実験結果より，単語の共起情報から疑似ラベルを生成した場合では，2 つの文書集合において，全ての閾値で LDA と比べ精度が上がらなかった．また，ラベル数と評価値との相関関係が見られなかった．これらは，この手法で生成された疑似ラベルは，トピックの情報が反映できていないためであると考える．原因としては，抽出する単語数が少ないこと，もしくは，単語の共起情報のみでは文書集合の全てのトピックについて反映できないということが考えられる．閾値を低く設定した場合，トピックと関連の無い単語の共起情報も認識してしまい，トピックと関係の強い疑似ラベルを生成することが困難になり，一方で，閾値を高く設定した場合には，共起関係が非常に強い単語のみで疑似ラベルを作るため，生成された疑似ラベルはトピックとの関係は強いが，限られた文書にしかラベルを振り分けることができないとことが精度を悪くすると考える．また，今回はカテゴリごとに同じ数の文書を用意したが，文書数に偏りがあった場合には，トピックの情報を反映した疑似ラベルを生成することが，さらに困難になるのではないかと考えられる．

文書の類似度から疑似ラベルを生成した場合，Crouch 法を用いた方が Leader-Follower 法を用いるよりも，全体的に安定した評価値を得ている．また，2 つの文書集合での実験結果に共通して，閾値 0.03 で Leader-Follower 法を用いた場合に，LDA を上回り最も良い結果を得ている．閾値 0.03 では，生成された疑似ラベルの数が最適トピック数と近い値となっている．また setB においては，閾値 0.03 よりも 0.7 の方が，最適トピック数に近い数の疑似ラベルを生成しているが，この 2 つの点における評価値は 0.03 の方が良い．このことから，Leader-Follower 法を用いて，閾値をできるだけ低く設定し，生成される疑似ラベルの数が最適トピック数と近い場合に良い精度が得られることが分かった．

## 6 おわりに

本研究では，単語の共起情報と文書の類似度の 2 つの表層的な情報から疑似ラベルを生成した．それぞれの手法によって生成した疑似ラベルを用いて実験を行い，文書分類の課題を通じて LDA との精度の比較，評価を行った．その結果，単語の共起情報を用いた手法では，LDA を上回る結果を得ることはできなかったが，文書の類似度を用いた手法では，閾値を変えることによって，一部で LDA よりも良い精度を得ることができた．2 つの文書集合における評価結果から，それぞれの手法における閾値と評価値の関係や生成された疑似ラベル数と評価値の関係を確認した．また，2 つの文書集合における，全ての手法について比較したところ，Leader-Follower 法を用いた文書の類似度を利用した手法で，閾値を小さく設定し，生成される疑似ラベルの

数が最適トピック数に近かった場合に LDA を上回り，最も良い精度が得られることが分かった．

今後の課題としては，単語の共起情報を用いた手法において，生成したラベルを振り分けることのできる文書数が限られる原因として，TF-IDF を用いて抽出する単語の数が少ないことが考えられることから，設定を変えた実験が必要であると考えられる．また全ての提案手法に共通して，分類課題における別の評価方法や，他の課題を通した精度の確認をしていきたい．

## 参考文献

[1] D. M. Blei, A. Y. Ng, M. I. Jordan:Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

[2] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning:Labeled LDA: A supervised topic model for credit attribution in multi-label corpora. EMNLP2009, pp. 248-256, 2009.

[3] Newman, David and Lau, Jey Han and Grieser, Karl and Baldwin, Timothy: Human Language Technologies, NAACL2010,pp. 100–108,Los Angeles, California, 2010.

[4] Gunes Erkan:Language Model-Based Document Clustering Using Random Walks, Association for Computational Linguistics,pp.479-456,2006.

[5] 岸田和明: 大規模文献集合に対して階層的クラスタ分析法を適用するための単連結法アルゴリズム, Library and Information Science,No.47 ,2002.

[6] 岸田和明:文書クラスタリングの技法, Library and Information Science,No.49, 2003.

# 半教師あり学習における教師データ選出とグラフ構成

# High-Quality Training Data Selection and Graph Construction for Graph-based Semi-supervised Learning

江里口 瑛子 [1] *　　小林 一郎 [2]
Akiko Eriguchi[1]　　Ichiro Kobayashi[2]

[1] お茶の水女子大学理学部情報科学科
[1] Department of Information Sciences, Ochanomizu University
[2] お茶の水女子大学大学院人間文化創成科学研究科理学専攻
[2] Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

**Abstract:** We try raising the accuracy of multi-class document categorization using graph-based semi-supervised learning (GBSSL). With this end in view, we propose two methods. The first one is a method to construct a similarity graph by employing both surface information and latent information to express similarity between nodes. The second one is a method to select high-quality training data for GBSSL by means of PageRank algorithm. We experimented on Reuters-21578 corpus. We have confirmed that our proposed methods work well for raising the accuracy of multi-class document categorization.

## 1　序論

機械学習手法は，教師あり学習，教師なし学習，半教師あり学習などがある．中でも，グラフ構造に基づく半教師あり学習 (Graph-Based Semi-Supervised Learning: GBSSL) 法は，Support Vector Machine (SVM)[4] などの学習法と比べてより有効な手法であることが知られている [6]．GBSSL 法の精度は，一方で，グラフ構成の仕方によって左右され，他方で，その精度はどのような教師データ (ラベルありデータ) を与えるかによっても左右される．前者に関連して重要となるのは，グラフのノード間の関係性をどのように表現するかである [11]．後者に関連して重要となるのが，情報量の大きい教師データをどのように選出するかである．その良い事例が能動学習法であり，これは教師データの数は少ないが，質の高い教師データを選出する方であり，これによって GBSSL 法の精度が向上することが知られている [9]．

本研究は，多クラス文書分類における GBSSL 法の精度向上を目指すものであり，この手段として二つの方法を提示する．

まず第一は，グラフ構成に関連するものであり，グラフ構成に必須の要件である類似度に，文書間の潜在的な類似度を取り入れる．一般にこれまで，テキストデータからなるグラフを構成する際には，文書間の表層的な類似度が多く採用されてきたが，我々はこれに加えて新たに，文書間の潜在的な類似度を加えたものをノード間の類似度として採用する．

第二は，教師データの選出に関連するものであり，教師データからなる類似度グラフにおいて，各ノードにスコアを付けて質の高い教師データを選出する．すなわち，潜在情報を加味したグラフ上で，PageRank[2] 手法を用いて，質の高い教師データを選出する．ちなみに，ここで用いるグラフのノード間の類似度は，先述の文書間の潜在的な類似度を加味したものである．

以上の手法をマルチラベルを有するテキストのカテゴリ分類に適用し，PRBEP を算出し，我々の手法の有効性を各カテゴリ毎に評価し，かつ，それら全体の精度の向上を検討する．

*連絡先： お茶の水女子大学理学部情報科学科
〒 112-8610 東京都文京区大塚 2-1-1
E-mail: g0920506@is.ocha.ac.jp

## 2　グラフに基づく文書分類手法

### 2.1　グラフ構成

本研究におけるグラフ構成においては，テキストデータを対象にしたグラフ構成を行う．したがって，各文書はグラフのノードとみなされる．そのノード (文書) 間の関係は類似度として表され，その類似度をグラフの辺の重みとするような重み付き無向グラフ $G = (V, E)$

を構成する．ここで $V$ と $E$ は，それぞれグラフのノード集合と辺集合を表す．

グラフ $G$ は隣接行列 $\boldsymbol{W}$ の形で表現することができ，$w_{ij} \in \boldsymbol{W}$ はノード $i$，ノード $j$ 間の類似度を表すとする．特に，GBSSL 法の場合には，その類似度はノード $i$ の $k$-近傍点集合 $K(i)$ からなるものとし，$w_{ij} = sim(\boldsymbol{x}_i, \boldsymbol{x}_j)\delta(j \in K(i))$ とする．ここで，$\delta(z)$ は $z$ が真ならば 1，偽ならば 0 とする．

## 2.2 グラフにおける類似度

テキストデータにおける文書間の類似度を測る指標として，表層情報に基づく類似度と潜在情報に基づく類似度の二種類の類似度を採用する．文書の表層情報としては，文書に含まれる単語の出現頻度に着目した *tfidf* ベクトル [5] が多く用いられる．ここでは，表層情報に基づく類似度を，*tfidf* ベクトルのコサイン類似度 ($sim_{cos}$) の値とする．また，文書の潜在情報として，複数文書内に隠れトピックが存在することを仮定し，その隠れトピックに関して生起する単語の確率分布 (トピック分布) を用いる．ここでは，潜在情報に基づく類似度を，式 (2) によって得られる値 ($sim_{JS}$) とし，トピック分布間の距離は Jensen-Shannon ダイバージェンス ($D_{JS}$) を用いて求める．トピック分布の推定には，Latent Dirichlet Allocation (LDA) 法 [1] を用いる．

本研究では，この従来の類似度 ($sim_{cos}$) に新たに，文書の持つ潜在情報に基づいた類似度 ($sim_{JS}$) を $\alpha(0 \leq \alpha \leq 1)$ の割合で付加する．これら $sim_{JS}$ と $sim_{cos}$ を $\alpha : (1-\alpha)(0 \leq \alpha \leq 1)$ の割合で合算した値を，ノード間 (すなわち，文書 S と文書 T 間) の類似度 ($sim_{nodes}$) とする (式 (1))．P と Q は，それぞれ文書 S と文書 T に対するトピック分布を表す．

$$\begin{aligned} sim_{nodes}(S,T) &\equiv \alpha * sim_{JS}(P,Q) \\ &\quad +(1-\alpha) * sim_{cos}(tfidf(S), tfidf(T)) \quad (1) \end{aligned}$$

$$sim_{JS}(P,Q) \equiv 1 - D_{JS}(P,Q) \quad (2)$$

## 2.3 質の高い教師データの選出

質の高い教師データの選出法として，北島ら [12] によって提案された TopicRank 法を採用して行う．TopicRank 法とは，グラフ構造を用いた重要文抽出法の一つであり，類似度グラフのノードを単文とし，辺の重みを文間の潜在情報に基づく類似度として構成したグラフに対して，PageRank[2] の概念を用いて式 (3) により各ノード (各単文) の重要度を算出し，各ノードの順位付けを行う手法である．ここで，$d$ は制動係数 (damping factor) である．

本研究では，類似度グラフのノードを単文から文書 (文の集合) に置き換えて用いることとする．このため，式 (3) において，$N$ を対象文書群の総文書数，$adj[u]$ を文書 $u$ の隣接ノード集合とする．$sim_{nodes}(u,v)$ は，式 (1) によって求めた文書 $u$ と文書 $v$ の類似度である．その上で，文書のトピック分布を考慮した，教師データのみをノードにもつグラフをカテゴリ毎に作成し，TopicRank スコアが高いデータから順に，GBSSL 法で用いる教師データとしていく．

$$\begin{aligned} r(u) &= d \sum_{v \in adj[u]} \frac{sim_{nodes}(u,v)}{\sum_{z \in adj[v]} sim_{nodes}(z,v)} r(u) \\ &\quad + \frac{1-d}{N} \quad (3) \end{aligned}$$

## 2.4 ラベル伝搬法

本研究における GBSSL 法として，ラベル伝搬法 [7, 10] を採用する．ラベル伝搬法は,「グラフ上において，辺で繋がるノード同士は同じカテゴリに属す」という仮定に基づき，カテゴリラベル未知のノード (すなわち，テストデータ) について予測を行う手法である．

類似度行列を $\boldsymbol{W}$，ノード数を $n$ 個 (このうち教師データ数は $l$ 個) とする．$n$ 個のノードに対する予測値 $\boldsymbol{f}$ は，以下の最適化問題の目的関数 (式 (4)) の解 (式 (6)) として求まる．式 (4) の第 1 項は，各ノードの予測値と教師データの正解値の差を表し，第 2 項は，類似度グラフ上で隣接するノード同士の予測値の差を表す．$\lambda(>0)$ は両項のバランスをとる定数である．

式 (4) は $\boldsymbol{L}$ を用いて，式 (5) と変形できる．$\boldsymbol{L}(\equiv \boldsymbol{D} - \boldsymbol{W})$ はラプラシアン行列と呼ばれ，対角行列 $\boldsymbol{D}$ は $\boldsymbol{W}$ の各行 (又は列) の和を対角成分に持つ行列である．

$$\begin{aligned} J(\boldsymbol{f}) &= \sum_{i=1}^{l} (y^{(i)} - f^{(i)})^2 \\ &\quad + \lambda \sum_{i<j} w^{(i,j)} (f^{(i)} - f^{(j)})^2 \quad (4) \\ &= ||\boldsymbol{y} - \boldsymbol{f}||_2^2 + \lambda \boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} \quad (5) \end{aligned}$$

$$\boldsymbol{f} = (\boldsymbol{I} + \lambda \boldsymbol{L})^{-1} \boldsymbol{y} \quad (6)$$

# 3 実験

## 3.1 実験仕様

テキスト分類問題の対象データには，Reuters-21578 (Reuters)[1] を用いる．Reuters は 135 のトピックカテゴリからなる Reuters newswire の英文記事を集めたデータセットである．本実験では “ModApte”分割に従って，本文とタイトルのみからなる記事データを抽出し，全データに対してストップワードの除去とステミング処理を行う．その後，同じデータセットを用いて GBSSL 手法でマルチラベル文書分類を行っている Subramanya ら [6] の実験仕様に合わせ，10 種のカテゴリ **earn**，**acq**，**money-fx**，**grain**，**crude**，**trade**，**interest**，**ship**，**wheat**，**corn** に対する分類精度を求める．Reuters の記事データはマルチラベルを有するため，ここでは各カテゴリ毎に one-versus-rest 法を適用した二値分類を行い，一定の閾値以上のカテゴリラベルを文書に付与するラベルとして採用する．

データセットは，テストデータ (ラベルなしデータ)$u = 3299$ 個を共通とし，これに教師データ $l = 20$ 個を加えたものを 11 セット用意する．データセットに含まれるデータ総数は $n = 3319$ 個である．教師データとして加えるカテゴリは，上記 10 種のカテゴリにそれら以外のカテゴリ (**others**) を加えた全 11 種とする．データセットに加える教師データ $l$ 個のカテゴリは 11 種のカテゴリからランダムに選択するが，全 11 種のカテゴリの教師データが少なくとも 1 個ずつ含まれるように選択する．

TopicRank 法を用いる際の LDA 法における潜在トピックの推定方法には，ギブスサンプリングを用い，その反復回数は 200 回とする．トピック数はパープレキシティの値を算出し，その 10 回平均の値で決定する．また，TopicRank 法で用いるグラフは，ノード数 $|V|$ =(カテゴリ毎の教師データの総数)，辺数 $E = |V \times V|$ の完全グラフとする．パラメータ $\alpha$ は，0.0 から 1.0 まで 0.1 刻み毎の値を与え，制動係数 $d$ は Brin ら [2] の結果を参考に 0.85 とする．カテゴリ毎に各文書の TopicRank スコアを算出し，テストデータに加える教師データのカテゴリ数にしたがって，スコアの高い教師データから順にデータセットに加えていく．$\alpha = 0$ のときは文書の表層情報のみを扱い，推定を行う必要がない．このため，類似度が一意的に決まるのでスコアは 1 回のみ算出する．他方，$\alpha \neq 0$ のときは文書の潜在トピックの推定を行うため，類似度が一意的に決まらない．このため，5 回平均の値をスコアとする．

ラベル伝搬法で用いた類似度グラフのノード数は $|V| = n(= 3319)$ であり，ノード間の類似度は，パラメータ $\alpha = 0$ とし，表層情報のみからなるものとする．$k$-近傍グラフの大きさのパラメータ $k$ は $\{2, 10, 50, 100, 250, 500, 1000, 2000, n\}$，ラベル伝搬法のパラメータ $\lambda$ は $\{1, 0.1, 0.01, 1e-4, 1e-8\}$ の範囲を動かす．最初のデータセットによって，各カテゴリに対する最適パラメータ $(k, \lambda)$ の組を決定した後，それらのパラメータの値を用いて，残り 10 セットに対して文書分類を行い，各カテゴリ毎に PRBEP を求め，各試行毎の各カテゴリに対する PRBEP の平均値を算出する．指標 PRBEP は，*Precision*(適合率) と *Recall*(再現率) が一致するときの値である．

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

## 3.2 実験結果

$[0, 1]$ における 0.1 刻み毎の各 $\alpha$ の値に対して，カテゴリ毎に決定した最適パラメータ $(k, \lambda)$ を表 1 に示す．各カテゴリに対し，これらの最適パラメータを用いて行った実験結果を図 1〜10 に示す．横軸は $\alpha$ の値を表し，縦軸は PRBEP の値を表す．図 1〜10 は，各 $\alpha$ の値に対して行った 10 回の試行の各カテゴリ PRBEP の平均値を示している．図 11 は $\alpha = 0$ の PRBEP を指数 100 とした際の，各 $\alpha$ に対する各カテゴリにおける PRBEP の割合の変移を示している．各 $\alpha$ 毎に全カテゴリの PRBEP を合算して求め，その平均値の変移を図 12 に示す．図中のエラーバーは標準偏差を表す．

全ての図において，$\alpha = 0$ の場合は，表層情報のみを用いた場合の結果である．また，$\alpha = 1$ の場合は，潜在情報のみを用いた場合の結果である．それ以外 ($\alpha \neq 0$ または 1) は，潜在情報と表層情報を一定の割合 $(\alpha : (1 - \alpha))$ で混合した場合であり，両情報を用いた結果を示している．

まず，図 1〜10 に関連しては次の通りである．$\alpha = 0$ の時よりも，$\alpha \neq 0$ の時の PRBEP が大きい値をとるのは，図 4，5，6，8，10 である．ただし，図 4 では，$\alpha = 1$ の時の PRBEP は $\alpha = 0$ の時よりも小さい値をとる．他方，逆に $\alpha = 0$ の時よりも，$\alpha \neq 0$ の時の PRBEP が小さい値をとるのは，図 2，7 である．そして，$\alpha = 0$ の時の値に対して，$\alpha \neq 0$ の時の PRBEP が上下の変動を繰り返し，一意的な相関関係を見て取るのが難しいものは，図 1，3，9 である．

次に，図 11 から分かるように，$\alpha$ が増加するにつれて，正の相関が見られるものに関連して，PRBEP が最善で正方向に 200% も増加し，精度の向上が見られるものもあれば，他方，負の相関が見られるものもあり，最悪で約 80% 減殺され，PRBEP が減少しているものもある．

最後に，図 12 からは以下のことが分かる．マクロ平均値の極大値は 46.2，46.9，45.0(それぞれ $\alpha = 0.2, 0.6, 0.9$) であり，最大値は 46.9($\alpha = 0.6$ の時) である．最小値は

表 1: カテゴリ毎の最適パラメータ $(k, \lambda)$

| カテゴリ \ $\alpha$ | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *earn* | (500, 1) | (50, 1) | (1000, 1) | (1000, 1) | (50, 1) | (50, 1) | (50, 1) | (50, 1) | (50, 1) | (50, 1) | (50, 1) |
| *acq* | (100, 0.01) | (100, 0.01) | (100, 0.01) | (2, 1) | (100, 0.01) | (100, 0.01) | (100, 1e-8) | (100, 1e-8) | (100, 1e-8) | (100, 1e-8) | (100, 1e-8) |
| *money-fx* | (250, 0.01) | (100, 1e-8) | (10, 1e-4) | (100, 1e-8) | (2, 0.1) | (2, 0.1) | (2, 1e-8) | (250, 1e-8) | (2, 0.1) | (2, 1e-8) | (250, 1e-8) |
| *grain* | (250, 0.1) | (2000, 1e-4) | (100, 1) | (250, 0.1) | (100, 1) | (50, 1) | (250, 1) | (50, 1) | (50, 1) | (50, 1) | (100, 1) |
| *crude* | (50, 0.1) | (2, 1) | (250, 0.01) | (50, 1e-8) | (10, 0.01) | (250, 0.01) | (250, 0.01) | (250, 1e-8) | (10, 0.01) | (250, 0.01) | (250, 0.01) |
| *trade* | (2, 1) | (10, 0.1) | (50, 0.01) | (10, 1e-8) | (10, 1e-8) | (10, 1e-8) | (50, 1e-8) | (10, 1e-8) | (10, 1e-4) | (10, 0.1) | (10, 0.1) |
| *interest* | (10, 1) | (50, 1e-8) | (50, 1e-8) | (10, 1) | (2, 0.1) | (250, 1e-8) | (250, 0.01) | (250, 0.01) | (2, 1) | (2, 0.1) | (500, 1e-8) |
| *ship* | (3318, 1) | (50, 1) | (50, 1) | (250, 0.1) | (50, 0.1) | (50, 0.1) | (50, 1e-8) | (50, 1e-8) | (100, 0.1) | (100, 0.1) | (50, 0.01) |
| *wheat* | (500, 1e-8) | (500, 1e-8) | (250, 1e-8) | (500, 1e-8) | (500, 0.01) | (1000, 0.01) | (500, 1e-8) | (250, 1e-8) | (250, 1e-8) | (250, 1e-8) | (250, 1e-8) |
| *corn* | (10, 1e-8) | (100, 1e-8) | (250, 1e-8) | (10, 1e-8) | (250, 1e-8) | (250, 1e-4) | (500, 1e-8) | (100, 1e-8) | (250, 1e-8) | (50, 0.01) | (250, 1e-4) |

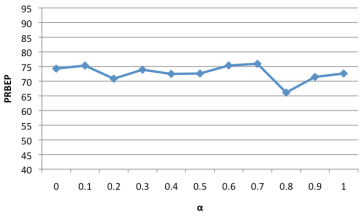


図 1: **earn** の平均 PRBEP

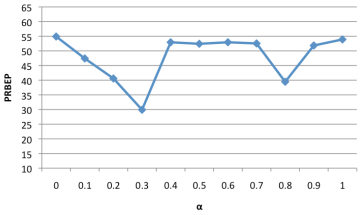


図 2: **acq** の平均 PRBEP

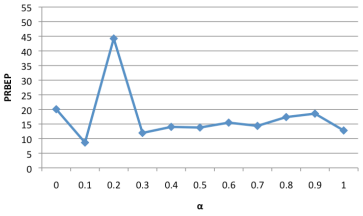


図 3: **money-fx** の平均 PRBEP

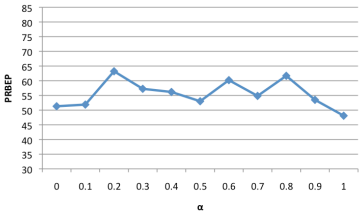


図 4: **grain** の平均 PRBEP

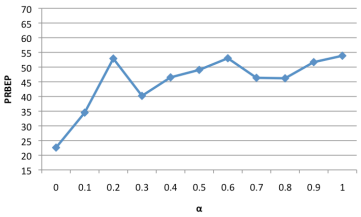


図 5: **crude** の平均 PRBEP

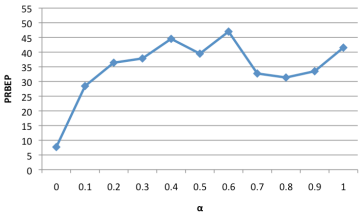


図 6: **trade** の平均 PRBEP

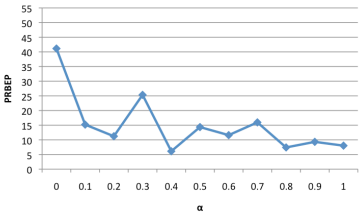


図 7: **interest** の平均 PRBEP

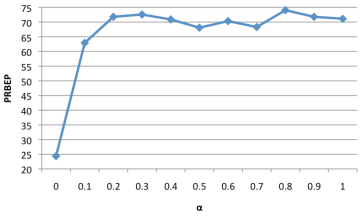


図 8: **ship** の平均 PRBEP

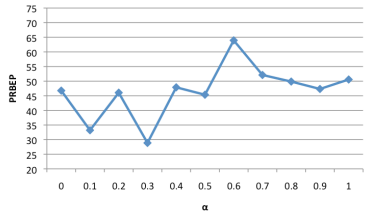


図 9: **wheat** の平均 PRBEP

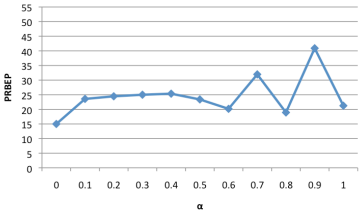


図 10: **corn** の平均 PRBEP

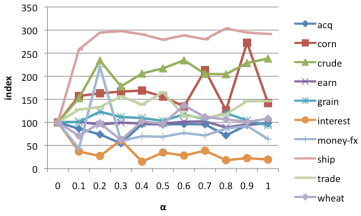


図 11: $\alpha$=0 の PRBEP を指数 100 とした時の PRBEP の割合

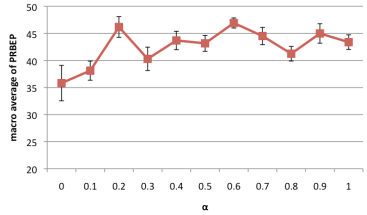


図 12: 全カテゴリの平均 PRBEP

35.8($\alpha = 0$) である．ただし，$\alpha = 1$ の時の値は 43.4 である．したがって，最大マクロ平均値 46.9($\alpha = 0.6$) は $\alpha = 1$ の時より 3.5%高く，更に $\alpha = 0$ の時より 11.1%高いことが分かる．また，$\alpha = 0$〜0.2 の時，マクロ平均値は単調増加しており (35.8 → 46.2)，$\alpha = 0.2$ 以上では，マクロ平均値は一定の範囲 (40.3〜46.9) を浮動している．以上から，$0.1 \leq \alpha \leq 1$ の時のマクロ平均値は $\alpha = 0$ の時よりも大きい．更に重要なことは，$\alpha = 0.2, 0.4, 0.6, 0.7, 0.9$ の時のマクロ平均値は，$\alpha = 0$ の時よりも大きいことは勿論，$\alpha = 1$ の時よりも大きいということである．

## 4 考察

図 1〜10 の各図において，PRBEP が最大値をとる時の $\alpha$ の値は各カテゴリ毎に異なっており，一律ではない．故に，精度が最大となる時の，$\alpha$ の値 (すなわち表層情報と潜在情報の混合割合) を一意的に決めることは難しい．しかしながら，全体的に見た場合には，一定の傾向性や関係性が見て取れる．図 11 においては，半数以上のカテゴリで PRBEP は増加傾向を示している．更に，大半のカテゴリにおいて，$\alpha \geq 0.1$ における PRBEP は $\alpha = 0$ の時の値 100 より大きい．この故に，全カテゴリのマクロ平均値をとった時には右肩上がりの変化パターンが期待される．図 12 は，全カテゴリのマクロ平均 PRBEP が示している．$\alpha = 0$ の時のマクロ平均値をベースラインとすると，$\alpha = 1$ の時の値は 7.6%増加しており，$\alpha = 0.6$ の時の最大値では 11.1%も増加している．加えて，$\alpha = 0.2, 0.4, 0.6, 0.7, 0.9$ の時の値は $\alpha = 1$ の時の値よりも大きい．

以上のことから，教師データの選出を行う際には，表層情報のみを用いるよりも潜在情報を用いる方が GBSSL 法の精度は向上することが分かる．また，両情報を用いる方が，潜在情報のみを用いるよりも精度は向上する．したがって，両情報の混合割合 $\alpha$ の最適値が求まりさえすれば，単に表層情報や潜在情報のみを用いる場合よりも，高い精度が得られるだろう．

## 5 結論

我々は，表層情報と潜在情報に基づく類似度グラフの構成法，並びに，GBSSL 法で用いる質の高い教師データの選出法を提案した．Reuters-21578 コーパスを用いた実験の結果から，教師データの選出には，表層情報と潜在情報のどちらかだけを用いるよりも，両情報を混合させて同時に用いた方が GBSSL 法における文書分類の精度を向上させることが分かった．

今後の課題としては，我々が今回得た結論 (表層情報と潜在情報の両情報を用いる方がそれらを単体で用いるよりも精度が高い) を他のデータセットを用いて検証することであり，最適パラメータ $(k, \lambda)$ における決定の仕方を改善することであり，ラベル伝搬法で用いるグラフ構成にも潜在情報を活用することによって，更なる精度の向上を図ることである．

## 参考文献

[1] Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation, *Journal of Machine Learning Research* (2003)

[2] Brin, S., Page, L.: The Anatomy of a Large-scale Hypertextual Web Search Engine., *Computer Networks and ISDN Systems*, pp. 107–117 (1998)

[3] Erkan, G., Radev, D. R.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, *Journal of Artificial Intelligence Research 22*, pp.457-479 (2004)

[4] Cortes, C., Vapnik, V.: Support-vector networks, *Machine Learning*,20: 273-297 (1995)

[5] Salton, G., McGill, J.: Introduction to Modern Information Retrieval, McGraw-Hill (1983)

[6] Subramanya, A., Bilmes, J.: Soft-Supervised Learning for Text Classification, in *Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp.1090–1099 (2008)

[7] Zhou, D., Bousquet, O., Lal, T. N., Weston J., Schölkopf B.: Learning with Local and Global Consistency, in *NIPS 16* (2004)

[8] Zhu, X., Ghahramani, Z.: Learning from Labeled and Unlabeled Data with Label Propagation, Technical report, Carnegie Mellon University (2002)

[9] Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions, in *Proc. of the International Conference on Machine Learning (ICML)* (2003)

[10] Zhu, X., Ghahramani, Z., Lafferty, J. Semi-supervised learning using Gaussian fields and harmonic functions, In *ICML* (2003)

[11] Zhu, X.: Semi-Supervised Learning with Graphs, PhD thesis, Carnegie Mellon University (2005)

[12] 北島理沙, 小林一郎: 潜在的意味を考慮したグラフに基づく複数文書要約, *Proceeding of ARG WI2*, (2012)

# 単語の共起グラフを用いた潜在的意味に基づく効果的な文書分類への取り組み

# A Study on Efficient Text Classification Based on Latent Semantic Used a Graph of Co-occurring Terms

小倉由佳里 [1*]　　小林一郎 [2]
Yukari Ogura[1]　　Ichiro Kobayashi[2]

[1] お茶の水女子大学理学部情報科学科
[1] Dept. of Information Sciences, Faculty of Science, Ochanomizu University
[2] お茶の水女子大学大学院人間文化創成科学研究科理学専攻
[2] Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

**Abstract:** In this paper, we propose a method to raise the accuracy of text classification based on latent topics, reconsidering the techniques necessary for good classification – for example, to decide important sentences in a document, the sentences with important words are usually regarded as important sentences. In this case, *tf.idf* is often used to decide important words. On the other hand, we apply the PageRank algorithm to rank important words in each document. Furthermore, before clustering documents, we refine the target documents by representing them as a collection of important sentences in each document. We then classify the documents based on latent information in the documents. As a clustering method, we employ the k-means algorithm and investigate how our proposed method works for good clustering.

## 1　はじめに

近年，インターネットの発達に伴い，爆発的に増大した莫大な量のテキストデータを扱う問題がある．そのため大量のテキストを，自動でカテゴリごとに分類できるような文書分類手法が必要とされている．本研究では，文書の潜在的意味を考慮した分類手法を提案する．文書分類の方針として，まず語彙の重要度に基づき重要文抽出を行い，元の文書を重要文のみで構成し，分類対象となる文書の精錬化を図る．語彙の重要度を決める指標としては，一般に $tf \cdot idf$ や語彙の頻度などが用いられるが，本研究では，語の共起関係からグラフを構成し，PageRank アルゴリズムを用いて重要語の決定を行う．次に，潜在的意味解析手法を用いて，文書の潜在トピックごとの確率分布をもとに，k-means 法でクラスタリングを行う．実験を行い，語の重要度の決定に PageRank を用いた場合と，$tf \cdot idf$ を用いた場合の文書分類の精度を比較することにより，提案手法の有効性を検討する．

*連絡先：お茶の水女子大学理学部情報科学科小林研究室
〒 112-8610 東京都文京区大塚 2-1-1
E-mail: g0920509@is.ocha.ac.jp

## 2　関連研究

文書分類の研究において，分類精度を上げるため数多くの研究がなされており，特に，文書中の語の重要度を決めるアルゴリズムを改良することにより，分類精度の向上が出来ることが報告されている．Hassan ら [1] は，n-グラムを用いて，単語の共起関係に基づき構成したグラフに，PageRank アルゴリズムを用いることにより，文書分類の精度が向上することを示した．Zaiane ら [2] や，Wang ら [3] は，文書分類における，語の重要度の決定手法を提案した．Wang ら [3] は，語の重要度の決定に PageRank アルゴリズムを用いることが，文書分類に有効であることを示した．PageRank アルゴリズムは，センチメント分析や，トピック推定にも用いられており，Kubek ら [4] は，語の共起関係に基づき構成したグラフに，PageRank アルゴリズムを用いることにより，トピック推定を行っている．語の重みづけは，文書要約やにおいても重要な課題である．Erkan ら [5] は，LexRank や TextRank と呼ばれる，PageRank アルゴリズムを用いた文書要約の手法を提案している．文をノードとしてグラフを構成し，高い PageRank スコアを持つ，中心性の高い文を抽出することにより，文書要約を行っている．

本研究では，文書を潜在情報に基づいて分類することを目的とし，Newman ら [8] による潜在的情報の首尾一貫性は単語の共起関係により形成されるという報告を参考に，共起語からなるグラフを構築し，それに PageRank アルゴリズムを適用することにより，抽出された重要語から重要文を決定する．その重要文を用いて，潜在情報に敏感な文書群を再構成し，文書分類を行う手法を提案する．

# 3 提案方法

## 3.1 PageRank アルゴリズムによる重要語の決定

PageRank とは，Brin ら [6] によって提案された，Web ページ間に存在するハイパーリンク関係を利用することでページの順位付けを行うアルゴリズムである．PageRank の基本的な考え方は，推薦である．例として，図 1 の場合，$V_a$ から $V_b$ へリンクが張られているため，これは $V_a$ から $V_b$ への推薦と考えることができる．他の重要な Web ページから推薦されている Web ページは重要である，という考え方が PageRank において中心となっている概念である．Web ページをノード，ページ間のリンク関係をエッジとした有向グラフとして構成され，このグラフに基づいて順位のスコアが計算される．グラフ $G = (V, E)$ が与えられたときに，$In(V_a)$ は，点 $V_a$ を指している点の集合，$Out(V_a)$ は，点 $V_a$ が指している点の集合である．点 $V_a$ の PageRank スコアは，式 (1) を反復的に処理することにより，全てのノードの PageRank スコアを求める．$d$ は，制動係数 (dumping factor) であり，ある一定の割合でリンクのないノードからの影響を考慮するパラメータであり，$[0, 1]$ の値をとる．

$$S(V_a) = \frac{(1-d)}{N} + d * \sum_{V_b \in In(V_a)} \frac{S(V_b)}{|Out(V_b)|} \qquad (1)$$

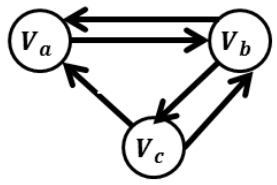


図 1: リンク関係の例

反復計算には，べき乗法を用いる．べき乗法とは，行列の主固有値と主固有ベクトルを見つけるための反復法であり，マルコフ連鎖の定常ベクトルがマルコフ行列の左側主固有ベクトルであること，および，求めたい PageRank ベクトルが Web ページ間のリンク関係を表した推移行列をもつマルコフ連鎖の定常ベクトルであることにより，PageRank の計算に用いられる．

語の重要度を決定するのには，$tf \cdot idf$ などが頻繁に用いられるが，語同士の様々な関係をグラフ構造で表現し，語の重要度を決定する手法が提案されている [3][1][10]．特に，Hassan ら [1] は，PageRank を用いてランクづけされた語の重要度は $tf \cdot idf$ よりも重要度を明確に差別化できることを示している．本研究でも彼らの手法を参考にして，語の重要度を PageRank アルゴリズムを用いて決定する．

## 3.2 潜在情報による分類

文書内の潜在的トピックの確率分布を表わすモデルとして Latent Dirichlet Allocation(LDA)[7] がある．このモデルでは，文書内にはいくつものトピックが潜在しており，トピックごとに出現しやすい単語があると考える．各トピックはそのトピックに対する出現確率を持った単語群で表され，複数文書内に存在している総単語に対して，各トピックごとに総和が 1 になる出現確率が割り当てられる．トピック自身にも文書セット内において出現確率の総和が 1 となるトピック比率として確率が付与される．本研究においては，文書に対する潜在トピックの確率分布を用いて，各文書をトピックで構成されるベクトルで表現し，文書間の類似度を測る．

## 3.3 提案手法における処理の流れ

本手法における，文書分類の流れを説明する．

**step1** 単語の共起関係の抽出
文書を文で区切り，文脈を考慮して，文中の単語の共起度を自己相互情報量 (PMI:Point-wise Mutual Information) に基づき算出する．

**step2** 重要単語の決定
step1 で得られた共起関係に基づき，ノードを単語，エッジの重みには PMI を用いたグラフを構成する．図 2 は，共起関係を基に構成したグラフの一例である．ここで，グラフを単語間の PMI で構成する理由は，文書分類を潜在的意味に基づき行うとしており，潜在トピックの一貫性は語の共起関係が影響を与えているとする Newman ら [8] の研究に基づき，潜在トピックを考慮した単語の重要度を算出するためである．このグラフに対し，多くの単語と高い共起度を持つ単語は重要

であると考え，PageRank アルゴリズムを用い，単語の重要度のランク付けを行う．

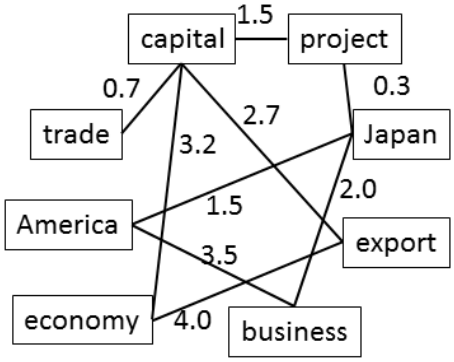


図 2: 類似度グラフ

**step3** 重要文の抽出
step2 で得られた単語のランキングに基づき，ランキング上位の単語を含む文を重要文とみなし，これを文書から抽出し，元の文書を重要文のみで構成する．

**step4** 分類
新たに構成された文書群に対し，LDA を用いてそれぞれの文書の潜在トピックごとの確率分布を得る．各文書のトピックに基づくベクトルを Jensen-Shannon 距離を用いて類似度を測り，k-means 法により分類する．

# 4 実験

## 4.1 実験仕様

実験対象データには，Reuters-21578[1] のテストセットからタグを除去したものを使用した．提案する手法は，対象文書から重要文を抽出し，文書を精錬してから文書分類を行うため，文数の少ない文書では提案手法の効果が判別できないため，1 文書中の文章数が 5 文以上である文書を利用した．カテゴリは，文書分類の他研究 [9]，[11] においても用いられている上位 10 件のカテゴリ，acq，corn，crude，earn，grain，interest，money-fx，ship，trade，wheat を利用した．その結果，文書数 792 件，語彙数 15,835 語，カテゴリ数 10 の文書群を対象に，ステミング処理とストップワード除去を施し実験を行った．

また，LDA で用いるパラメータは，$\alpha = 0.5$，$\beta = 0.5$ とし，サンプリングにはギブスサンプリングを用い，イテレーションは 200 回とした．トピック数は，パープレキシティにより決定することにした．トピック数を 1 から 30 まで変化させたときのパープレキシティの値の 10 回の平均をとり，パープレキシティが最小になるときのトピック数を最適トピック数とした．計算の結果，元の文書群のトピック数が 11 となった．重要文の抽出を行わない元の文書群の分類精度をベースラインとするため，実験に使用するトピック数は 11 とした．分類手法には，k-means 法を用い，トピックで構成された文書ベクトルを用いて分類を行う．

[1] http://www.daviddlewis.com/resources/testcollections/reuter21578

## 4.2 評価手法

評価には，文献 [9] を参考にして，正解率と F 値の 2 つの評価指標を用いる．文書 $d_i$ に関して，$l_i$ はクラスタリングアルゴリズムにより $d_i$ に与えられたラベル，$\alpha_i$ は $d_i$ の正解のラベルである．そのとき，正解率は式 (2) で表される．

$$\text{正解率} = \frac{\sum_{i=1}^{n} \delta\left(map\left(l_i\right), \alpha_i\right)}{n} \tag{2}$$

$\delta(x, y)$ は，$x = y$ ならば 1 となり，そうでなければ 0 となる関数である．$map(l_i)$ は，k-means 法により $d_i$ に与えられるラベルである．

評価には，各カテゴリの F 値を求め，全カテゴリの平均を算出した．カテゴリ $c_i$ の F 値は，精度を $P(c_i)$，再現率を $R(c_i)$ とすると，式 (3) のように表される．

$$F\left(c_i\right) = \frac{2 \cdot P\left(c_i\right) \cdot R\left(c_i\right)}{P\left(c_i\right) + R\left(c_i\right)} \tag{3}$$

カテゴリごとの F 値 (式 (3)) を測り，全カテゴリの平均を評価指標として用いた．(式 (4))

$$F = \frac{1}{|C|} \sum_{c_i \in C} F\left(c_i\right) \tag{4}$$

また，k-means 法において初期値には，それぞれのカテゴリの正解データの文書ベクトルをランダムに選び，1 つ与えることにする．分類する際，文書群におけるカテゴリ数 k を事前に知っていること，それぞれのカテゴリから 1 つだけ正解例を見つけることは，計算コストがかからないことから，妥当な方法であると判断できる．この方法により，分類結果のクラスが，どのカテゴリであるか判断できるようになる．

## 4.3 実験結果

k-means 法を 10 回行い，その平均値を測った．ただし，LDA を用いて，文書のトピックごとの確率分布から分類を行う場合には，出力される確率分布 $\theta$ が毎回

変化する．そのため 1 つの $\theta$ に対して k-means 法を 10 回行い，これを 8 セット行ったときの平均値を測った．確率分布を用いて分類を行う場合には Jensen-Shannon 距離を，文書ベクトルを用いて分類を行う場合にはコサイン類似度を用いた．重要文抽出を行った場合の結果を表 1，行っていない場合の結果を表 2 に示す．また，重要文を行った後の文書群の語彙数，文数の変化をそれぞれ表 3，表 4 に示す．また，表 5，表 6 では，PageRank，$tf \cdot idf$ のそれぞれの指標を用いて重要単語のランクを決定し，それに基づき，同じ数だけ文を抽出し，トピック数を変化させて分類を行った場合の実験結果を示す．

表 1: 重要文抽出した場合

| 単語の重要度 | 類似度指標 | 正解率 | F 値 |
|---|---|---|---|
| PageRank | Jenshen-Shannon 距離 | **0.5671** | **0.4852** |
| | コサイン類似度 | 0.2870 | 0.2906 |
| $tf \cdot idf$ | Jenshen-Shannon 距離 | 0.5500 | 0.4347 |
| | コサイン類似度 | 0.2753 | 0.2701 |

表 2: 重要文抽出しない場合

| 類似度指標 | 正解率 | F 値 |
|---|---|---|
| Jenshen-Shannon 距離 | 0.5177 | 0.4262 |
| コサイン類似度 | 0.2875 | 0.3048 |

表 3: 語彙数の変化

| 手法 | 1 語 | 2 語 | 3 語 | 4 語 | 5 語 |
|---|---|---|---|---|---|
| PageRank | 12,268 | 13,141 | 13,589 | 13,738 | 13,895 |
| $tf \cdot idf$ | 13,999 | 14,573 | 14,446 | 14,675 | 14,688 |

表 4: 文数の変化

| 手法 | 1 語 | 2 語 | 3 語 | 4 語 | 5 語 |
|---|---|---|---|---|---|
| PageRank | 1,244 | 1,392 | 1,470 | 1,512 | 1,535 |
| $tf \cdot idf$ | 1,462 | 1,586 | 1,621 | 1,643 | 1,647 |

表 5: 正解率

| トピック数 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|
| PageRank | 0.525 | 0.535 | **0.566** | 0.553 | 0.524 |
| *tf.idf* | 0.556 | 0.525 | 0.557 | 0.550 | 0.541 |

# 考察

実験結果より，Jenshen-Shannon 距離を用いて分類を行った場合においては，重要文抽出を行った場合の方が，行わない場合よりも正解率，F 値ともに値が良くなるということが分かった．このことから，重要文抽出することにより，文書が精錬されていることが確認された．文書が精錬されたことにより，文書の特徴を表現するのに必要な文のみが残り，文書のトピックごとの確率分布の差が測りやすくなったのではないかと考えられる．また重要文抽出に関して，$tf \cdot idf$ を用いた場合に比べ，PageRank を用いて重要文の抽出を行った場合に文書分類の精度の向上が見られた．このことから，文書の 3 文中での単語の共起関係からグラフを構成し，単語の重要度を PageRank アルゴリズムを用いて決定することにより，文脈を考慮した単語の重要度が得られていると考えられる．

また表 3，表 4 から，重要文抽出したあとの語彙数，文数の比較では，$tf \cdot idf$ と比較して，PageRank を用いた場合に，より語彙数，文数が減っていることが分かる．$tf \cdot idf$ の場合，特定の文書に多く出現している単語の値が高くなるため，$tf \cdot idf$ が高い単語は，その文書中の多くの文に出現している可能性が高い．そのため，$tf \cdot idf$ の高い単語を含む文を抽出すると，自然と多くの文を抽出することになるのではないかと考えられる．

コサイン類似度で分類を行った場合において，精度が良くなかった原因としては，実験結果から，Jenshen-Shannon 距離と比較すると，文書間の類似度の値の差が小さいことが観測されており，そのため異なるカテゴリの文書の判別がうまくいかなかったのではないかと考えられる．

文書群から抽出する文数を同じにし，トピック数を変化させて分類を行った場合，トピック数が 9,10,11 の時に $tf \cdot idf$ を用いた場合よりも，PageRank を用いた場合に分類精度が上回った．また，PageRank，$tf \cdot idf$，どちらを用いた場合でもトピック数を 10 に設定した時に，高い精度となった．これは，文書群の正解カテゴリが 10 であることから，正解のカテゴリ数と設定したトピック数が一致しているため，分類精度が高くなったのではないかと考えられる．トピック数 10 の時に PageRank と $tf \cdot idf$ での結果を比較すると，PageRank を用いた場合に良い精度となっている．このことからも，PageRank アルゴリズムにより決定した単語の重要度が，分類精度の向上に寄与することが分かる．

表 6: F 値

| トピック数 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|
| PageRank | 0.431 | 0.431 | **0.467** | 0.460 | 0.434 |
| *tf.idf* | 0.466 | 0.430 | 0.461 | 0.435 | 0.445 |

## 5 おわりに

本研究では，PageRank を用いた重要語の抽出を行い, それに基づいて重要文を抽出し, 潜在的意味によるクラスタリングを行う手法を提案した．分類対象データとして，Reuters-21578 を用いて実験を行った．提案手法の有効性を検証するため，重要文の抽出に語の重要度を PageRank，または $tf \cdot idf$ を用いて行い，重要文によって再構成された文書集合に対して，潜在情報または表層情報に基づき，k-means 法を用いてクラスタリングを行った．その結果，全体として表層情報よりも潜在情報を用いた分類の方が精度が良く，重要文を抽出する際に，語の重要度を PageRank を用いて決める方が分類精度が向上することがわかった．重要文抽出した後の語彙数の比較から，PageRank を用いた場合のほうが抽出される文章数が少ないということが推測できるため，より文脈を考慮した重要文の抽出がなされているのではないかと考えられる．

今後の課題としては, どの程度の重要文をどのように選択したかにより, 分類の精度が変化すると考えられるため, 適切な重要文選別方法を考察するつもりである．また, 現在は k-means 法での分類しか行っていないため, 他の多クラス分類手法との比較を行うつもりである．

## 参考文献

[1] Samer Hassan, Rada Mihalcea, Carmen Banea.: Random-Walk Term Weighting for Improved Text Classification, (2007)

[2] Osmar R.Zaiane, Maria-luiza Antonie.: Classifying Text Documents by Associating Terms with Text Categories, *In Proc. of the Thirteenth Australasian Database Conference(ADC'02)*, pp. 215–222

[3] Wei Wang, Diep Bich Do, and Xuemin Lin.: Term Graph Model for Text Classification, *Springer-Verlag Berlin Heidelberg 2005*, pp. 19–30 (2005)

[4] Mario Kubek, Herwig Unger.: Topic Detection Based on the PageRank's Clustering Property, *IICS'11*, pp. 139–148 (2011)

[5] Gunes Erkan.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, *Journal of Artificial Intelligence Research 22*, pp. 457–486 (2004)

[6] Sergey Brin, Lawrence Page.: The Anatony of a Large-scale Hypertextual Web Search Engine, *Computer Networks and ISDN Systems*, pp. 107–117 (1998)

[7] David M. Blei, Andrew Y. Ng, Michael I. Jordan, John Lafferty.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, p. 993–1022 (2003)

[8] Newman David, Lau Jey Han, Grieser karl, Baldwin Timothy.: Automatic evaluation of topic coherence, *Human Language Technologies :The 2010 Annual Conference of the North Ametican Chapter of the Association for Computational Linguistics*, pp. 100–108 (2010)

[9] Gunes Erkan.: Language Model-Based Document Clustering Using Random Walks, *Association for Computational Linguistics*, pp. 479–486 (2006)

[10] Christian Scheible, Hinrich Shutze.: Bootstrapping Sentiment Labels For Unannotated Documents With Polarity PageRank, *Proceedings of the Eight International Conference on Language Resources and Evaluation*, (2012)

[11] Amarnag Subramanya, Jeff Bilmes.: Soft-Supervised Learning for Text Classification, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 1090–1099, Honolulu (2008)

# 電子カルテにおける新人とベテランの特徴比較支援システム

# Feature Interpretation Support for Comparing Newcomers and Veterans in Electronic Health Records

谷恵里香　砂山渡*
Erika Tani　　Wataru Sunayama

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:** The electronic medical records are written by nurses. There is a difference between the newcomer's description and veteran's description. In this study, the system creates the medical record sets of newcomers and veterans. The system supports users to discover features and differences in each medical record set. By using the system, newcomers can learn how veterans write the electronic medical record. Specifically, the system shrinks by the year of the length of experience and keywords that medical records include. The system draws and displays maps from shrunk electronic medical record sets. In addition, the system displays the words that are contained in medical record sets or in the map.

## 1　はじめに

看護士が記述する電子カルテ（看護記録）には，患者の多くの情報が記入される．例えば，患者の関心，注目すべき行動，気がかりな事，問題，心配，重要な出来事，患者に起こった，あるいは起こりうる状態などが挙げられる．これは，医師および看護士自身が患者の病態を把握し，適切な治療を行うために必要なものとなっている．そのため，看護士だけが理解できるものでは用をなさない．また，記述者によって内容の質や量が異なり，記入漏れがあるといった問題も示唆される．更に，カルテを記述する上で患者の注目すべき点等は，経験を積み重ねなければわからない．そこでより良いカルテを作成するために電子カルテの監査が行われている．

しかし，看護記録は患者一人一人に対して毎日記述を行うため，監査を行うべきカルテの情報量は膨大になる．そのため監査を行うのに多くの時間がかかり非効率的である．そこで本研究では，新人とベテランのカルテに注目し，新人では着目できない点や，もっと記述しておくべき点などをベテランのカルテと比較を行うことで見つけ出す．本研究では，カルテで記述された単語に着目し，単語から新人とベテランそれぞれの特徴を見出す．単語から，新人とベテランの特徴が発見できればカルテを監査するための基準（記述上必須単語，所見すべき患者の体の箇所など）を作ることができると考えた．また，この監査すべき基準ができれば本システムで単語抽出が行えるので，カルテを自動で評価して結果を提示することができる．

*連絡先：〒 731-3194 広島県広島市安佐南区大塚東 3-4-1

## 2　関連研究

### 2.1　カルテの監査に関する研究

カルテの監査をチェック項目を使用して行う研究 [1, 2] がある．これら研究は，一つのカルテに対して記述しなければいけない項目を満たしているかどうかをチェックする監査方法であった．チェックを行うことでカルテの記述において不十分な点を明らかにし，より良いカルテの記述に近づけようといった研究であった．本研究はカルテの内容の良しあしという基準は設けず，カルテ集合同士の比較を行うという点で異なる．

### 2.2　新人看護士へのカルテ記述支援に関する研究

過去のカルテから経過記録が類似する患者情報をシステムで発見し，情報を提示する研究 [3] がある．これは，過去の経過記録と類似している患者に対して今後起こりうる事象を予測し教示するものである．また，新人看護士の記述したカルテを対象にしているという点で本研究と類似している．電子カルテの監査において新人とベテランのカルテを比較することで新人が記述したカルテの中で足りない部分や，今後を予測してどういったことを記述しておくべきかを考えるための

支援という点で一致している．本研究では，カルテ集合から監査を行いたい人間が自ら条件を入力し，比較したいカルテ集合を作成することができる．また，作成したカルテ集合の特徴を地図の可視化によって視覚的にわかりやすくするという点で異なる．

### 2.3 テキストデータマイニングにおける電子カルテの監査に関する研究

テキストデータマイニングによって電子カルテの監査を行う研究 [4] がある．これはカルテの文章から病状経過や治療行為に関する単語を抽出し，単語の関係性を地図として描き可視化する．可視化によって膨大な量の文章を人間が視覚的に解釈しやすくするための支援となる．本研究は，カルテ集合の特徴を地図の可視化で示すという点で類似している．しかし，異なる点は，条件の入力によって，比較を行いたいカルテ集合を作成する．そのカルテ集合から地図が作成できるという点が１つ．比較を行う２つのカルテ集合のデータからカルテを特徴づける単語を提示し，新たな着目点の発見を助けるという点．地図の表示および単語の表示の２つの可視化画面を利用することでさらに監査を行う際に注目すべきであろう単語を視覚的に示しているという点で異なる．

## 3 単語特徴比較支援システム

### 3.1 単語特徴比較支援システムの構成

図１に単語特徴比較支援システムの全体構成を示す．本システムでは，カルテのテキストデータを扱う．図１の入力１，入力２で，ユーザがカルテの絞り込み条件を入力する．システム動作部分では，条件によって絞り込んだ新人とベテランのカルテ集合の作成を行い，地図の作成および単語の抽出を行う．出力部分では，２つのカルテ集合より得た単語から単語比較画面を出力する．出力結果から新人とベテランの特徴を発見し，この特徴から新人とベテランの比較を行うことでカルテの監査の支援を行う．

以下では本システムの各処理について述べる．

### 3.2 カルテの絞り込み条件

カルテを絞り込むための条件には，新人の経験年数と，ベテランの経験年数およびカルテに含まれるキーワードを指定して与える（図１入力１，２）．この条件入力により電子カルテデータを絞り込み，新人とベテランの電子カルテ集合を作成する．なおカルテの絞り込みには，カルテを記述した看護士の名前と経験年数のデータを用いる．また電子カルテには，記述者名，記述日時，診療科名，患者の発言や症状，処置内容，検査結果，結果から考えられること，今後の治療方針といった入院患者の経過内容が含まれることを想定している．

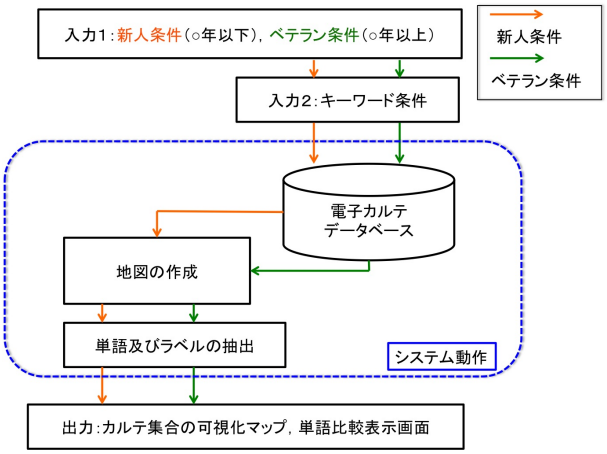


図 1: 単語特徴比較支援のシステム構成

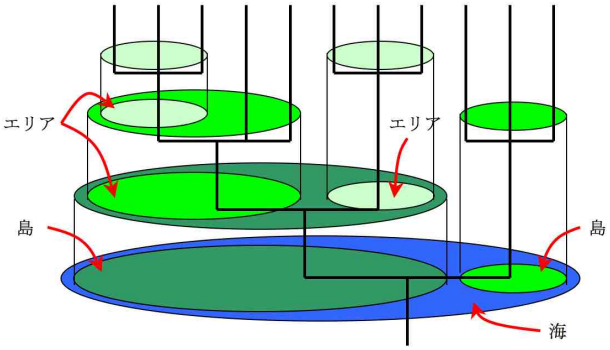


図 2: 逆さにしたデンドログラムと島，エリアとの関係

### 3.3 地図の生成

前節で作成された新人とベテランの電子カルテ集合を把握しやすくするため、それぞれの電子カルテ集合をクラスタリングによって分類して地図として可視化する．地図の生成には，既存手法の再帰的クラスタリング [5] を用いて，できるだけ複数のクラスタにテキストをばらけさせカルテ集合の構造を把握しやすくする．

図２に，クラスタリングの結果（デンドログラム）と地図生成のためのエリアとの関係図を示す．また生成されたマップの例を図３中央に示す．マップ上では，１つのカルテをノード（●）として表し，カルテ内で最も頻度が高い名詞をノード名として表示する．また，クラスタリングの結果として得られるクラスタ（カルテの集合）を白線で囲みエリアとして，そのエリアのラベルを，クラスタ内で最も多くのカルテに出現する名詞として表示する[1]．

[1]エリアのラベル名は，上位（多くのカルテを含むクラスタ）か

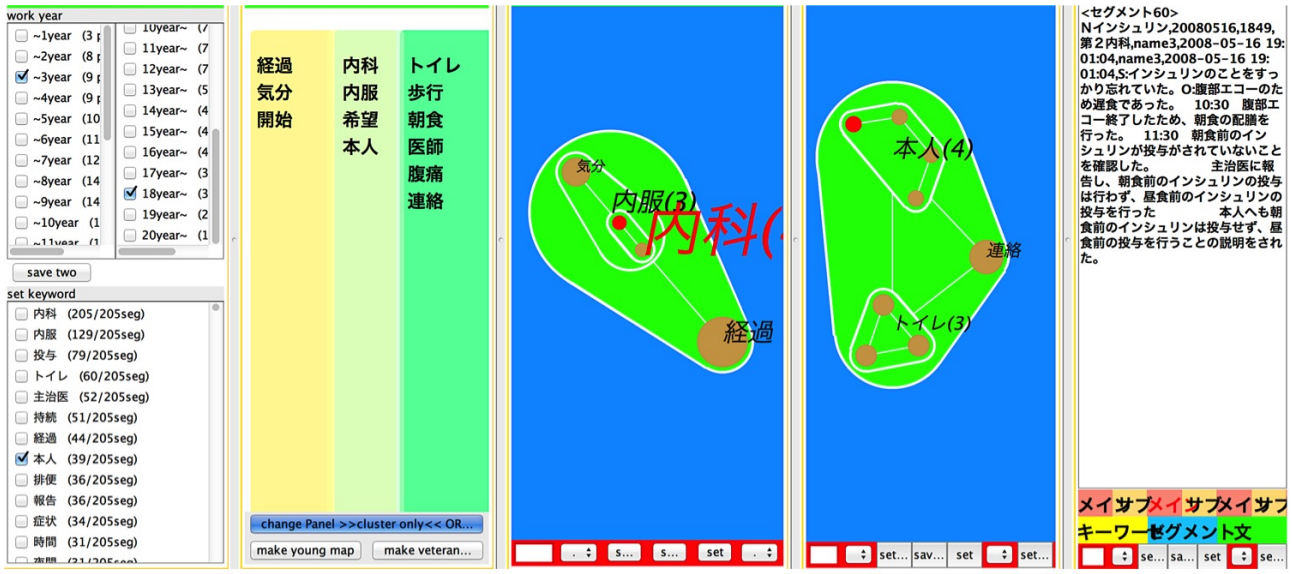


図 3: 出力表示画面 ((a):選択パネル (b):単語比較パネル (c):新人のカルテ集合から得たマップ表示パネル (d):ベテランのカルテ集合から得たマップ表示パネル (e):カルテ内容提示パネル)

## 3.4 単語の抽出

本節では，絞り込んだ複数のカルテ集合を比較するための単語を抽出する方法について述べる．比較に使用する単語は，電子カルテ内で用いられているすべての単語を対象とする場合と，前節で述べた地図のラベルを用いる場合の 2 種類があり，このそれぞれについて説明する．

### 3.4.1 全単語の抽出

全単語の抽出では，新人とベテランのカルテ集合それぞれで使用された全ての単語を比較対象として抽出する．また，比較インタフェース上で，新人，ベテランのそれぞれの特徴を表す順に表示するために，各単語の出現頻度を用いる．

### 3.4.2 ラベリングした単語の抽出

ラベリングした単語の抽出では，地図の生成を行った際にラベリングされたノード名とクラスタ名を抽出する．そして，ノード名とクラスタ名を合わせてラベル単語という．また，比較インタフェース上で，新人，ベテランのそれぞれの特徴を表す順に表示するために，各単語の出現頻度を用いる．ラベル単語のみを用いることで，全単語を対象とした場合に比べ，電子カルテ集合の特徴を表す単語に絞って比較を可能にすることを意図している．

## 3.5 出力インタフェース

本節では，本システムの出力インタフェースおよび使用例を説明する．

出力インタフェースは図 3 で，左から選択パネル，単語比較画面，新人の可視化マップ表示画面，ベテランの可視化マップ表示画面，カルテ情報の表示画面となっている．

### 3.5.1 単語比較画面

単語比較画面（図 3 の (b)）は，抽出を行った単語に対して，単語の出現頻度を求め，新人とベテランそれぞれで求められた頻度の差を求める．求めた差から，新人の出現頻度が高いものを表示画面左部のエリアに，ベテランの出現頻度が高いものを表示画面右部のエリアに，頻度が同程度のものを真ん中のエリアに表示する．

単語比較画面は 2 パターンある．1 つは，全単語を抽出し，そこから頻度の算出，表示を行うパターン．もう 1 つは，ラベル単語を抽出し，そこから頻度の算出，表示を行うパターンである．

ら優先して与え，ラベル名には重複がないようにしている．

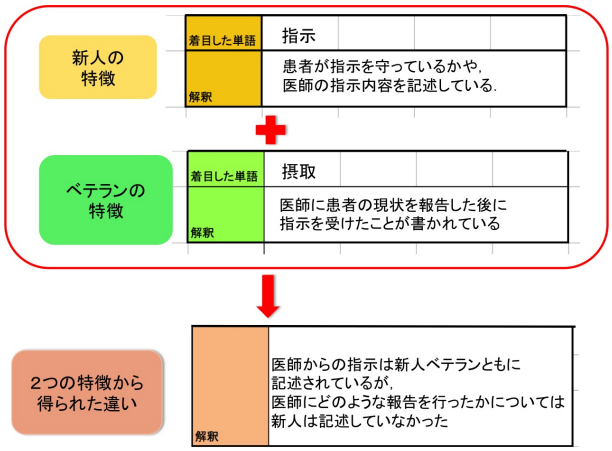


図 4: 新人とベテランの特徴の解釈および違いの解釈の記述例

### 3.5.2 カルテ集合の可視化マップ

カルテ集合から作成した可視化マップは２つ表示画面を提示しており（図 3 の (c),(d)），出力画面 (c) には新人のカルテ集合の可視化マップを表示し，出力画面 (d) には，ベテランのカルテ集合の可視化マップを表示している．

### 3.5.3 システム使用手順

システム使用手順の例を以下に示す．

1. 経験年数の条件入力およびキーワード条件入力（出力表示画面 (a)）
2. 可視化されたマップの違いの解釈，マップからカルテ情報の表示（出力表示画面 (c,d,e)）
3. 単語比較画面の出力単語を見て着目したい単語の決定（出力表示画面 (b)）
4. カルテの特徴の解釈

条件入力で，新人の経験年数の選択（図 3 条件選択パネル A）およびベテランの経験年数の選択（図 3 出力表示 B）を行う (1)．また，適宜キーワードの選択（図 3 出力表示 C）を行う (1)．その後可視化されたマップの見た目の違いや，マップから得られる情報（カルテの内容，クラスタ名，ノード名，どのカルテが関係しているのか）を見る (2)．また，単語比較画面で表示されている単語に注目する (3)．(1)，(2)，(3) の手順を何度か行い，(4) で絞り込まれたカルテの内容の特徴を解釈する（4）．

表 1: 経験年数ごとの看護士の人数とカルテデータ数

| 経験年数 (年) | 人数 (人) | データ数 |
|---|---|---|
| 1 | 3 | 21 |
| 2 | 5 | 57 |
| 3-6 | 3 | 32 |
| 7-9 | 3 | 24 |
| 10-15 | 2 | 11 |
| 16-20 | 3 | 45 |
| 合計 | 19 | 204 |

## 4 単語特徴比較支援システムの有効性を検証する評価実験

本章では，提案システムの有効性を検証するための評価実験について述べる．

### 4.1 実験手順

ある病院の 204 の電子カルテデータ（表 1）を用いて，20 名の大学生，大学院生に以下の手順に従って新人とベテランの違いの解釈を行ってもらう実験を行った．

**Step1** 新人およびベテランのカルテの特徴を解釈する

**Step2** Step1 で記述した内容を比べて，新人とベテランのカルテの特徴の違いを解釈する．

被験者には，入力した条件によって絞り込まれた新人，ベテランのカルテ集合の特徴を解釈してもらった．その後２つの特徴を踏まえたうえで新人とベテランの違いについて解釈してもらいこれらを解答用紙に記述してもらった．新人，ベテランそれぞれの特徴を記述する際，解答用紙には，選択パネルから選択した経験年数，選択したキーワード，また，単語比較画面パネルから着目した単語を記述してもらった．その後，選択した条件からどういった解釈を得たかを最大５項目記述してもらった．新人，ベテランの特徴から違いを解釈してもらう際，解答用紙には，解釈に使用した項目番号と，解釈内容を記述してもらった．また，解釈に使用する単語の違いによる解釈結果の違いを考察するため，20 名を 10 名ずつの 2 グループに分け，全単語を使用するグループと，地図状のラベル単語を使用するグループとに分けて実験を行った．

## 4.2 実験結果

### 4.2.1 新人およびベテランの特徴

被験者によって挙げられた新人とベテランの特徴の解釈結果例を表 4 と表 5 に示す．得られた特徴を 4 つの項目に分類することができた．4 つの項目に分類した際の記述数を表 2，表 3 に示す．

- 患者の状態：患者の発言，症状，容態について
- 処置，処方：患者に対して行った治療内容や，薬の投与，検査について
- 記述方法：内容というよりは端的に記述しているかどうか，主観的な記述の仕方をしているかについて
- その他：診断結果，医師とのやりとり，医師の行動について

また，上で挙げた項目を複数組み合わせて記述していた場合もあった．表 2，表 3 の分類項目数を比較すると，被験者の多くは，記述内容（患者の状態，処置，処方についてなど）に注目し，新人，ベテランそれぞれの特徴を解釈していた．表 2，表 3 の各項目における記述数を比較するとラベル単語比較グループは全単語比較グループに比べ上記項目の単体で解釈を行うだけでなく，複数の観点から解釈を行っていた．対して全単語比較グループはカルテの内容に観点を重点的に置いて解釈を行っていた．また，各パターンでの解釈時間には差が無いにも関わらずラベル単語比較グループはより多くの観点で特徴を解釈できている．これは，ラベル単語比較の単語表示によって効率よく解釈が行え，内容だけでなく記述方法などの他の観点に目を向ける余裕ができたのだと考えられる．以上より，さまざまな観点で解釈を行う際は，ラベル単語比較パターンの使用が時間を効率的に使い，より深い考察ができると言える．また，内容に特化した解釈を行う際は，全単語比較パターンの使用がより多くの特徴を列挙することができると言える．

### 4.2.2 新人とベテランの特徴の比較

前節より得た新人，ベテランの特徴を踏まえ，違いを解釈してもらった．

解釈内容についても，新人とベテランの特徴を分類したときと同様の基準で 4 つの項目に分類した結果を，表 7 に示す．

表 7 に被験者が解答した解釈分類と解釈数を示した．これより，全単語グループでは解釈数に対して内容に関する解釈が多く，ラベル単語グループでは記述書式についての解釈が多いことがわかった．

表 2: 全単語比較グループの解釈分類と記述数

| 分類 | 新人 | ベテラン | 合計 |
|---|---|---|---|
| 患者の状態 (1) | 12 | 16 | 28 |
| 処置，処方 (2) | 14 | 11 | 25 |
| 記述方法 (3) | 6 | 6 | 12 |
| (1),(2) の組合せ | 3 | 6 | 9 |
| (1),(3) の組合せ | 1 | 0 | 1 |
| その他 | 4 | 3 | 7 |
| 合計 | 40 | 42 | 82 |

表 3: ラベル単語比較グループ解釈分類

| 分類 | 新人 | ベテラン | 合計 |
|---|---|---|---|
| 患者の状態 (1) | 13 | 9 | 22 |
| 処置，処方 (2) | 7 | 8 | 15 |
| 記述方法 (3) | 9 | 6 | 15 |
| (1),(2) の組合せ | 11 | 7 | 17 |
| (1),(3) の組合せ | 3 | 6 | 9 |
| (2),(3) の組合せ | 0 | 3 | 3 |
| (1),(2),(3) の組合せ | 0 | 1 | 1 |
| その他 | 1 | 2 | 4 |
| 合計 | 44 | 42 | 86 |

全単語比較は，解釈数に対して「患者の状態についての詳細さ」および「処置，処方についての詳細さ」といった内容に関する解釈が多かった．これは，全単語比較画面でしか出現しない単語が存在するためラベル単語では出現しなかった出現頻度の高い単語を見つけることができたために内容に関する解釈を多くなすことができたと考えられる．

ラベル単語比較は，ラベル名，つまりカルテの特徴を示した単語を提示している．ラベル名の提示によってより効率的に解釈を行うことができ，カルテの内容の他に記述書式といった他の観点に着目できたのだと考える．

つまり，全単語表示画面では単純に多く使用されている単語が表示されるため，ここで表示された単語の中から経験を経ることでしか得られない着目点や注意すべき点に関係した単語を発見することができる．そして，ラベル単語表示画面ではカルテの特徴を示す単語を発見することができ，また，書式にも着目できるため，記述すべき基本内容や基本書式に関係した単語を発見することができる．ラベル単語比較で示した単語（クラスタ名，ノード名）はカルテの特徴を示した単語となっている．ラベル単語グループでは，カルテの特徴を表す単語を示したために，

以上より，経験を経ることでしか得られない着目点

表 4: 全単語比較グループの被験者記述例

| 種別 | 記述例 |
|---|---|
| 新人 | 患者が気分を訴えた際，患者がどう言ったかは報告されているが，具体的には書かれていない． |
| 新人 | 患者に投与した薬の量や患者の尿量について書いている． |
| 新人 | ほぼ患者なのかもしれないが，主語がない文が多い． |
| ベテラン | 患者本人への（処置や薬などの）説明だけでなく，親や家族への説明を行っていることも記述してある． |
| ベテラン | 時間の情報とともに，投与した薬の量や体温などの情報もセットで記述しているところが多く見られる． |
| ベテラン | 説明が長い人と短い人がいる． |

表 5: ラベル単語比較グループの被験者記述例

| 種別 | 記述例 |
|---|---|
| 新人 | 患者本人の症状についての記述が多く，症状以外の記述は少ない． |
| 新人 | 目標量のクリアについて書かれている．持続投与出来るかどうか書かれている． |
| 新人 | 主観的な感想が多い． |
| ベテラン | 患者の行動や状態に対して，表情や時間なども細かく記載されているものが多く見られた． |
| ベテラン | 時間を正確に区切り，予定をしっかりとどめている．具体的な容量をとどめている． |
| ベテラン | 必要事項もにを端的に記している． |

表 6: 解釈に用いられた単語の手法間の重複率

| 手法 | 重複率 |
|---|---|
| 全単語 | 31/60(52%) |
| ラベル単語 | 33/64(52%) |

表 7: 新人とベテランの特徴比較時の解釈分類

| | 全単語 | ラベル単語 | 合計 |
|---|---|---|---|
| 患者の状態 (1) | 16 | 18 | 50 |
| 処置，処方 (2) | 8 | 8 | 16 |
| 記述方法 (3) | 6 | 14 | 20 |
| その他 | 5 | 2 | 7 |
| 合計 | 35 | 42 | 77 |

や注目すべき点といったより内容に特化した内容の特徴単語を発見したい場合には，全単語比較の表示画面を用いたシステムの利用が有効的であると考える．

また，基本内容，書式といったより一般的な内容の特徴単語を発見したい場合にラベル単語比較の表示画面を用いたシステムの利用が有効的であると考える．

#### 4.2.3 監査基準の作成

実際に単語比較システムを使用して得られた考察から監査基準の作成例を考える．

まず，全単語比較グループについて監査基準を作成する．表 8 の全単語 4 を見ると「母親」「家族」「説明」という単語に着目し，ベテランは家族への説明を行ったという内容もきちんと記述されていた．そこで監査基準を「家族への説明をきちんと行ったかどうか」とし,「家族」や「説明」という単語を監査基準単語とする．この単語の使用の有無を確かめることによって新人がこの監査基準をクリアしているかどうかを発見することができると考える．

次に，ラベル単語比較グループについて監査基準を作成する．表 8 のラベル単語 4 を見るとベテランはカルテの「Object」客観的データの部分をきちんと客観的に記載していた．そこで監査の基準を「曖昧な表現を使用していないか，主観的な考察を行っていないか」とし,「など」「よう」「思う」という単語を監査基準単語とする．この単語の使用の有無で監査基準をクリアしているかどうかを発見することができると考える．

## 5 結論

電子カルテの情報から新人とベテランのカルテ集合の可視化を行い，特徴比較を行うためのシステムを提案した．評価実験により，医学の知識を持っていない人でもさまざまな観点からカルテの特徴比較を行うことができることを確認した．今回の実験では経験年数を選択可能にしていた．しかし，経験年数が低くても質の高いカルテを記述する看護士もいるため，さらに看護士の名前で条件を絞り込むことによって更により詳細な比較ができると考える．また今後は，この比較結果をもとに電子カルテの自動監査につなげるため，実際の医師らに使用してもらうことを想定している．

表 8: 新人とベテランの違いの解釈の記述例

| グループ | 記述例 |
|---|---|
| 全単語 1 | 同じ患者本人が訴える症状を記録するにしても，新人は言った言葉そのままを記録しているが，<br>ベテランはそれに加えてさらに詳しく記録している． |
| 全単語 2 | 投薬以外の治療について，ベテランの方が詳しく記録している． |
| 全単語 3 | 新人は主語がぬけていたり，患者についてもおおまかに表現していることがあった．<br>ベテランは誰が誰にいつ何をしたかを明確にできている． |
| 全単語 4 | 新人は患者本人に関する対応についてのみ記述される傾向にあるが，<br>　ベテランは，患者本人に関する対応だけでなく，<br>患者の母親や家族への説明を行ったことも記述される傾向にある． |
| ラベル単語 1 | 病院側の目標については新人の方がかけているが，患者一人一人の状態については<br>ベテランの方がよく書けているように感じた． |
| ラベル単語 2 | 検査の内容や処置方法は書かれている．ベテランは結果も書かれている． |
| ラベル単語 3 | ベテランは主観をいれず事実を記載する． |
| ラベル単語 4 | ２年以下の新人は，医師の指示を意識した処置内容，<br>その後の患者の様子が記載されているが若干曖昧な印象がある．<br>しかし，ベテランの方では，同様に患者の様子が記載されているが，<br>客観的な記載のされ方がしてあり，患者の容態の理解がしやすくなっている． |

## 参考文献

[1] 東富佐乃，飛田敦子：看護記録監査を基にした記録改善への取り組み，看護研究発表論文集録，pp.137-140(2005)

[2] 元永智子，末森節子，藤井美智子，川石文子，江本しず子：記録監査表からみた精神科看護記録の現状と分析と改善への取り組み:山口大学医学部附属病院看護部研究論文集会 82 巻，AA12138758，pp.102-107，(2012)

[3] 木原崇博，仲谷善雄:問題志向型看護録に基づく新人看護師への看護推薦支援の試み:情報処理学会第 73 回全国大会報告，2Z-3，(2011)

[4] 串間宗夫，荒木賢二，鈴木斎王，荒木早苗，二鎌照絵：電子カルテ入院患者看護記録のテキストデータマイニング，第 26 回人口知能学会全国大会報告，3K2-NFC-3-1，(2012)

[5] 砂山渡，濱岡秀平，奥田澄：情報収集のためのテキストデータ集合の再帰的クラスタリング，日本知能情報ファジィ学会誌, Vol.24, No.3, pp.697 – 706, (2012)

[6] 山本浩子，岡田淳子，小池伝一，吉田和美，川西美佐，楠広子：新人看護師の電子カルテを用いた診療記録活用における課題，日本赤十字広島看護大学紀要，第 12 巻，pp.19-26，(2012)

[7] 森田敏子，松永保子：社会・医療・看護の変化と看護記録記載基準，月刊看護きろく，第 16 巻 8 号，pp.3-13，(2006)

[8] 田中粛美，山本紀代子，藤野純子，花田千鶴美，黒田由利子：看護過程支援システムの現状と課題：従来の記録と比較・検討して，院内看護研究発表会集録，pp.78-82，(2000)

# 多人数向けメッセージからの失礼表現の自動抽出

# Extraction of Impolite Expression from Messages to Multipersons

安藤律子　砂山渡*
Ritsuko Ando　　Wataru Sunayama

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:** It is desired that messages don't become impolite when you send a message towards many unspecified persons. This paper aimed to extract the most likely sentence that contains impolite expression from message such as BBS and Twitter. The system uses the set of words that are more likely to be impolite expression to judge whether a message is impolite. The result of the system promotes the reconsideration of message contents to user. In addition, the result helps to avoid a contribution of the impolite expression.

## 1　はじめに

近年，文章を用いたインターネットコミュニケーションが，盛んに行われている．例えば，電子掲示板やTwitterが挙げられる．これらの利用者は，不特定多数の人に向け，様々な内容のメッセージを発信する．しかし，メッセージの発信者は，読みだけに徹しているの読者の存在に気付きにくいことや，ネット特有の匿名性によって，無責任な発言や批判的な発言をしてしまう [1]．そういった状況下で，良好なコミュニケーションとるためには，メッセージの発信者が，発信するメッセージの表現を読者に対し，失礼にならないように配慮をすることが望まれる．しかし，メッセージの表現が，知らないうちに読者への配慮が足りない表現となることや，発信するメッセージ内容が，大量な場合，失礼な表現が含まれる文の存在に気付きにくく，見落とし易くなることが考えられる．そこで本研究では，電子掲示板やTwitter等，多人数向けメッセージの集合を対象に，失礼表現が含まれている可能性が高い文 (以下，失礼文と記述する) を抽出することを目的とする．システムの使用者に，発信予定のメッセージ集合に含まれる失礼文の存在を提示し，使用者の失礼文の見落としを防ぎ，メッセージ内容の見直しをアシストする．

## 2　関連研究

### 2.1　電子掲示板，Twitterに関する研究

本節では，電子掲示板，Twitterに関する研究について述べる．電子掲示板を対象としたコミュニケーションの雰囲気に関する研究 [2, 3] がある．これらの研究は，電子掲示板に投稿されている発言から，電子掲示板の雰囲気を判定し，システムの使用者に提示を行うものであった．また，Twitterから利用者個人の特徴を抽出する研究 [4, 5] がある．これらの研究は，投稿されているツイートから，ツイートした人物の特徴を得て，システムの使用者に提示を行うものであった．4つの研究を挙げたが，電子掲示板やTwitterに関する研究は，既に投稿された文から，読者に対し，有益な情報を与える研究が多かった．本研究では，電子掲示板やTwitterに文を投稿する前に，文を投稿する人に対して，その文が失礼文かどうかの情報を与える．

### 2.2　文章の抽出に関する研究

本節では，文の抽出に関する研究について述べる．各研究で，対象とする文の抽出を行う研究 [6, 7] がある．これらの研究では，手がかり語や語句間の関係性，係り受け関係にある文節に着目し，対象とする文を抽出していた．本研究では，語句間の関係や文節係り受け関係には着目せず，文中で使用されている単語に着目して，失礼文の抽出を行う．

*連絡先：〒 731-3194 広島県広島市安佐南区大塚東 3-4-1

### 2.3 表現の判定に関する研究

本節では，表現の判定に関する研究について述べる．入力に対し，誹謗中傷表現を判定する研究 [8]，不満表現を判定する研究 [9]，有害情報を判定する研究 [10] がある．これらの研究は，人手で収集した単語に，表現の度合いを示す値を付与したり，収集した単語とは逆の意味を持つ単語を再度収集，比較したりすることで表現の判定を行うものであった．本研究では，これらの研究と同じく，アンケートをとり，人手で単語収集を行った．しかし，表現の判定には，失礼表現の度合いを示す値の付与や，失礼表現とならない単語の収集の手間を省き，その単語が 1 文に含まれているか，含まれていないかで行っている．

## 3 失礼文自動抽出システム

### 3.1 失礼文の定義

本研究で抽出の対象としている失礼文の定義について述べる．一般に失礼とは，他人に接する際の心得をわきまえていないこと，また，礼儀に欠けることを指す．すなわち，人間関係や社会生活の秩序を維持するために守るべき行動が欠けていることが失礼に当たると考えられる．そこで本研究では,『他者や物に対して，必要以上に否定的な表現，または蔑視的な表現』を失礼表現と定め，このような失礼表現を含む文を失礼文として定義する．

### 3.2 システム構成

本節では，システムの構成について述べる．図 1 にシステムの構成を示す．本システムは，入力された多人数向けメッセージの集合から，辞書内の単語を 1 つでも含む文を失礼文と判定し，色付けをして出力する．失礼文の判定には，失礼単語辞書を用いる．失礼単語辞書は，失礼表現となる可能性がある単語を集めた辞書である．3.4 項にて，失礼文を抽出する際に用いる失礼単語辞書の作成について述べる．

### 3.3 入力：多人数向けメッセージの集合

本節では，失礼文自動抽出システムへの入力について述べる．本システムには，入力として，多人数向けメッセージを記述したテキストを与える．本研究では，多人数向けメッセージを電子掲示板や Twitter 等，多くの人々に読まれる可能性がある場に投稿するメッセージとする．

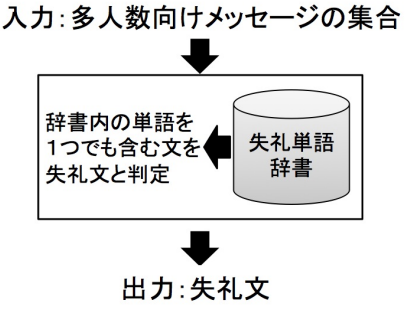


図 1: 失礼文自動抽出システムの構成

### 3.4 失礼単語辞書の作成

本節では，失礼分の判定の用いる失礼単語辞書を作成した方法について述べる．以下，失礼単語辞書の作成した方法について、既存の辞書からの失礼単語の収集と失礼単語の追加の順に述べる．

#### 3.4.1 失礼単語の収集

失礼単語の収集について述べる．失礼単語は，失礼文の定義に沿った名詞，形容詞とし，単語感情極性対応表 [11] から収集する．単語感情極性対応表とは，ある単語が一般的に良い印象を持つか悪い印象を持つかを，感情極性値-1 から 1 で表したものになる．この感情極性値が正の値ならば良い印象を，負の値ならば悪い印象を表している．一般的に悪い印象を持つ単語は，失礼表現となる単語の可能性が高いと考え，その単語を元にした．失礼単語は，単語感情極性対応表内の感情極性値が-0.48 以下の単語，14035 単語についてアンケートをとり，その結果から収集した．アンケートは，情報科学を専攻する大学生・大学院生の男女 6 人に，単語について，その単語を含む文が次の 4 つの項目のいずれに当てはまるかを評価してもらった．

(1) ほぼ間違いなく失礼な文になる

(2) 場合によっては失礼な文になる

(3) 基本的に失礼な文にならない

(4) 単語の意味がよく分からない

アンケートの回答に点数を割り振り，(1) の 1 回答につき 3 点，(2) の回答数が 1 つにつき 1 点とした．この時点で，1 単語当たりの回答人数や合計点から単語を振り分け，それぞれ以下の条件が当てはまる 10 種類の辞書 (辞書を構成する単語数) を用意した．

これらの辞書について，適切な辞書を決定するための予備実験を行った．10 種類の辞書を用いて失礼文を

表 1: 用意した辞書とアンケート結果との関係一覧

| | (1) の人数 | (2) の人数 | 単語数 |
|---|---|---|---|
| 辞書 1 | 3 人以上 | | 325 |
| 辞書 2 | 3 人以上 | または，3 人以上 | 878 |
| 辞書 3 | 1 人以上 | | 1068 |
| 辞書 4 | 1 人以上 | または，3 人以上 | 1380 |
| 辞書 5 | 1 人以上 | または，1 人以上 | 2014 |
| 辞書 6 | 2 人以上 | または，3 人以上 | 1052 |
| 辞書 7 | 1 人以上 | または，2 人以上 | 1644 |
| 辞書 8 | 人数× 3 点 | × 1 点：計 8 点以上 | 406 |
| 辞書 9 | 人数× 3 点 | × 1 点：計 5 点以上 | 744 |
| 辞書 10 | 人数× 3 点 | × 1 点：計 2 点以上 | 1115 |

表 2: 辞書の適合率，再現率，抽出正解文数

| | 適合率 | 再現率 | 抽出正解文数 |
|---|---|---|---|
| 辞書 1 | 0.71 | 0.38 | 33 |
| 辞書 2 | 0.56 | 0.48 | 42 |
| 辞書 3 | 0.56 | 0.48 | 42 |
| 辞書 4 | 0.52 | 0.51 | 45 |
| 辞書 5 | 0.45 | 0.51 | 45 |
| 辞書 6 | 0.57 | 0.51 | 45 |
| 辞書 7 | 0.49 | 0.52 | 46 |
| 辞書 8 | 0.62 | 0.41 | 36 |
| 辞書 9 | 0.50 | 0.41 | 36 |
| 辞書 10 | 0.61 | 0.53 | 47 |

抽出し，その適合率，再現率を求めた．抽出に用いた文は，電子掲示板からの 200 文と Twitter から 200 文の計 400 文とした．予備実験に使用した文は，失礼文の数が極端にならないよう，選択，収集をした．このとき，大学生 3 人に抽出に用いた文を読んでもらい，うち 2 人が失礼な文と判断した 88 文を正解文とした．

表 2 に，作成した 10 個の辞書の適合率，再現率，抽出正解文数を示す．表 2 から，抽出正解文数が多く，再現率が高い辞書は辞書 7 と辞書 10 であることがわかる．しかし，適合率をみると辞書 10 の方が優れている．よって，辞書 10 の単語を失礼単語辞書に用いることとした．表 3 に，この時点で収集できた失礼単語，名詞 942 単語と形容詞 173 単語，計 1115 単語からそれぞれ一部抜粋したものを示す．

表 3: 収集した失礼単語 (一部抜粋)

| 品詞 | 失礼単語 |
|---|---|
| 名詞 | 生意気　最低　邪魔　不細工　低能<br>音痴　非常識　無神経　外道　非道<br>ろくでなし　厄介　不評　ごみ　下手 |
| 形容詞 | つまらない　みっともない<br>あざとい　鈍い　陰気臭い　とろい<br>薄気味悪い　どぎつい　騒がしい |

### 3.4.2　失礼単語の追加

前節の辞書では再現率の値が十分でなく，その理由は，辞書の構築に用いた単語感情極性対応表がブログ記事を元に作られたものであり，電子掲示板やつぶやきによく含まれる単語でも，辞書内には含まれていない単語が見られたことによる．そこで，電子掲示板と Twitter のメッセージから，失礼表現を以下の手順によって追加した．

手順 1 電子掲示板と Twitter からメッセージの集合 (合計 1000 文[1] ) を収集し，辞書 10 を用いて失礼文を抽出．

手順 2 失礼文と判定されなかった文を大学生 3 人に読んでもらい，その中から失礼な文を選択してもらう．

こうして収集した 1000 文から，大学生 3 人に失礼単語となる可能性が最も高い単語を，1 文につき，1 単語取り出してもらった．結果として，各文において 3 人中 2 人以上が抽出した単語が存在したため，合計 1000 単語を追加の候補とした．その後，収集した 1000 文の中で 2%以上の文から回答された 37 単語を失礼単語として，辞書に追加した．表 4 に追加した単語の一部を示す．

最終的に作成された失礼単語辞書は，名詞 973 単語，形容詞 179 単語の合計 1152 単語から構成される．

[1]失礼単語の収集が目的のため，失礼文が多く存在しそうな文を収集した．

## 3.5　失礼単語辞書による失礼文の判定

前節で作成した失礼単語辞書を用いて，入力として与えられた文が失礼単語辞書内の単語を 1 単語以上含んでいる文を失礼文として判定する．

## 3.6　出力：失礼文の表示

システムの出力例を図 2 に示す．システムは，失礼文と判定した文を橙色で色づけして出力する．また，桃色で色づけされている単語が失礼単語となる．システムのユーザは，この出力を確認して，実際にメッセージとして投稿するか否かを判断することに役立てる．

表 4: 追加した失礼単語 (一部抜粋)

| 追加した失礼単語 |
|---|
| ウザい　キモい　デブ |
| ケチ　ショボい　ダサい　DQN |
| ニート　変態　ゆとり　無職 |

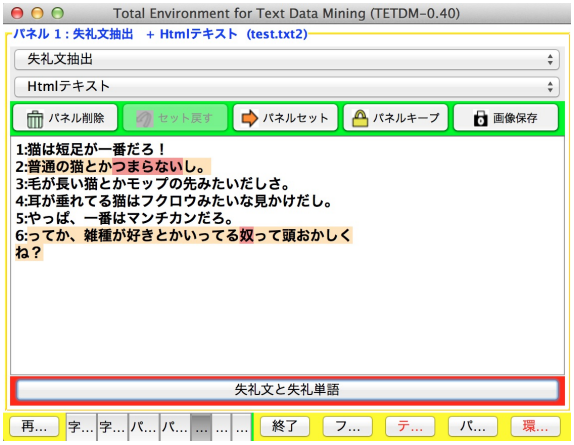


図 2: システムの出力例

## 4　システムの評価実験

本章では，システムがどの程度失礼文の抽出に効果があるかを検証した評価実験について述べる．

### 4.1　実験内容

提案するシステムによって，失礼文をどの程度抽出できるのかを明らかにするために，実験を行った．被験者は，情報科学を専攻する大学生・大学院生の計 20 名とした．表 5 に示す電子掲示板「2 ちゃんねる[2] 」のスレッドと，表 6 に示すニュース記事に対して，投稿された Twitter のツイートからそれぞれ 400 文，計 800 文を用いた．

手順 1　800 文全てを読んでもらう．この時，自分，誰か，もしくは何かに対し，失礼となる文を選択．(想定所要時間:40 分)

手順 2　システムが提示する失礼文抽出結果を参考に，手順 1 での回答を修正．(想定所要時間:20 分)

手順 2 では「手順 1 では選択したが，手順 2 では色付けされていなかった文」と「手順 1 では選択していなかったが，手順 2 では色付けされている文」の 2 パターンを修正の対象とした．また，被験者 20 名を 10 名ずつ，提案グループと比較グループの 2 グループに分けた．提案グループには，失礼単語辞書 (以下，提案辞書と記述する) を用いた失礼文抽出の結果を提示した．比較グループには，単語感情極性対応表の感情極性値が-0.48 以下となる 14035 単語を用いて作成した辞書 (以下，比較辞書と記述する) を用いた失礼文抽出の結果を提示した．

[2]http://www.2ch.net

表 5: 実験に使用した文を収集した電子掲示板のスレッド一覧

| |
|---|
| ★鉱物・石ヲタスレ★ Vol.35 |
| ネコたんと一緒の布団で添い寝したい PART11 |
| これだけは理解できないってある？ 2 |
| 100 円ショップの園芸モノってどうよ？ その 23 |
| 東京の食いもんはマズい！マズすぎる！166 |
| 【アイス総合】アイス大好きっ子スレ★その 97 |
| チェリオのブルースコーヒーが不味すぎて困る |
| 調理師として生きるということ part26 |

表 6: 実験に使用したツイートが投稿されているニュース記事

| |
|---|
| 世界でもっとも任天堂グッズを<br>集めた男がギネスに認定 |
| 【女性編】こんな人とカラオケに<br>行きたくない！ランキング |
| イチロー選手のバッド窃盗容疑<br>神戸の公園、19 歳逮捕 |
| 当てはまったら要注意！<br>「パソコン依存症」な人の特徴 5 つ |
| 任天堂「wii U」が救世主となれない 3 つの理由 |
| 「民間だったら当たり前」は<br>「民間のブラック企業なら当たり前」 |
| かつて『ハリー・ポッター』作者も<br>もらっていた「生活保護」 |
| ビンタ擁護論<br>「これで一斉に廃止したらどうなっちゃうのか」 |

### 4.2　実験結果と考察

表 7 に，提案辞書と比較辞書をそれぞれ用いた時のシステムが抽出した失礼文数と，各辞書の適合率，再現率を示す．正解文は，手順 2 の時点で，被験者 20 名中 10 名以上に失礼であると回答された 186 文とした．

表 7 より， 提案辞書，比較辞書それぞれを用いたシステムの再現率の間に差はないが，適合率では大きな差がみられた．理由として，提案辞書を用いたシステ

表 7: 各システムの適合率と再現率

| | 適合率 | 再現率 |
|---|---|---|
| 提案システム | 0.51(156/306) | 0.84(156/186) |
| 比較システム | 0.25(170/677) | 0.91(170/186) |

表 8: 両グループ被験者の適合率, 再現率

| | 適合率 | 再現率 |
|---|---|---|
| 提案:手順 1 | 0.75(101/135) | 0.54(101/186) |
| 手順 2 後 | 0.74(139/188) | 0.75(139/186) |
| 手順 2 のみ | 0.72(39/54) | — |
| 比較:手順 1 | 0.69(105/153) | 0.56(105/186) |
| 手順 2 後 | 0.61(142/234) | 0.76(142/186) |
| 手順 2 のみ | 0.46(37/80) | — |

ムの総抽出文数が比較辞書を用いたシステムの出力文数より少なかったことと, 提案辞書が少ない出力文数に対し, 失礼文を多く抽出できていたことの 2 つが挙げられる. 辞書の単語数によって, 提案辞書の出力文数が, 比較辞書の出力文数よりも少なくなったと思われる. これに加え, 出力中の正解失礼文の数をみると, 提案辞書と比較辞書との間に大きな差がみられない.

以上より, 提案辞書は比較辞書に比べ, 単語数は少ないが, 出力文数を抑えつつ, 総正解文数 186 文中から 156 文, 約 8 割の失礼文を抽出できたといえる. 従って, 提案する失礼単語辞書は, 失礼文の抽出に有効な単語を収集されており, 提案する失礼単語辞書による失礼文の抽出が, 一定の精度で行えているといえる.

表 8 に, 提案グループと比較グループの被験者の適合率, 再現率の平均を, 表 9 に提案グループと比較グループの修正時に被験者が追加した回答数とその中で正解だった回答数, また削除した回答数とその中で不正解だった回答数の平均を示す. 表 8 より, 提案, 比較ともに, 手順 2 後の再現率はほぼ同じ値となっているが, システムの出力を参考にした手順 2 における適合率には差があり, 比較グループの被験者は非常に多くの失礼ではないと考えられる文も, システムの影響を受けて抽出していたことがわかる. また, 表 9 より, 提案は比較と比べて, 修正時に適切な回答の追加と回答の削除を行えていることがわかる. このことから, 提案辞書を用いたシステムにより, 効率的に失礼文を抽出できることがわかった.

## 5 結論

電子掲示板や Twitter 等, 多人数に向けメッセージから, 失礼表現が含まれる可能性が高い文を抽出する

表 9: 両グループ被験者の回答追加, 回答削除数

| | 追加数 | 正解数 | 削除数 | 不正解数 |
|---|---|---|---|---|
| 提案 | 60 | 40 | 6 | 6 |
| 比較 | 84 | 38 | 4 | 3 |

システムを提案した. また, 失礼表現が含まれる可能性が高い文を抽出する方法として, 文に含まれる失礼表現となる可能性の高い単語に着目し, その単語を収集した失礼単語辞書を作成した. 作成した失礼単語辞書による, 失礼表現が含まれた文を抽出が, 有効なことを明らかにした.

## 参考文献

[1] 大坊郁夫: ネットワーク・コミュニケーションにおける対人関係の特徴, 対人社会心理学研究, No.2, pp. 1–14 (2002)

[2] 濱岡秀平, 砂山渡: 単語特性辞書を用いた電子掲示板の雰囲気の同定, 日本知能情報ファジィ学会誌, Vol.24, No.3, pp. 707–716 (2012)

[3] 一藤裕, 今野将, 曽根秀昭: 電子掲示板の雰囲気を考慮する発言分類, 電子情報通信学会技術研究報告, Vol.109, No.438, pp. 125–128 (2010)

[4] 松田有史, 今井倫太, 大澤博隆: Twitter タイムライン解析による存在感の抽出, 全国大会講演論文集 *2011*, No.1, pp. 169–171 (2011)

[5] 奥川巧, 大石哲也, 長谷川隆三, 藤田博, 越村三幸, 倉門浩二: Twitter のリスト機能を用いたユーザの特徴抽出, 全国大会講演論文集 *2011*, No.1, pp. 687–689 (2011)

[6] 古瀬蔵, 廣島伸章, 山田節夫, 片岡良治: ブログ記事からの意見文検索, 情報処理学会研究報告, 自然言語処理研究会報告 *2006*, No.124, pp. 121–128 (2006)

[7] 池田和史, 柳原正, 松本一則, 滝嶋康弘: 係り受け関係に基づく違法・有害情報の高精度検出方式の提案, 第 *2* 回データ工学と情報マネジメントに関するフォーラム *(DEIM2010)*, C9-5 (2010)

[8] 石坂達也, 山本和英: Web 上の誹謗中傷を表す文の自動検出, 言語処理学会第 *17* 回年次大会 発表論文集, pp. 131–134 (2011)

[9] 坂井俊之, 藤村孝: ブログに記述された不満表現からの潜在ニーズの発見, 情報処理学会論文誌, Vol.52, No.12, pp. 3806–3816 (2011)

[10] 松葉達明, 里見尚宏, 桝井文人, 井須尚紀: 学校非公式サイトにおける有害情報検出, 電子情報通信学会技術研究報告, *NLC,* 言語理解とコミュニケーション, Vol.109, No.142, pp. 93–98 (2009)

[11] 高村大也, 乾孝司, 奥村学: スピンモデルによる単語の感情極性抽出, 情報処理学会論文誌, Vol.47, No.2, pp. 627–637 (2006)

[12] 松本裕治, 山下達雄, 平野義隆, 松田寛, 高岡一馬, 浅原正幸: 形態素解析システム『茶筌』, Ver.2.4.0, 使用説明書 (2007)

# 組合せ発想のための意見交換の発散支援システム

## Opinion Divergence Support System for Combination Creation

宮原 和也　　砂山 渡
Kazuya Miyahara　　Wataru Sunayama

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:** As the general flow which performs exchange of opinions, there are an emission phase for which an idea is made to conceive broadly and a convergence phase which makes many ideas collect. In this study, we build the system which supports the broad way of thinking of an idea and its combination in the emission phase of exchange of opinions. That is, we display in a network the idea and its combination which the participant in exchange of opinions enumerated, and propose the system which presents and suggests the new idea and its combination which have not been enumerated yet. By the evaluation experiment, we checked that there was an e ect to which the contents which the system presented and suggested urge a participant's broad way of thinking.

## 1　はじめに

現代の社会生活において，意思決定をするために意見交換をする機会が多くなっている．しかし，複数人で意見交換を行う際に考える問題として，話が主題からそれてしまい関係のない意見のやり取りが増えてしまう点や，同じ意見を繰り返し発言してしまい，時間の浪費が多くなる点や，人間関係や立場上の問題により率直な意見が発言できなくなってしまう点がある．そこで意見交換において，スムーズかつ幅広いアイデアの列挙を支援する環境が望まれている．

一般的な意見交換の全体の流れとして,「発散フェイズ」と「収束フェイズ」の 2 つのフェイズが存在する．まず発散フェイズは，多数のアイデアを幅広く発想させることにより，多くのアイデア候補を検討していく．次に収束フェイズは，発散フェイズで挙げられた多数のアイデアをまとめあげることにより，複数の挙げられていたアイデア候補から 1 つの候補，いわゆる結論に絞り込んでいく．

ここで，先行研究として「RFID タグを用いた意見交換の収束支援システム」[1] を挙げる．この研究では，意見交換における収束フェイズに着目し，列挙された選択肢を絞り込む過程を，RFID タグを用いた意見交換環境により，スムーズな意見交換の進行と，多くの参加者が納得できる結論の決定を支援した．よって，意見交換の一連の流れのうち「収束フェイズ」を支援する研究は上記の通り完成しているため，今回は，残りの「発散フェイズ」を支援する研究を行う．

そこで本研究の目的は，意見交換の発散フェイズにおいて，アイデアとその組合せの幅広い発想を支援するシステムの構築とする．具体的には，参加者が列挙したアイデアとその組合せをネットワークで表示し，未列挙の新しいアイデアとその組合せを提案する．

以下，2 で関連研究，3 で発散支援システム，4 で発散支援システム評価実験，5 で結論を述べる．

連絡先： 広島市立大学大学院情報科学研究科
〒 731-3194 広島市安佐南区大塚東 3-4-1

## 2　関連研究

### 2.1　意見交換の発散支援に関する研究

膨大なアイデアを発想させる研究として,「ブレインストーミング法習得のためのカードゲーム開発とストレス軽減およびルール学習効果の検討」[2] が挙げられる．本研究との相違点として，この研究は参加者同士だけでのアイデアの発散を促す方法だが，本研究は，システム側からアイデアに関する関連単語を提示したことで，更なるアイデアの発散を促せる点で異なる．

アイデアの発散を支援する研究として,「GroupMind: Supporting Idea Generation through a Colllaborative Mind-mapping Tool」[3] が挙げられる．本研究との相違点として，この研究は一から候補を考えるブレインストーミングの手法をとっているが，本研究では既に候補の発想を限定できる点になる．

キーワードマップを利用して新たな意思決定を促す研究として,「関連バランス制御機能を組み込んだキーワードマップによる意思決定方略に応じたデータ分析

の支援」[4] が挙げられる．本研究との相違点として，この研究はマップ表示により自由に発想していくのだが，本研究ではマップ表示を行い関連度の高い候補を提示することで，適切な発想を支援できる点で異なる．

キーワードマップを利用して新たな意思決定を促す研究として,「Poker-Maker モデル：ユーザの検索意図を反映するキーワードマップと情報収集エージェントの連携による探索的情報検索」[5] が挙げられる．本研究との相違点として，この研究は参加者の検索意図に応じて情報収集や可視化を行うことで支援するが，本研究では参加者の検索意図は考慮せず，参加者が列挙したアイデアに関連するキーワードを用いて支援する点で異なる．

### 2.2 関連情報による発想支援に関する研究

オンライン上の情報を利用し意見交換を支援する研究として,「オンラインゲームにおけるコミュニケーション支援のための Web を用いた情報抽出」[6] が挙げられる．本研究との相違点として，この研究は関連語の抽出元がチャットログや Web 内で未知語となる単語を対象とするのだが，本研究では関連語の抽出元が検索エンジンでのヒットページになり頻度の多い単語を対象とする点で異なる．

関連情報の提示による意見交換を支援する研究として,「MAKOTO：ソーシャルグラフを用いたコミュニケーション支援システムの提案」[7] が挙げられる．本研究との相違点として，この研究はお互いの共通する情報を可視化して参加者自身の情報を提示し支援しているが，本研究では参加者が挙げたアイデアと関連する情報を可視化して支援する点で異なる．

列挙された単語同士の関連情報を用いて意見交換を支援する研究として,「会話中の名詞の関連情報を用いた対面型異文化間コミュニケーション支援システムの構築と評価」[8] が挙げられる．本研究との相違点として，この研究は，対話内で抽出した名詞画像，関連名詞とその画像を提示して支援するが，本研究では列挙された名詞に対する関連単語と関連を示すネットワークを提示して支援する点で異なる．

### 2.3 アイデアの発想支援に関する研究

アイデアの組合せ発想を支援する環境に関する研究として,「組み合わせ発想を刺激するイノベーションゲーム」[9] が挙げられる．本研究との相違点として，この研究は，異なるテーマのアイデア同士を組合わすイノベーションゲームを構築して支援するが，本研究は，同じテーマで列挙されてない組合せから新たな組合せを示唆できる点で異なる．

多様な情報を利用してアイデア発想を促進させる研究として,「情報の多様性がアイデア生成に及ぼす影響の検討」[10] が挙げられる．本研究との相違点として，この研究は，関連が低い情報でも利用する支援をしたが，本研究では，関連の高い情報を利用して発想を支援できる点で異なる．

創造的なアイデアの組合せ発想を支援する研究として,「Generating Creative Ideas Through Crowds: An Experimental Study of Combination」[11] が挙げられる．本研究との相違点として，この研究はアイデアを挙げた順に一つずつ組合わせることで新しいアイデアを生成するが，本研究では挙がったアイデアをランダムに組合わせる事ができる点で異なる．

## 3 発散支援システム

本章では，本研究で扱う発散支援システムについて述べる．

本研究では，意見交換の発散フェイズにおいて，参加者が列挙したアイデアをネットワークで表示し，参加者が列挙していない新たなアイデア候補を提示するインタフェースを構築する．また，意見交換においてはアイデアを発言する参加者と，参加者の行動によりシステムを操作し，参加者への指示を行う司会者がいる．

今回対象とする意見交換は，上記に述べた発散フェイズの流れにより，一般的に話し合いの場が設けられ，かつ組合せを考える必要があるテーマとする．また「発散フェイズ」は 2 つに分かれ，前半では，参加者は主題に関連するアイデアを，単独で幅広く列挙してもらう．後半では，前半部分で挙げられたアイデアから，幅広く組合せを列挙してもらう．

### 3.1 発散支援システムの環境

本節では，発散支援システム環境について述べる．

本研究では，意見交換の各参加者に RFID タグ付きカードを選択肢カード 8 枚，発言終了カード 1 枚で，及びカード提示用のボードが割り当てられる．RFID タグ付きカードを利用することで，対面での意見交換を阻害しない媒体として，他のデバイスの利用による使用，入力方法への意識を不要にするとともに，お互いのカード提示行動を確認できる環境で，スムーズな発言順序の構築を支援できる．

参加者とカードの情報の取得方法は，図 1 を用いて説明する．エリア（Area）内に各参加者のボードが置いてあり，そのボード上に置かれたカードの情報が PC 上に送られ，スクリーン上のインタフェース上に表示される．ここで，ハードウェアの制約上，参加できる人数は 4 から 8 人とする．

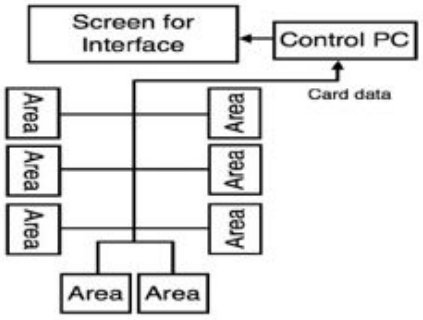


図 1: 発散支援システムの環境


図 2: 発散支援システムのインタフェース

## 3.2 発散支援システムのインタフェース

本節では，発散支援システムのインタフェースについて述べる．本システムのインタフェースを，図 2 に示す．

①の部分は，テーマと参加者登録時の残り時間を示す．②の部分は「情報視覚化パネル」で，列挙された意見の情報を見ることができる．③の部分は「ユーザ情報パネル」で，現時点での参加者の状況を確認できる．このエリアは 4 つに分かれており，現時点で参加者が「(左から) 発言エリア／待ち行列／待機エリア／終了エリア」の状況のいずれに属するかを表している．またアバター画像の上に対応する各参加者の番号も表示している．各エリアでの参加者の状況を表 1 に示す．

## 3.3 発散フェイズ前半

本項では，発散支援フェイズの前半のシステム構成について述べる．本システムの発散フェイズ前半の構成を図 3 に示す．以下，参加者と司会者が行うこととあわせて説明する．

表 1: 参加者の状況と表示エリア

| エリア名 | 現時点での参加者の状況 |
|---|---|
| 発言エリア | 意見を発言中の人 |
| 待ち行列 | 意見を発言する意思があり待っている人 |
| 待機エリア | 特に何もしていない人 |
| 終了エリア | 意見を発言する意思のない人 |

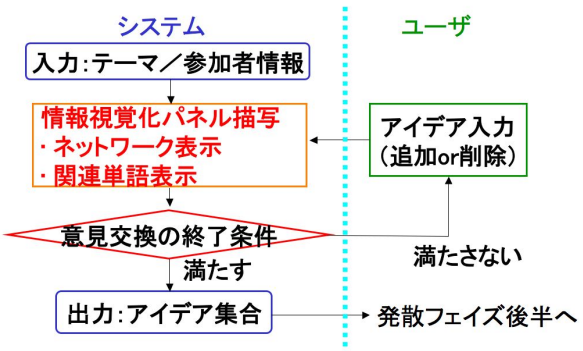


図 3: 発散フェイズ前半のフローチャート

### 3.3.1 入力：テーマ／参加者情報

入力は，意見交換のテーマと参加者情報をシステムに与える．参加者情報は，カードを参加者が 1 枚ボード上に提示することで取得できる．「待機エリア」に参加者番号に対応したアバターが表示される．ここで参加できる参加者の数はハードウェアの制約により 4 人から 8 人とし，全員が参加できたら意見交換を開始する．

### 3.3.2 アイデア入力

意見を発言したい参加者は，カードを 1 枚ボード上に提示すると「待ち行列」に，現在発言中の人がいなくなると「発言エリア」に自分のアバターが表示される．表示された参加者は，自分の意見として，必ず 1 回単語で 1 個（＝アイデア）列挙しその理由を 30 秒以内で述べる．ここで司会者は，列挙されたアイデアの単語を 1 つずつ入力していく．終了するまでは，上限 8 個までアイデアを追加したり削除したりできる（追加 or 削除自体は司会者が行う）．

### 3.3.3 情報視覚化パネル描写

システム側がアイデア同士の関連を表すネットワークと参加者が挙げたアイデアと関連単語をその上に提示することで，新たなアイデアを参加者に発想させる．ここでの描写方法の詳細は 3.5 節で，特に描写例は 3.5.3 項で述べる．

#### 3.3.4 意見交換の終了条件

参加者全員が 1 度以上発言しかつ発言終了カードを提示した場合に，前半が終了する．発言する意思のない参加者は発言終了カードを提示すると,「終了エリア」に参加者番号と発言終了カードが表示される．

#### 3.3.5 出力：アイデア集合

司会者は前半終了の旨を参加者に伝え,「要素列挙 ⇔ 組合せ」ボタンを押し発散フェイズ後半へと移る．ここで，それまでに列挙されたアイデア集合を出力する．

### 3.4 発散フェイズ後半

本項では，発散支援フェイズ後半のシステム構成について述べる．本システムの発散フェイズ後半の構成を図 4 に示す．

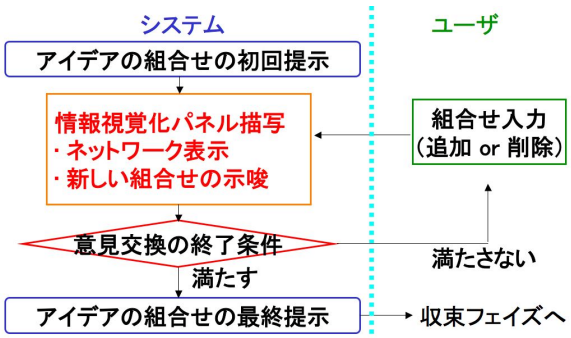


図 4: 発散フェイズ後半のフローチャート

#### 3.4.1 アイデアの組合せの初回提示

発散フェイズ後半開始時に，前半で列挙されたアイデアのリストを参考に，参加者は組合せを列挙する．方法は，参加者は組合わせたい符号のついた選択肢カードを全てボード上に提示することで，アイデアの組合せのシステムへの入力ができる．

#### 3.4.2 情報視覚化パネル描写

システム側から参加者が列挙した組合せを表すためのネットワークを表示する．各参加者はここで組合せを選んだ理由を 30 秒以内で述べる．発言方法は，前半の 2 と一緒とする．さらに同時にシステム側から，新たな組合せ候補を示唆しその組合せに対する関連単語を提示することで，新たな組合せを参加者に発想させる．ここでの描写方法の詳細は 3.6 節で，特に描写例は 3.6.3 項で述べる．

#### 3.4.3 組合せ入力（追加 or 削除）

新たな組合せを追加したり列挙した組合せを削除できる（追加 or 削除自体は司会者が行う）．

#### 3.4.4 意見交換の終了条件

終了条件は，前半の 4 と一緒とする．

#### 3.4.5 アイデアの組合せの最終提示

参加者はもう一度現時点での最終的な意見としての組合せを提示する．提示方法は，後半の 1 と一緒とする．ここで，意見交換を終了する．

### 3.5 発散フェイズ前半での情報視覚化パネル描写

本節では，発散フェイズ前半における情報視覚化パネルの描写について述べる．ここでは，アイデア間の関連を表すネットワークを表示しアイデアに対する関連単語を提示する．

#### 3.5.1 ネットワーク表示方法

本項では，発散フェイズ前半におけるネットワークの表示方法について述べる．ネットワークを表示する目的として，各アイデアの全体の中での位置づけを確認し，幅広くアイデアを考える参考にさせる．

ネットワークとして，各アイデアとアイデアの関連を表示する．アイデアが列挙されると，各アイデアの座標をバネモデルにより計算し描写を行う．その後，アイデア同士の関連度を計算し，列挙された全アイデア同士の関連度の平均値を上回った関連度のアイデア同士のみ線を引き，さらに関連度の高いアイデア同士は距離を縮めて線を太く表示する．

ここで，関連度の計算方法を記述する．列挙されたアイデアをキーワードとして，検索エンジン（Yahoo! JAPAN）からヒット件数を取得し，以下の式 1 で，アイデア A と B の関連度 $relate(A,B)$ を求める．

$$relate(A,B) = \frac{P(I_A \cap I_B \cap T)}{\sqrt{P(I_A \cap T) \quad P(I_B \cap T)}} \qquad (1)$$

ただし上式 1 で使用した記号については以下の通り．

- $P(A)$：$A$ の検索ヒット件数
- $P(A \cap B)$：$A$ と $B$ の $AND$ 検索ヒット件数
- $I_A$：アイデア $A$ の単語
- $T$：テーマに含まれる全名詞

さらに，テーマに含まれる全名詞数を $n$ とすると，

$T$={$T_1$∩..∩$T_n$}（$T_n$：テーマに含まれる $n$ 番目の単語の検索ヒット件数）で計算し，名詞は，形態素解析による日本語自然言語処理システム *chasen* により取得する．

### 3.5.2 関連単語について

本項では，発散フェイズ前半で提示する関連単語について述べる．

関連単語を提示する目的は，あるアイデアから新たなアイデアを連想させるためとする．関連単語を取得するアルゴリズムは，以下の通りになる．また，毎回の検索で同様にヒットする「ページ，検索，結果，サーチ，情報，設定，タグ，ログイン，スポンサー」は表示する際に除去する．

1. Yahoo JAPAN を利用して，テーマに含まれる単語と列挙されたアイデアとの *AND* ヒット検索を行う．
2. ヒットした 1 ページ目（検索結果数：10 件）に含まれている表示される文章（スニペット）を抜き出す．
3. 抜き出した文章に *chasen* を用いて形態素解析を行い，名詞と判断された単語を出現回数とともに取得する．
4. 出現回数の多い名詞から順番に，インタフェース上に水色の枠を作りその中に表示する．

### 3.5.3 情報視覚化パネル描写例（インタフェース例）

発散フェイズ前半での情報視覚化パネル描写例（インタフェース例）を図 5 に示す．テーマが表示され，ユーザ情報パネル上に参加者の情報が表示される．アイデアが挙げられていくと，アイデア同士の関連を示すネットワークを表示させ，各アイデアの関連単語を水色の枠内に 20 個表示している．「ドイツ-フランス」「イタリア-スイス」は関連が強い．また関連単語について，アイデア「イタリア」を例にとって考えると，「スペイン」が新たなアイデアとして連想できる．

## 3.6 発散フェイズ後半での情報視覚化パネル描写

本節では，発散フェイズ後半における情報視覚化パネルの描写について述べる．ここでは，アイデアの組合せを表すネットワークと組合せの関連単語を表示する．

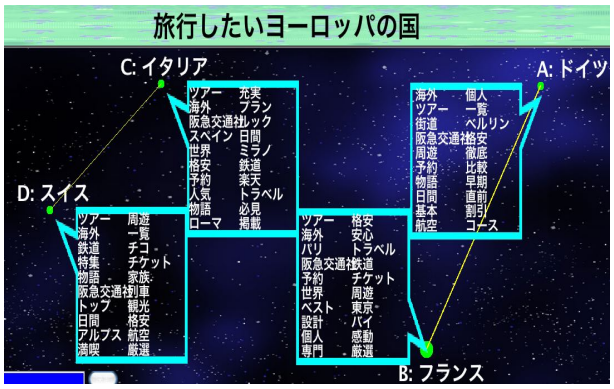


図 5: 発散フェイズ前半のインタフェース例

### 3.6.1 ネットワーク表示方法

本項では，発散フェイズ後半における情報視覚化パネルの描写方法について述べる．ネットワークを表示する目的として，同じ組合せを列挙した参加者や列挙されていない組合せを確認させる．

表示するネットワークは，前半で描写された各アイデアをそのまま利用し，参加者が列挙した組合せを参加者の色で全て実線でアイデア同士に引く．

実線で参加者全員の組合せがネットワーク上に表示された後，システム側が新しい組合せ候補として 2 種類の組合せを示唆する．これは，今列挙されていない組合せの中で一般的な組合せを赤点線，面白い組合せを青点線でアイデア同士に引く．赤点線は，参加者が忘れがちな組合せを気づかせるため，最も関連度（3.4.1 項の式 (1) 参照）が高いアイデア同士に引く線とする．青点線は，参加者の視野を広げてくれそうな面白い組合せを気づかせるため，テーマの単語を含めない関連度（式 1 から $T$ を省いた式）とテーマの単語を含めた関連度（式 1）を比較し，後者が高くかつ関連度が高いアイデア同士の組合せを選択し線を引く．

### 3.6.2 関連単語について

本項では，発散フェイズ後半で提示する関連単語について述べる．関連単語を提示する目的は，幅広く組合せを考えるきっかけを促すためとする．

関連単語を抽出するアルゴリズムは，発散フェイズ前半で用いたもののステップ 1 を下記で置き換えたものを用いる．

1. Yahoo JAPAN を利用してテーマに含まれる単語と点線を引いたアイデアとの *AND* ヒット検索を行う．

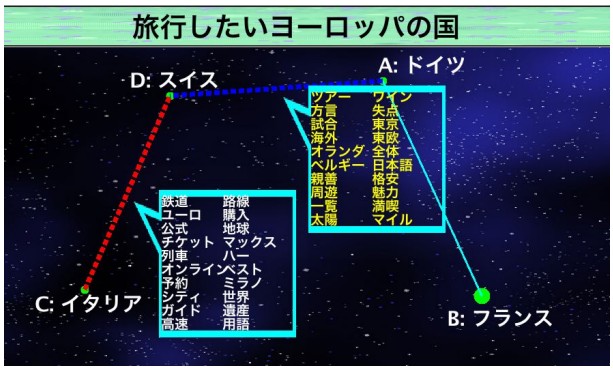


図 6: 発散フェイズ後半のインタフェース例

ただし，単語は全て名詞のみ扱い，省く単語は発散フェイズ前半と同様にし，赤点線においては白文字で，青点線においては黄文字で表示している．

### 3.6.3 情報視覚化パネル描写例（インタフェース例）

発散フェイズ後半での情報視覚化パネル描写例（インタフェース例）を以下図 6 に示す．参加者が列挙したアイデアの組合せをネットワーク上に実線で表示する．列挙されていない組合せは「ドイツ-フランス」以外の 5 通りある．その中から最も関連度が高い「イタリア-スイス」に赤点線（一般的な組合せ）が引かれ，テーマ含む関連度がテーマ含めない関連度を超える場合において関連度の高い「ドイツ-スイス」に青点線（面白い組合せ）が引かれる．

また関連単語について，赤点線「イタリア-スイス」を例にとって考えると，提示された中の「ユーロ」「ミラノ」からサッカー関連で結ばれた可能性が高いと分かり，この組合せを考えるきっかけを与えられる．

## 4 発散支援システムの評価実験

本章では発散支援システム評価実験について述べる．実験目的として，発散フェイズにおいてシステムを利用し意見交換を行う際に，幅広くアイデアまたはその組合せが列挙されたか，そしてシステムが提示する内容が参考になったかに着目して評価を行う．

### 4.1 実験準備

テーマに対して解答の候補を幅広く挙げる意見交換を，提案システムと比較システムそれぞれ被験者 5 人と司会者 1 人で行った．意見交換は全 6 テーマ（表 2 参照）行った．被験者は大学生，大学院生 20 名とし，各テーマで参加した被験者の組合せはテーマにより適

表 2: 実験に使用した意見交換のテーマ

| テーマ | テーマ内容 |
|---|---|
| 1 | 鍋パーティをすることになりました．<br>食材 3 つまでなら，先生が買って来てくれるそうです．<br>何を買って来てもらうかを相談して下さい． |
| 2 | ヨーロッパに一週間（移動日を除くと 5 日間）<br>の卒業旅行に行くことになりました．<br>まわる国の組合せを相談して下さい． |
| 3 | 年度末に研究室の大掃除をすることになりました．<br>掃除機とぞうきん掛け以外に，<br>どんな掃除をすればよいでしょうか？ |
| 4 | 母校の高校生が，大学に見学に来ることになりました．<br>1 時間引率して，大学の施設をいくつか見てもらうときに，<br>どこを見せてまわりますか？ |
| 5 | 無人島に 3 つだけものをもっていけるとした場合，<br>何をもっていきますか？ |
| 6 | 結婚相手に何を求めますか？ |

表 3: 列挙されたアイデア一覧（提案システム）

| テーマ 1 | テーマ 2 | テーマ 3 | テーマ 4 | テーマ 5 | テーマ 6 |
|---|---|---|---|---|---|
| 豚肉 | イギリス | ハタキ | 学食 | ナイフ | 家事 |
| 白菜 | ドイツ | ホウキ | 図書館 | 飲料水 | 一般常識 |
| 豆腐 | フランス | 窓拭き | 喫茶店 | ろ過装置 | 共通趣味 |
| 牡蠣 | スペイン | シンク | パソコン室 | ライター | 計画性 |
| 鶏肉 | ギリシャ | フィルター | 講義棟 | 火打石 | 容姿 |
| 大根 | イタリア | 壁拭き | 芸術棟 | 釣竿 | 子育て |
| 水菜 | ポルトガル | 激落ちくん | 体育館 | ロープ | 共通感性 |
| ネギ | スイス | - | 売店 | 望遠鏡 | 一定収入 |

切に割り振った．実験手順は，3.3 節で記述した内容と一緒の方法で行ったため，ここでは省略する．

提案システムは，前半において，アイデア同士のネットワークとアイデアに対する関連単語を表示し，後半においては，組合せを表示したネットワークと，システム側からのアイデアの新たな組合せの示唆と対応する関連単語を提示したものとする．

比較システムは，前半において，アイデア同士のネットワークを表示し，後半においては，組合せを表示したネットワークを提示したものとする．

### 4.2 実験結果と考察

発散フェイズ前半において列挙されたアイデア候補の一覧を，表 3，表 4 に示す．テーマ 4, 6 については，提案では最大の 8 個挙げられていたが，比較では 6 個と 7 個だった．一方，テーマ 3 については，比較では最大の 8 個挙げられていたが，提案では 7 個だった．だが，5 人が被験者だったため最低 5 個以上列挙されれば良いと指示を出していたので，新たなアイデア候補が思いつかない場合にはすぐ発言終了の意思を出したと考えられる．

提案システムと比較システムの両システムにおいて，

表 4: 列挙されたアイデア一覧（比較システム）

| テーマ 1 | テーマ 2 | テーマ 3 | テーマ 4 | テーマ 5 | テーマ 6 |
|---|---|---|---|---|---|
| 豚肉 | イタリア | ミニモップ | グラウンド | ライター | お金 |
| うどん | ルーマニア | ゴミ | エネルギーセンター | 包丁 | 知性 |
| 豆腐 | ドイツ | 窓拭き | 学生会館 | ひも | 性格 |
| 白菜 | イギリス | ロッカー | 実験室 | ろ過装置 | 容姿 |
| スープ | スイス | 机拭き | 芸術学部 | ビニールシート | 相性 |
| 白ネギ | フィンランド | コロコロ | 公園 | 釣竿 | 貞操観 |
| 水菜 | オランダ | 業者 | - | 銛 | 愛 |
| しいたけ | スペイン | フィルター | - | 鍋 | - |

表 5: 全アイデア同士の関連度の平均

| テーマ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 提案システム | 0.171 | 0.653 | 0.062 | 0.106 | 0.139 | 0.109 |
| 比較システム | 0.196 | 0.665 | 0.114 | 0.092 | 0.162 | 0.307 |

各テーマに対する全アイデア同士の関連度の平均を比較したものを表 5 に示す．

テーマ 4 以外，全アイデア同士の関連度の平均値が，提案システムの方が低かったことが分かる．よって似たアイデアが少なく様々な種類のアイデアが列挙されたと考える．特にテーマ 2 に関しては，関連単語として表示した国名が候補として列挙されたアイデアが 8 個中 4 個（フランス／ドイツ／スイス／ポルトガル）ヒットした．残りのテーマに関しても，システムが提示した関連単語は，アイデア候補を挙げる際に役に立った．

一方，テーマ 4 に関しては，関連度の平均値が提案システムの方が高かったため，提示した関連単語の表示が役に立たなかったと考える．原因として，テーマ 4 が「広島市立大学」と範囲をかなり狭くしたことにより，提示する関連単語が上手に拾えなかった．よって，単語として拾えるものを増やすため，検索エンジンから取得できる単語の視野を広くすればよいと考える．

発散フェイズ後半について，被験者が最終的に列挙された組合せにおいて，1 人だけ挙げた組合せ（以下，単独リンク）を比較したものをそれぞれ表 6 に示す．表中の「追加」は，初回提示で列挙された組合せにはなく最終提示で列挙された組合せに現れたものを表す．

テーマ 3 以外，単独リンクにおいて提案システムのみ新しい組合せが追加された．つまり比較システムでは参考する情報が無かったので，幅広く組合せを列挙できなかったと考える．逆に提案システムでは新しい組合せの示唆と関連単語の提示により，新たな組合せの発想を促せたと考え，単独リンクが増えたことから幅広く組合せが列挙されたことが分かる．

一方，テーマ 3 に関しては，比較システムでも単独リンクが 2 本追加されたため，ネットワークを見せたことが新たな組合せを促せたと考える．さらに，比較システムの方が単独リンクの総数が多いため幅広い組合せが列挙されたと考える．以下，テーマ 3 において，比較システムの方が幅広く組合せを列挙された理由を，共通で列挙された組合せの視点から考える．

表 6: 列挙された単独リンク数

| テーマ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 提案（合計） | 9 | 5 | 3 | 7 | 8 | 11 |
| 提案（追加） | 2 | 4 | 2 | 4 | 3 | 1 |
| 比較（合計） | 7 | 1 | 5 | 4 | 0 | 2 |
| 比較（追加） | 0 | 0 | 2 | 0 | 0 | 0 |

表 7: 列挙された共通リンク数

| テーマ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 提案（合計） | 3 | 4 | 5 | 2 | 3 | 9 |
| 提案（追加） | 0 | 0 | 2 | 0 | 0 | 0 |
| 比較（合計） | 3 | 4 | 1 | 6 | 3 | 13 |
| 比較（追加） | 0 | 0 | 0 | 0 | 0 | 0 |

被験者が最終的に列挙された組合せにおいて，共通で列挙された組合せ（以下，共通リンク）を比較したものを表 7 に示す．表中の「追加」は，初回提示で列挙された組合せにはなく最終提示で列挙された組合せに現れたものを表している．

テーマ 3, 4, 6 以外，最終提示時の総数は両システムとも変わらず，さらにテーマ 3 以外は新しい組合せも無いため共通リンクに差はなかったと判断できる．

一方，テーマ 3 に関して，最終提示時の総数は提案システムの方が多かった．よって，提案システムの方では他人と同じ組合せを選んだ被験者が少なかったと考えられる．つまりテーマ 3 では単独リンクとして上記に述べた通り，比較システムの方が幅広く組合せが列挙され，提案システムの方は，皆同じ発想で組合せを列挙したと考えられ幅広く組合せを列挙できなかった．

また最終提示の総数が変わったテーマ 4, 6 に関して，最終提示時の総数は比較システムの方が多かったが新たな組合せの追加はなかった．よって，提案システムの方では他人と同じ組合せを選んだ被験者が少なかったと考えられ，他のテーマ以上に幅広く組み合わせ候補が挙げられたと考えられる．

最後に，システム側が提案した組合せ（以下，提案リンク）を表 8 に示す．表内の青文字は，実際に被験者が採用した組合せを示している．

赤点線は，テーマによらず一般的な組合せを表示するため，テーマに沿っていない場合や，意図的に外している場合には採用されない可能性があり，実際にはテーマ 3 で一つが採用されるのみとなった．青点線は，

表 8: システムが提案したリンク

| | テーマ 1 | テーマ 2 | テーマ 3 |
|---|---|---|---|
| 赤点線 | 豆腐 - 大根 | ドイツ - ギリシャ | シンク - フィルター<br>ホウキ - 窓拭き |
| 青点線 | 白菜 - ネギ<br>豆腐 - 鶏肉<br>鶏肉 - 大根<br>鶏肉 - ネギ | イギリス - ギリシャ<br>スペイン - ギリシャ<br>スペイン - ポルトガル | ホウキ - 窓拭き<br>ホウキ - シンク<br>ホウキ - フィルター |
| | テーマ 4 | テーマ 5 | テーマ 6 |
| 赤点線 | 図書館 - 体育館 | ロープ - 望遠鏡 | 共通感性 - 一定収入 |
| 青点線 | - | ナイフ - 望遠鏡 | - |

一般的には組み合わせられる可能性が低く，テーマに関連する際に組合せが強くなるもので，テーマに関する見落としを防ぐ意味がある．実際には 4 つのテーマでシステムからの提案が出力され，うち 3 つのテーマでその打ちの 1 つのリンクが採用される結果となった．

これらのリンクは，実際にリンクとして採用されるものを提示すること以外にも，膨大な組合せの中から，客観的な指標に基づいて有効な可能性が高い組合せについて，採用されたものと採用されなかったものが存在する結果となったことから，被験者による提案リンクの吟味が行われ，見落としがないことの確認を促す効果もあったと考えられる．これらの結果から，システムによるリンクの提案にも一定の効果があったと考えられる．

## 5 おわりに

意見交換をする際に，開発した発散支援システムにより，参加者に役に立つアイデアを幅広く発想させることが可能だと検証した．まず発散フェイズ前半では，アイデア間の関連を示すネットワークを表示し，各アイデアに対する関連単語を提示したことにより，自分の発想にない新しいアイデア候補を幅広く発想させる事ができた．次に発散フェイズ後半では，参加者が列挙した組合せをネットワークで表示させ，新たな組合せ候補を示唆し，かつその組合せに対する関連単語を表示させる事により，自分の発想にない新しい組合せを幅広く発想させる事ができた．

そこで今後は，考察で挙げたテーマの設定方法と示唆する組合せの閾値といった検討すべき点を考慮し，多種多様なテーマに関して意見交換をスムーズに行えてかつ幅広くアイデアを発想させられるシステム環境を作り出すことを課題として挙げる．さらに，先行研究 [1] と合わせて 1 つのシステムとすることにより，意見交換全体を支援する．

## 参考文献

[1] 砂山渡, 清水允文 : RFID タグを用いた意見交換の収束支援システム, 人工知能学会論文誌, Vol.26, No.5, pp.527 – 535 (2011)

[2] 西浦和樹, 田山淳: ブレインストーミング法習得のためのカードゲーム開発とストレス軽減およびルール学習効果の検討, 日本教育工学会論文誌, Vol.33, pp.177 – 180 (2009)

[3] Patrick C. Shih, David H. Nguyen, Sen H. Hirano, David F. Redmiles, Gillian R. Hayes: GroupMind: Supporting Idea Generation through a Colllaborative Mind-mapping Tool, *International Conference on Supporting group work Pages*, pp.139 – 148, (2011)

[4] 梶並知記, 槇原崇, 小笠原敏之, 高間康史: 関連バランス制御機能を組み込んだキーワードマップによる意思決定方略に応じたデータ分析の支援 , 知能と情報, Vol.21, No.6. pp.1067 – 1077 (2009)

[5] 梶並知記, 高間康史: Poker-Maker モデル：ユーザの検索意図を反映するキーワードマップと情報収集エージェントの連携による探索的情報検索, 情報知識学会誌, Vol.20, No.3, pp.277 – 292 (2010)

[6] 高松雅彦, 荒木健治: オンラインゲームにおけるコミュニケーション支援のための Web を用いた情報抽出, 情報処理学会研究報告, Vol.2010, No.9, pp.1 – 7 (2010)

[7] 藤本義治, 星亮輔, 高宮浩平, 井口真朝, 岡本誠, 松原仁: MAKOTO：ソーシャルグラフを用いたコミュニケーション支援システムの提案 , 情報処理学会シンポジウム論文集, Vol.2011, No.3. pp.703 – 706 (2011)

[8] 岡本健吾, 吉野孝: 会話中の名詞の関連情報を用いた対面型異文化間コミュニケーション支援システムの構築と評価, 情報処理学会論文誌, Vol.52, No.3, pp.1213 – 1223 (2011)

[9] 高市暁広, 大澤幸生, 古田一雄, 定木淳, 青山和浩: 組み合わせ発想を刺激するイノベーションゲーム , 人工知能学会全国大会論文集, Vol.22, 1B2-8 (2008)

[10] 清河幸子, 鷲田祐一, 植田一博, Eileen Peng: 情報の多様性がアイデア生成に及ぼす影響の検討 , 認知科学, Vol.17, No.3. pp.635 – 649 (2010)

[11] Lixiu Lisa Yu, Jeffrey V. Nickerson: Generating Creative Ideas Through Crowds: An Experimental Study of Combination, *Thirty Second International Conference on Information Systems*, pp.1 – 16 (2011)

# 言語表現による時系列データ検索システムの提案

## A System for Retrieving Time-Series Data Based on Linguistic Expression

蓮井大樹 [1*]　　松下光範 [2]
Daiki Hasui[1]　　Mitsunori Matsushita[2]

[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 for retrieving time-series data based on a linguistic query given by a user. Our proposed system uses a line chart as a query. The system generates a linguistic query by verbalizing the chart first, then retrieves similar charts by using the obtained linguistic query.

## 1　はじめに

現在、インターネットを介してさまざまな時系列データや統計データを得ることが出来るようになってきた。しかしグラフの検索を行う際に、どのようなグラフを求めているのかを言語化し、それに適したグラフを得ることは難しい。

例えば、「全体的に凸型のグラフ」のように、頭の中には具体的なグラフの形が浮かんでいるが、それを適切に言語化することができないために曖昧な表現しかできない場合には、システムとユーザがインタラクションを繰り返しながら徐々にユーザ要求を明確化し、探索的に適切なグラフを見つける必要がある。また、「昨年 8 月頃に特徴的な変化をした株は?」「他の商品にくらべて価格変動が緩やかな商品は?」といったように、他と比較した値の変化の特徴を利用して探索的に条件を満たす対象を見つけるような場合では、ある対象が全体集合の中でどのような位置づけにあるかを理解し、それを考慮してユーザの意図や関心に沿った区間や粒度を特定する必要がある。

本研究のゴールは、様々な時系列データに対する柔軟な情報アクセス手段の提供である。そのために、ユーザが与えた検索要求から、その条件に見合った変動をしている時系列データを特定したり、特定の時系列データから該当する時期を見つけたりするための検索機構の実現を目指している。

現在そのひとつのアプローチとして、時系列情報に予め自然言語表現を付与しておき、それとユーザの検索要求とのマッチングによって適切な範囲・粒度の時系列情報を特定し、視覚化する手法について検討を進めている [6]。この手法では、(1) 時系列データに基づく言語表現の生成、(2) 自然言語で表現された質問の解釈、(3) これらふたつのマッチング方法の定式化、という 3 段階の枠組を想定している。本稿ではこのうちの (1) と (3) に焦点をあて、時系列データを対象に、あらかじめ用意したグラフから検索したいグラフと類似している形状や傾向の部分を指定することでグラフ自体をクエリとし、そのグラフを言語化することで検索を行うシステムを提案する。

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

## 2　関連研究

時系列データなどの数値情報の集合を直感的に理解・把握するために、自然言語を用いて表現する手法について、これまで様々な研究が行われている。

例えば、グラフの解析に関する研究として、Ahmad ら [1] や小林ら [5] の研究が挙げられる。Ahmad らは Wavelet 解析を行って変極点や変動サイクルなどのグラフ特徴を抽出し、それらを元にテキストを生成する手法を提案している [1]。小林らは SAX 法を用いて記号化し、複数のグラフを比較することで時系列データ間の関連性を捉え、その結果を言語化する手法を提案している [5]。

また、時系列データからのテキスト生成に関する研究として Kukich [3] や 小林ら [4] の研究などが挙げられる。Kukich は入力として与えられた数値情報列を元に、予め用意したドメイン知識ベースを参照してメッセージ集合を生成し、それを統合することで概況を生成する手法を提案している [3]。また、小林らは与えら

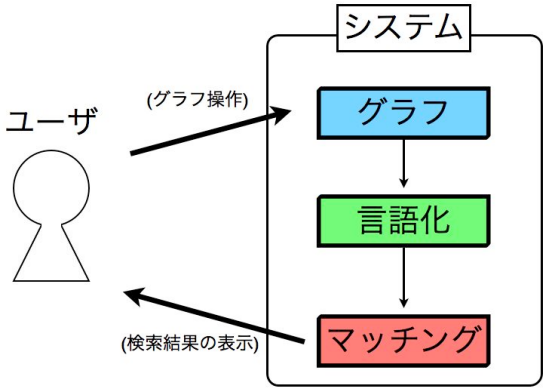


図 1: システムの構成

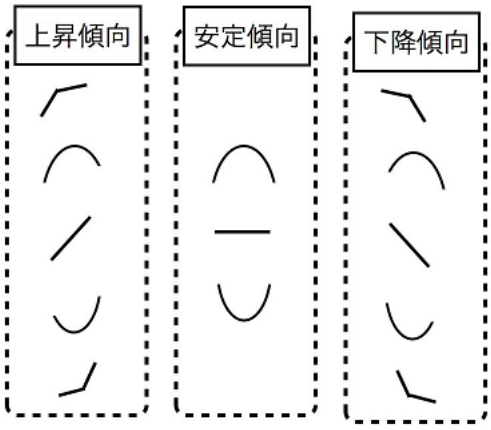


図 2: グラフの概形

れた焦点に基づいて生成する文書のパタンを決定し、そのパタンに応じたテンプレートを用いて説明文テキストを生成する手法を提案している [2, 4]。

## 3 提案手法

本章では、提案システムで用いたグラフの言語化と言語同士のマッチング手法について述べる。

大まかなシステムの構成を図 1 に示す。このシステムでは、検索の例となるグラフの一部分をユーザが指定し、その箇所を言語化して、それと予め蓄積された時系列データ集合に対応する言語表現とのマッチングにより、適切なグラフを見つけ出す。ユーザはこれを繰り返すことで探索的に統計データにアクセスする。以下では、各段階での処理について説明する。

### 3.1 グラフの言語化

本節ではまず、時系列データに対する言語表現特徴の付与方法について述べる。

本研究ではグラフに言語表現を付与するために、グラフの変動、変化の度合い、グラフの概形、の 3 つの特徴に着目している。グラフの変動には「上昇」「下降」「安定」の 3 つの表現を用いる。ユーザが指定したグラフの範囲の終点から始点の値を引き、その差がグラフ全体の上限と下限の差の 1/10 以下であれば「安定」とし、そうでない場合で差の値が正ならば「上昇」、負ならば「下降」と判断することとした。

変化の度合いは、終点から始点を引いた差の値を、指定した範囲の期間で割った値から導出している。傾きの値が 2.0 以上なら「大きく」変動している、傾きの値が 1.0 以下なら「小さく」変動しているとし、それ以外の傾きの場合は「なだらかに」変動していると判断した。

グラフの概形は、始点、中間点、終点の各位置関係をもとに導出している。

グラフの概形の分類は「山型」「谷型」「前半が急で、後半が緩やか」「前半がゆるやかで、後半が急」「直線形」の 5 つを基本とし、それぞれグラフの変動が、上昇、下降、安定の場合に分けることで 13 通りになる。すべてのグラフパターンを図 2 に示す。

それぞれのグラフの分類方法としては、中間点と始点の差と、終点と中間の差の方向がある範囲を越えて異なっている場合は、「山型」か「谷型」のどちらであるかを判断する。システム上でそこにグラフの変動を加えることで、例えば上昇傾向の山形のグラフの場合は「前半は上昇しているが、後半は下降している」と表示し、逆に上昇傾向の谷型のグラフの場合は「前半は下降しているが、後半は上昇している」と表示している。下降傾向の谷型のグラフの場合も、上昇傾向の谷型のグラフと同じように「前半は下降しているが、後半は上昇している」と表示されるが、システム内では別の概形として扱っている。グラフの変動が安定であった場合は「山形に変化するが、最終的に安定している」あるいは「谷型に変化しているが、最終的に安定している」と表示している。

前半と後半のグラフの変動が同じ場合で、前半あるいは後半の差が、全体の 1/8 より大きく、もう一方が全体の 1/12 よりも小さい場合には「前半が急で、後半が緩やか」あるいは「前半がゆるやかで、後半が急」だと判断している。システム上では、そこにグラフの変動を加えることで、例えば上昇傾向の前半が急で、後半が穏やかなグラフなら「前半は大きく上昇しているが、後半はあまり上昇していない」と表示し、上昇傾向で前半が穏やかで、後半が急なグラフなら「前半はあまり上昇していないが、後半は急に上昇している」と表示している。

以上の条件に当てはまらない場合は、「直線形」であ

ると判断している。システム上では、グラフの変動と変化の度合いを組み合わせることで例えば上昇傾向で、大きく変化しているグラフなら「全体的に大きく上昇している」と表示し、下降傾向で、小さく変化しているグラフなら「全体的に小さく下降している」と表示している。グラフの変動が安定であった場合は、変化の度合いに関係なく「全体的に安定している」と表示している。

### 3.2 マッチング手法

ユーザが指定したグラフの範囲をもとに、対象データで同じ範囲のグラフの言語化を行いマッチングを行う。マッチングでは、グラフの変動、変化の度合い、グラフの概形のそれぞれを比較し、0%–100% の間で一致度を導出している。

具体的には、グラフの変動が「安定」、「上昇」、「下降」のいずれかで一致しているかを調べる。このときお互いの傾向が一致しなかった場合、一致度は 0% となる。

一致していた場合は、変化の度合いとグラフの概形での一致度を調べる。それぞれ最大で 40%と 60%の一致度を割り当ててあり、どちらも完全に一致していた場合のみ 100%の一致度となる。今回は変化の度合いよりも概形が似ているグラフの方がユーザが似ていると感じると考えたため、形の方に一致度を多く割り振っている。また、複数の範囲で検索を行った場合は、それぞれに 100 を範囲の数で割った値を一致度の最大として割り当てて計算し、最後に合計する。

概形の一致度を求めるにあたり、「前半が急で、後半が穏やか」なグラフと「前半が穏やかで、後半が急」なグラフでは、グラフの変動が「上昇」であるか「下降」であるかによってグラフの形状が大きく異なっているため、グラフの変動を加味した計算を行っている。上昇傾向の前半が急で、後半が穏やかなグラフと山形のグラフでマッチングを行った場合、類似していると考え、一致度は高めに設定している。逆に下降傾向の前半が急で、後半が穏やかなグラフと山形のグラフでは、一致度は低く設定している。しかし、谷型のグラフとは類似していると考え、高めの一致度としている。前半が急で、後半が穏やかなグラフ同士であっても、グラフの変動が「上昇」と「下降」で異なっている場合は一致度を低くし、上昇傾向の前半が急で、後半が穏やかなグラフと、下降傾向で前半が穏やかで、後半が急なグラフとの一致度は高くしている。

傾きの一致度の導き方を表 1 に、概形の一致度の導き方を表 2 に示す。

表 1: 変化の度合いでのマッチング

| 度合い | 大きい | 普通 | 小さい |
|---|---|---|---|
| 大きい | **40** | 20 | 10 |
| 普通 | 20 | **40** | 20 |
| 小さい | 10 | 20 | **40** |

## 4 実装

本章では実装したプロトタイプシステムについて述べる。

### 4.1 対象データ

対象となるデータは統計局ホームページ[1]から収集した。統計局ホームページ内の総合統計データ月報から主な気象官署別の平均気温や降水量などを対象に 2009 年 3 月から 2012 年 2 月までの 4 年分のデータを収集した。

これらのデータはスケールを合わせるために、値が 0 から 400 程度になるように修正している。

### 4.2 デザイン指針

はじめに、提案システムでは時系列データへのアクセス手段として、言語を用いて検索を行う手法ではなく、グラフの形を提示することでそれを言語化し、類似したグラフを検索する手法を取ることとした。

そのために提案システムでは、見本となる様々な形状のグラフを用意し、検索したいグラフと類似しているグラフの場所を指定することで検索をしたり、探したいグラフの範囲と、その範囲で上昇や下降しているといった傾向を指定することで類似するグラフを検索する機能を設けることとした。

次に、ユーザが指定したグラフはどのようなグラフなのか、グラフの特徴を言語化する。提案システムでは、グラフの指定された範囲を、上昇傾向のグラフか、下降傾向のグラフか、どちらでもない安定傾向のグラフかといった増減傾向と、グラフが全体的に同じ増減傾向か、途中で増減が切り替わっているか、増減の度合いが変化しているかといった、グラフの概形の二つの要素から言語化することとした。

最後に、ユーザが指定したグラフと検索対象となるグラフのマッチングを行う。本システムではさきほど述べたグラフの増減傾向とグラフの概形の二つの要素でマッチングを行い、一致度が高い順に結果を提示する機能を実装することとした。

[1]http://www.stat.go.jp/index.htm

表 2: 概形のマッチング

| | 山 | 急→穏やか (上昇) | 穏やか→急 (下降) | 直線 | 急→穏やか (下降) | 穏やか→急 (上昇) | 谷 |
|---|---|---|---|---|---|---|---|
| 山 | **60** | 50 | 50 | 30 | 20 | 20 | 20 |
| 急→穏やか (上昇) | 50 | **60** | 50 | 30 | 20 | 20 | 20 |
| 穏やか→急 (下降) | 50 | 50 | **60** | 30 | 20 | 20 | 20 |
| 直線 | 30 | 30 | 30 | **60** | 30 | 30 | 30 |
| 急→穏やか (上昇) | 20 | 20 | 20 | 30 | **60** | 50 | 50 |
| 穏やか→急 (下降) | 20 | 20 | 20 | 30 | 50 | **60** | 50 |
| 谷 | 20 | 20 | 20 | 30 | 50 | 50 | **60** |

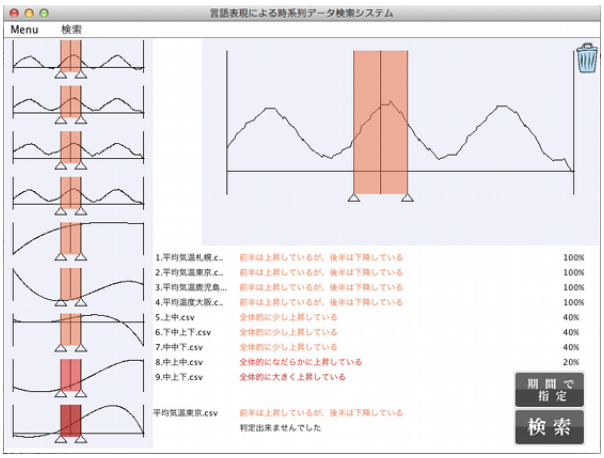


図 3: プロトタイプシステムのインタフェース

## 4.3 操作方法

図 3 が作成したシステムのインタフェースである。

右側に検索を行う際の見本となるグラフが表示され、左側には検索結果のグラフが上から一致度の高い順に表示される。左上のメニューを選択することでメニュー画面が表示され、グラフ名を選択することで見本のグラフを変更することが出来る。見本のグラフ上をドラッグすることで、検索を行うグラフの範囲が指定出来る。グラフの範囲は始点と終点となる縦線を掴むか、その縦線の下に表示されている三角形の部分を掴むことで左右に調整することが出来る。指定している範囲のグラフを言語化したものが下に表示される。右上のゴミ箱にドラッグすることで指定した範囲は消すことが出来る。

始点と終点の間は色が塗られ、その色はグラフが上昇しているか、下降しているか、安定なのか、によって変化し、それぞれその増減の度合いによって色の濃淡が決定される。詳細を表 3 に示す。

範囲の始点と終点以外の場所でドラッグすると、複数の範囲を指定することが出来る。複数の範囲が重なると始点と終点が掴みにくくなるので、掴んだ状態で上下にドラッグすることで下に表示されている三角形の位置をずらすことが出来る。

表 3: グラフの変動と度合いに対応する色

| | 小さい | 普通 | 大きい |
|---|---|---|---|
| 安定 | 薄黄 | 黄 | 濃黄 |
| 上昇 | 薄赤 | 赤 | 濃赤 |
| 下降 | 薄青 | 青 | 濃青 |

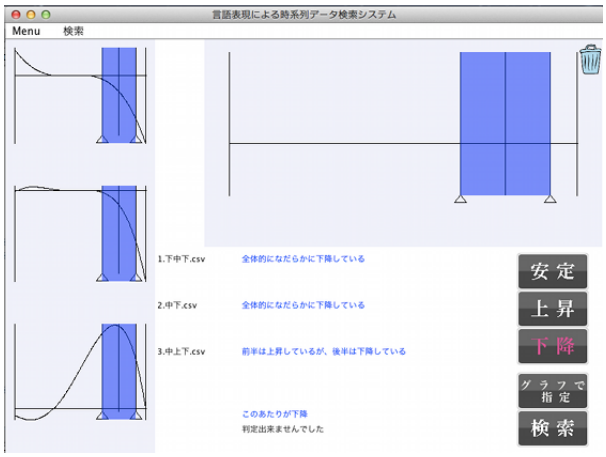


図 4: 期間と変動タイプを指定して検索するモードのインタフェース

右下にあるボタンをクリックすることで、グラフの期間とグラフの変動を指定することで検索を行うモードに切り替えることが出来き、もう一度クリックすることで元のモードに戻すことが出来る。このモードの状態を図 3 に示す。右下にあるボタンを上昇、下降、あるいは安定ボタンをクリックすることで探したいグラフの変動を選択することが出来る。ボタンを選択した状態で、右側の期間指定画面をドラッグすることでどの期間がどのような変動のグラフなのかを指定するこ

とが出来る。

範囲を指定している状態で右下にある検索ボタンをクリックすることで、類似するグラフの検索が行える。

## 5 おわりに

本稿では折れ線グラフの全体傾向や局所的特徴を言語化して、言語による検索を可能にするシステムの実現に向けて、クエリとなるグラフを言語化し生成された言語表現に基づいて検索を行うシステムを提案した。今後、被験者実験を通じてこのシステムの有用性を明らかにし、より的確な検索が行えるシステムへの改良につなげたい。

## 謝辞

本研究は科学研究費補助金基盤研究 (C)(課題番号: 22500209) の助成を受けた。記して謝意を表す。

## 参考文献

[1] Saif Ahmad, Paulo C F de Oliveira, Khurshid Ahmad: Summarization of Multimodal Information, *Proc. 4th International conference on Language Resources and Evaluation*, pp.1049–1052 (2004)

[2] Kobayashi, I.: A Study on Text Generation from Non-verbal Information on 2D Charts, *Proc. Computational Linguistics and Intelligent Text Processing 2nd International Conference*, pp.226–238 (2001)

[3] Kukich, K.: Design of a Knowledge-based Report Generator, *Proc. 21 st Annual Meeting on Association for Computatonal Linguistics*, pp. 145–150 (1983)

[4] 小林 一郎, 渡邉 千明, 奥村 奈穂子: グラフとテキストの協調による知的な情報提示手法: 日経平均株価テキストとグラフの提示を例にして (ヒューマンインタフェース基礎,< 特集 > インタラクション技術の原理と応用), 情報処理学会論文誌, Vol. 48, No. 3, pp.1058–1070 (2007)

[5] 小林瑞希, 小林一郎: 複数の時系列データの関連性発見に基づく言語化の一考察, 情報処理学会第 *74* 回全国大会講演論文集, No.4, pp. 629–630 (2012)

[6] 松下光範, 末吉れいら: 言語表現による時系列データ検索のための基礎検討, 第 *19* 回 *Web* インテリジェンスとインタラクション研究会, pp.31–32 (2011)

# 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

# Phickle: A System for Supporting Cross-Modal Information Access Triggered by Photographs

田中和広 [1*]　松下光範 [2]
Kazuhiro Tanaka[1]　Mitsunori Matsushita[2]

[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:** We aim to support cross-modal information access triggered by photographs. Toward this purpose, we propose a method to facilitate information retrieval based on content (e.g., people, objects, events) or meta-data (e.g., date, place) of photographs. Among content or meta-data, we focus on the date on which a photograph is taken. We introduce Phickle, a system that facilitates time-series information retrieval. When a user focuses on the date of a photograph, the system provides access to other information at the time the photograph is taken. We conducted an experiment to find the points of improvement of the system. On the basis of the results of this experiment, we obtained positive opinions about a function for browsing time-series information related to users' photographs. However, we found that the availability of information access based on the date of photographs needs to be improved, because most participants didn't use this function.

## 1　はじめに

ディジタルカメラが普及し、自身の活動や体験の記録が容易になった。写真は、後日見返すことで撮影当時の雰囲気や出来事を振り返るだけでなく、そこに写っている内容に触発されて情報探索行為へ移る際のトリガとなることもある。例えば、初めて学会発表した時の写真を見て、「今年はどんな発表があるのだろう」という興味を持ち、検索に移ることがあるだろう。加えて、発表内容を調べていると開催地が仙台であることを知り、「仙台の名物は何だろう」といった、それまでとは異なる新たな興味を抱くこともあるだろう。本研究では、このような探索過程で移り変わるユーザの興味に基づいて、求める情報へのアクセスを円滑にするシステムの実現を目指している [1]。しかし、情報はテキスト、画像、音声などが混在しているため、これらのモダリティの違いに関わらず、求める情報への横断的なアクセスの支援が必要となる。その一環として、本稿では写真に写るものを手がかりとして得られる人物やもの、出来事などの情報 (コンテンツ情報) と、EXIF 情報のような撮影日や撮影場所などの情報 (メタ情報) をトリガとした情報アクセスの支援を志向したシステム Phickle を提案する。それと共に、実験を通じて本システムの改善点を整理し、今後の展望を考察する。

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

## 2　関連研究

情報検索技術が向上し、キーワード検索エンジンが普及している。キーワード検索は、ユーザが自らの要求を言語化するが、要求や目的が明確でない場合や、明確でも調べることを顕在化できない場合には十分な支援ができない。このような曖昧な要求に基づく情報探索を Exploratory Search [2] と呼ぶ。Exploratory Search を行う人々は、(1) 目的としている分野や領域をよく理解していない (つまり、その分野や領域の知識を付ける必要がある)、(2) 目的を達成する方法 (技術または手順) について確信がない、(3) 目的が不確かである、という特徴がある [3]。この不確かさが Exploratory Search では重要であり、その明確化が課題の一つとなる。Marchionini は、そのために行われる行為を Lookup、Learn、Inves-

tigate の 3 種類に分類している [2]。Lookup はユーザが生成したクエリと適合する情報へアクセスする行為であり、既知の情報を検索したり、質問に対する解答を得たりする。Learn は、単なる情報の獲得ではなく、情報を新たな知識にする知的活動を含む。これは探索過程で得た情報の意味や考えの理解・解釈、情報や概念の比較などの行為に当たる。Investigate は、既存の情報を新たな知識や情報へ加工するために、知識を分析、統合、評価する行為である。これは、単なる知識獲得ではなく、知識を活用する高次な知的活動に当たる。これらの行為の中で、Exploratory Search では特に Learn と Investigate といった知的活動を含む行為が重要であり、これらを繰り返すことで探索者の知識は増大し、情報を得る度に探索者の要求は変化する。

Exploratory Search と同様に、探索過程で変化する情報要求を考慮したモデルに Berrypicking[4] がある。図 1 に、Berrypicking における探索者の行動モデルを示す。このモデルでは、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、探索者が考えや情報要求を変化させながら、目的の達成や問題解決を図る。Exploratory Search での行動も、これと類似している。本研究では、このような要求が明確でないユーザが変化する興味に基づいて行う情報探索を対象としている。その支援に向けて、本稿では写真をトリガとした情報アクセスに着目し、コンテンツ情報やメタ情報を利用した支援を試みる。写真が持つこれらの要素を利用した研究がいくつかなされている。

PLUM[5] は、大量の写真を撮影場所や日時に基づいて地図上に配置することで撮影者の移動経路を表示し、撮影者の行動を観察可能にする写真閲覧システムである。このシステムでは、写真を撮影場所と日時によってクラスタリングし、その中から代表画像を選択できるようにすることで、写真同士が重ならないようにし、写真と移動経路の閲覧を損なわないようにしている。

Crandall ら [6] や小関ら [7] は、撮影した位置情報と画像の特徴によって、撮影スポットを推薦するシステムを提案している。Crandall らは、大量の地理情報付き写真と画像特徴を用いて、多くの人が訪れる人気スポットや、ランドマークのある主要地域が得られることを示している。小関らは、位置情報と画像特徴の他に時間情報も反映させることで、特定の地域・期間によく撮影されるスポットを推薦する研究を進めている。

捧ら [8] は、時間、空間、人間関係の 3 つの要素を利用したライフログ写真の閲覧手法を提案している。この手法は、写真の撮影日、場所、人物のいずれかを指定することで写真をクラスタリングする。これにより、ユーザが探したい写真に関する記憶が曖昧でも、効率的に写真を探索できるシステムの実現を目指している。

Yee ら [9] は、大量の画像コレクション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　amenco 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階層的な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。

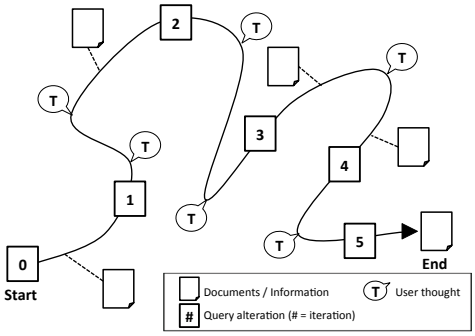


図 1: Berrypicking モデル (文献 [4] 参照)

これらの研究は、写真のコンテンツ情報やメタ情報を利用しており、本研究の目的である写真をトリガとしたアクセスを検討する上で有用なものである。しかし、これらの研究はいずれもアクセス対象が写真のみであり、様々なモダリティの情報を対象としていない。

写真をトリガとした異なる情報へのアクセスの支援として、本研究では写真の撮影日を利用し、写真と共に撮影当時のニュースを提示するシステム PHOTMOSPHERE を提案した [10]。本システムでは、写真から思い出される “記憶” と外在化された “記録” を紐付けることで記憶を豊かにする支援を試みた。しかし、本システムの情報アクセスは one-shot であり、繰り返される情報アクセスを考慮しなかった。また、一枚の写真をトリガとした場合のみを対象としたため、複数の写真によって生まれる興味を考慮していなかった。

以上を踏まえ、本研究では複数の写真を扱い、ユーザの移り変わる要求や興味を考慮し、様々なモダリティの情報へのアクセスを円滑にするシステムを提案する。

# 3　システムの実装

## 3.1　対象とするインタラクション

本節では、本研究で対象とする情報探索の例を述べる。

B さんは、自らが研究発表した時の写真を見て「今年はどんな発表があるのだろう」と興味を持った。発表タイトルを調べると、興味のある発表がいくつか見られた。発表を聴きに行きたいと思い、開催場所を見ると仙台であることを知り、「名物には何があったかな」と思い始めた。名物を調べると、牛タンが有名であると知り、美味しい店を調べ始めた。さらに、仙台で他のものも食べたいと思い、仙台の料理を探索し始めた。

この例では、B さんの興味は得られる情報により変化し、最初の学会発表への興味が、最終的には仙台の料理への興味に変わっている。この探索で、B さんが興味を持った要素がトリガの情報とアクセスする情報の両方に含まれている。例えば、学会について調べている際に開催場所の仙台をトリガとして、仙台の名物へとアクセスしている。この点に着目し、本研究ではユーザがトリガの情報に含まれる要素に興味を持った際に、その要素を元にアクセスできる情報の候補を提示することでシームレスな情報アクセスの支援を試みる。

## 3.2 システムデザイン

ここで、写真に含まれる要素を整理する。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写真に写る人物やもの、その時に起きた出来事などのことである。現在、人物やものを識別するために画像認識技術が研究されている [11]。また、iPhoto[1] や Picasa[2] などでは、顔認識によって人物を認識するサービスが提供されている。しかし、この技術は人物が誰か、ものが何かの特定は実用化できていないため、これらの主観的な要素は、ユーザが入力する方式をとっている。また、写真のメタ情報とは EXIF 情報[3] として規格化されている情報のことである。EXIF 情報とは、ディジタルカメラで撮影した時に、画像データと併せて保存される付属情報のことであり、撮影日時や機種名、シャッタースピード、絞り値の設定といった撮影に関する情報と、圧縮モード、色空間、画素数などの主画像のデータを読み取るための情報が含まれている。

本稿では、コンテンツ情報やメタ情報をトリガとした情報アクセスの支援に向けて、時間情報に着目する。その理由は、写真が過去を想起させる情報であり、撮影日という時間情報が重要であると考えたためである。加えて、時間情報に基づいて移り変わる興味による情報探索を考慮し、時系列情報の探索を題材とする。

その支援に向けて、横断的なアクセスを想定し、写真以外にテキストや音楽の情報も利用する。今回は、テキスト情報に YOMIURI ONLINE[4] のニュースを、音楽に関する情報に過去のヒット曲に関する情報を利用する。このニュースとヒット曲を合わせて時事情報と呼ぶ。時事情報は、2011 年 5 月から 2012 年 11 月までの期間を対象に人手で集め、その情報が発表された年月をメタ情報として付与した。また、時系列情報はある観点に基づいて時間の流れに沿って纏め上げた情報のことであるため、年月だけではなく「民主党政権」「野球」「ヒット曲」などのトピックも付与した。

[1] https://www.apple.com/jp/ilife/iphoto/
[2] http://picasa.google.com/
[3] http://www.cipa.jp/exifprint/contents\textunderscorej/01exif1\textunderscorej.html
[4] http://www.yomiuri.co.jp/index.htm

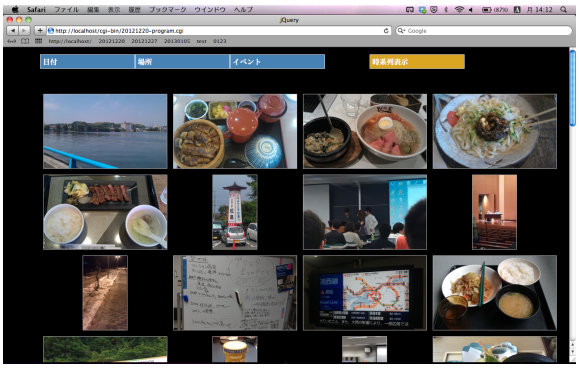


図 2: 写真の一覧表示


図 3: 時系列情報の比較

## 3.3 プロトタイプシステム

本節では、システムの実装のために行った 3 度のプロトタイプシステムの改良について述べる。この改良では、実施者 (著者) が機能を説明しながら被験者が操作し、問題点や改善点を指摘してもらった。この調査は合計 3 回行い、各調査に 4 人の被験者が協力した。被験者は筆者の研究室に所属する情報系の大学院生または教員である。3 度の調査で用いたプロトタイプシステムを各々 Ver.1、Ver.2、Ver.3 システムと呼ぶ。

Ver.1 システムは、写真閲覧のために最低限必要な機能を備え、時系列情報の探索を重視して実装した。

初期画面では写真を一覧表示できる (図 2)。写真は、図 2 上部の撮影日、撮影場所、イベントごとに観点を選んで表示させられる。例えば撮影日を選ぶとメニューが出現し、そこから日付を一つ選択するとその日付の写真を一覧できる。その中から写真を一枚選ぶと拡大表示できる。写真は図 2 右上の時系列表示ボタンを押すことで、時間軸に沿って表示させられる。時系列に並んだ写真の上にカーソルを持っていくとメニューが現れ、その撮影日を元にアクセスできる時事情報の見出しが提示される。その中に興味を持つものがあれば、選択することで詳細な情報を閲覧できる。この選択と同時にその情報が持つトピックに関する時系列情報が

表示される。これにより時系列情報へのシームレスなアクセスが可能となり、情報同士を比較できる (図 3)。

Ver.1 システムでは時系列情報の比較やアクセスを重視したが、本システムを用いた調査の結果、時間情報をトリガとしたアクセスや写真閲覧の支援が不十分であると指摘を受けた。挙げられた意見の抜粋を示す。

- 写真の表示切り替えを画面上部のプルダウンメニューで行うことは、特定の写真に興味を持った際のシームレスな情報アクセスに適していない
- 時事情報へのアクセスが時系列表示からしかできないのは、簡便なアクセスの支援とは言えない
- 時系列表示で、詳細な情報を見ようと見出しをクリックしたが、意図せず時系列情報も提示された

以上の指摘を元に、Ver.2 システムを実装した。Ver.2 システムでは、写真や時事情報の要素をトリガとしたアクセスに重点を置いた。そのため、時系列情報を探索する機能は敢えて省いて実装した。なお、Ver.2 システムの実装についての説明は省略する。本システムを用いた調査の結果、情報アクセスが簡便になった反面、以下のような情報探索における問題点が挙げられた。

- 前に見ていた情報に戻れないため探索しづらい
- 探索を進めていく内に、今何に関する情報を閲覧しているのかが分からなくなる

以上の指摘を元に、Ver.3 システムを実装した。Ver.3 システムでは、写真をトリガとした情報アクセスを改善すると共に、何の情報を見ているか、何の情報へアクセスできるかを提示するように改良した。加えて、時系列情報を探索する機能を追加し、戻る機能も取り入れた。本システムを用いた調査の結果、否定的な意見は少なくなり、以下のような機能向上の意見を得た。

- 写真が持つメタ情報やコンテンツ情報をもっと有効に使って、表示させる情報を変化させる
- 写真から時事情報へのアクセスについて、最初からアクセス可能な情報を出すのではなく、撮影日に基づいたアクセスが可能であることを示したボタンを設置し、それをクリックすることでアクセスできる情報の一覧を表示するようにする

以上の指摘を元に、システム Phickle を実装する。

### 3.4 Phickle

図 4 から図 9 にシステムの表示画面を示す。初期画面では、写真をイベントごとに纏めて表示させている

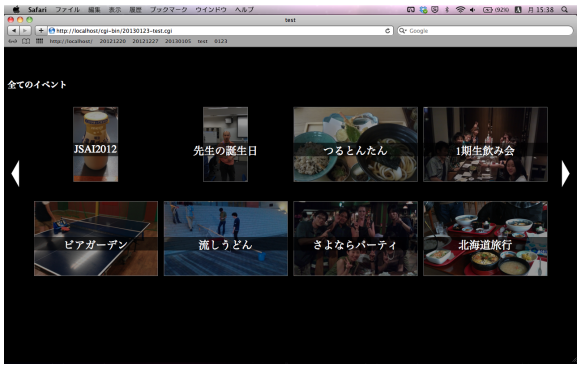


図 4: イベントごとの表示

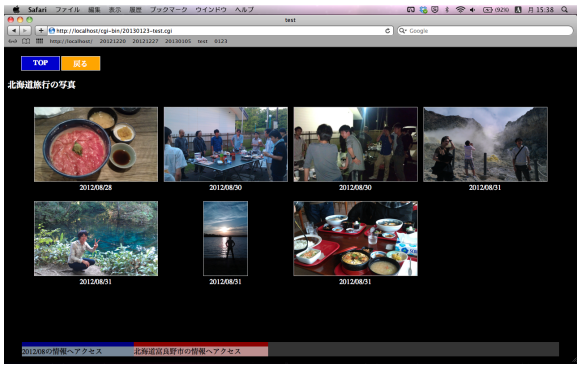


図 5: あるイベントの写真の一覧表示

(図 4)。見たいイベントを選ぶと、その際に撮影した写真を閲覧できる (図 5)。写真を一枚選択すると拡大表示できる。本システムでは、画面上部が今見ている情報を表すタグ、中央がコンテンツ情報、下部がメタ情報を示している。図 5 と写真を拡大表示させた画面の下部には、写真の撮影日と撮影場所を記載し、これらをクリックすると、そのメタ情報によってアクセスできる情報を取得できる (図 6 )。「時事情報を見る」を選択すると同じ月にあった時事情報の一覧 (図 7) へアクセスできる。時事情報の一覧から興味のある見出しを選択すると、詳細な情報へアクセスできる (図 8)。時事情報も図 6 と同様に画面下部にメタ情報からアクセス可能な情報の候補を取得できる。また、記事中のトピック名（図 8 の [民主党政権]）をクリックすると、そのトピックに関連する情報の一覧へアクセスできる。これは、時事情報の内容に興味を持った際の情報アクセスを想定したものである。Ver.3 システムでは時系列表示ボタンを常に表示させていたが、本システムは時系列表示にする必要がある時にだけ提示されるようにした。このボタンをクリックすると、閲覧していたトピックの情報を時系列に表示できる。時系列情報は以前に見たものを蓄積しておき、それらと比較して閲覧できるようにした (図 9)。加えて、様々なモダリティの情報へのアクセスを想定し、音楽の再生機能を備えた。

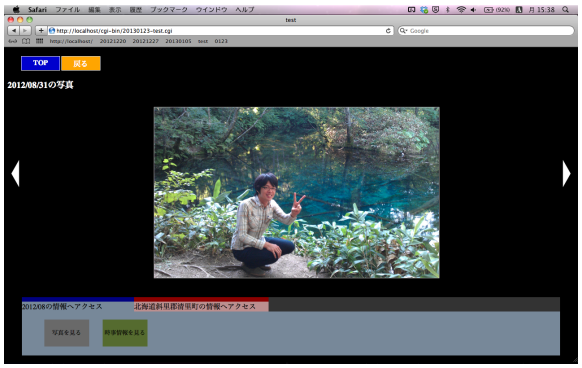


図 6: メタ情報をトリガにアクセス可能な情報の提示

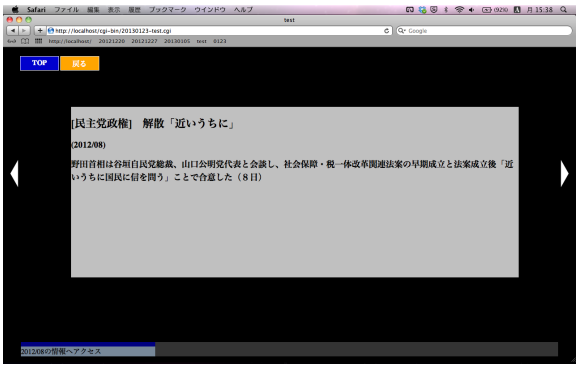


図 8: 時事情報の拡大表示

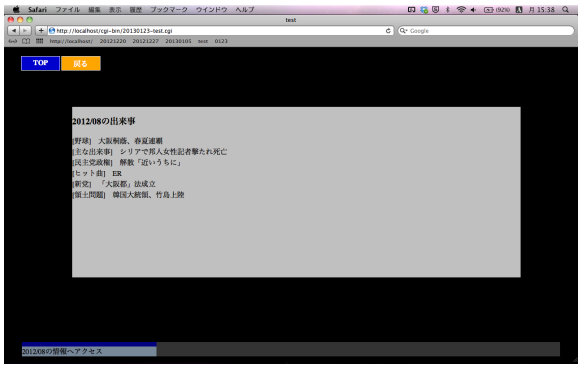


図 7: ある月の時事情報の一覧表示

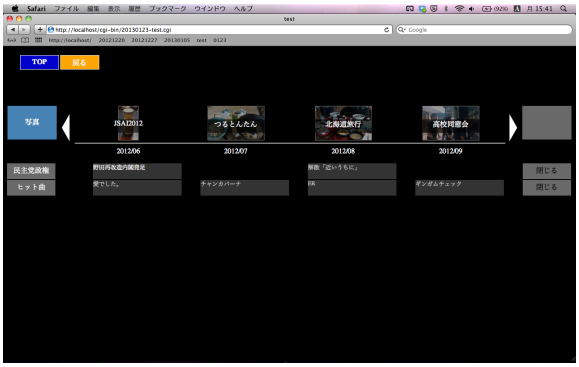


図 9: 時系列情報の比較

# 4 実験

## 4.1 実験の目的

本実験の目的は、Phickle の問題点や改善点の整理である。それに向けて、認知的ウォークスルーの手法を取り入れた。この手法は、エンドユーザの協力を必要としない評価手法であり、特にシステムの学習容易性に焦点を当てたシステムの評価に適している [12]。本来の認知的ウォークスルーは探査学習を対象としており、被験者に操作方法を教示しないで最初にタスクだけを提示し、そのタスクの達成のために行われる操作ステップへの問題点や改善点を収集する。しかし、本研究では探索過程で移り変わる興味を対象としているため、最初にタスクを与えるのではなく、探索過程で被験者にいくつかの興味を提示するように構成した。

## 4.2 実験の構成

Phickle は写真の閲覧、コンテンツ情報やメタ情報による情報アクセス、時系列情報の閲覧、横断的な情報アクセスが主な機能である。これらの機能を含めるように以下の 9 つの操作ステップを設定した。

**ステップ 1** イベント名をクリックして、そのイベントの写真一覧へ遷移する

**ステップ 2** 写真をクリックして、拡大表示させる

**ステップ 3** メタ情報をクリックして、撮影当時の時事情報へアクセスする

**ステップ 4** 個々の時事情報を拡大して表示させる

**ステップ 5** 時事情報中のジャンル名 [民主党政権] をクリックする

**ステップ 6** 時系列表示ボタンをクリックする

**ステップ 7** イベントの写真をクリックして、同じ月のイベントの写真を見る

**ステップ 8** 音楽を再生させる

**ステップ 9** 個々の時事情報を閲覧する

これらの操作を指示しないように、提示する興味を設定した。興味の構成は、目的や要求が曖昧なユーザを想定し、ブラウジングとその過程で生まれる興味による情報アクセスを考慮した。なお、各興味の文末には対応する操作ステップを記す。被験者にはこれらの興味を持った想定でシステムを操作するように促した。

1. あなたは、今自分が撮影した写真を整理しています。あなたは、全ての写真をイベントごとに纏め、名前を付けて管理しています。写真を整理している途中で、ふと過去の思い出を振り返りたいと思いました。【ステップ 1】【ステップ 2】
2. 思い出を振り返り、あなたは北海道へ旅行に行った時の出来事について詳しく思い出したくなりました。【ステップ 3】
3. あなたは、北海道へ旅行に行った時期に起きた時事情報についてより詳しく知りたいと思いました。【ステップ 4】
4. 時事情報について概観し、あなたは「民主党政権」に関する動向に興味を持ち、詳しく知りたいと思いました。【ステップ 5】【ステップ 6】
5. 時系列に並べられた「民主党政権」の情報を見て、あなたは時間の流れに沿った変化についての理解を深めたいと思いました。【ステップ 7】
6. 時間の流れに沿って情報を俯瞰し、あなたは TPP に交渉参加したニュースと京都大学との合同研究会に参加したことが同時期であったことに興味を持ち、その当時の出来事についてより詳しく知りたいと思いました。
7. 京都大学との合同研究会の時期にあった時事情報を概観し、あなたは「ヒット曲」に興味を持ちました。過去の「ヒット曲」にどのようなものがあったかを知りたくなり、先ほど見たのと同じように自分自身の過去の写真と見比べたいと思いました。【ステップ 8】【ステップ 9】

各操作への指摘を受けるために、質問シートを作成した。質問シートには、(1) ユーザは本システムによって、興味を達成するためにその操作をしようと試みるか、(2) ユーザは本システムを見て、その操作が利用可能であると正しく理解できるか、(3) ユーザは本システム上で自身の興味と操作手順を正しく関連付けることができるか、(4) 本システムからのフィードバックを元に、ユーザは興味に基づいた探索が行えたと理解できるか、という質問を記載した。実験後にはインタビュー形式で、システムの各表示画面、操作ボタン、時系列表示についてのアンケートに回答してもらった。

## 4.3 実験手続き

被験者は、我々の研究室に所属していない情報学専攻の大学院生 8 名 (男性 7 名、女性 1 名) である。実験前の説明では、被験者に本システムのユーザとなった想定で、問題点や改善点を指摘するように促した。そのため、被験者には提示される興味を本当に自身が抱いた興味であると想定し、システムをどう使いたいかを考え、操作するように依頼した。加えて、本実験では、実験者 (筆者) の所持する写真を用いたため、被験者には自身の写真を用いていると想定してもらった。

本実験では、前節で述べた興味を被験者に一つずつ与え、一つの興味が達成された段階で、次の興味を被験者に与えた。各興味を終えるごとに、その興味に含まれる操作ステップの質問シートを提示した。質問シートでは、4 つの質問に Yes か No で回答してもらい、問題点や改善点を指摘するためのコメントを記述してもらった。本実験は時間の制約を設けなかったため、被験者が興味を達成したと思った段階で操作を止めるように促した。また、実験者側でその興味内で想定している以外の操作を被験者が行った場合や、実験者側の判断で被験者が興味を達成したと思われた段階で、操作を止めるように求めた。被験者が操作を止めた際、実験者側で想定していた操作ステップを行わない場合もある。その場合には、質問 1 〜 4 の全てに No を選択させ、使用しなかった理由や原因を記述してもらった。

## 4.4 実験結果

各質問に対して Yes と回答した被験者数をステップ別に纏めたものを図 10 から図 13 に示す。この図における横軸の数字は、ステップの番号を表している。

写真の閲覧機能は全ての被験者が操作を行い、問題となる指摘は少なかった。本機能は、ステップ 1 やステップ 2 に当たる。両方のステップ共に、質問 (1) に対して全被験者が Yes と回答するなど、4 つの質問に対して Yes と回答した数は高い傾向にあった。しかし、被験者によってはこれらのステップが最初の操作であったため、何が出来るかをすぐに理解できず戸惑う場合もあった。また、本操作は被験者が類似した機能を使用した経験があったため利用できたが、慣れないユーザには分かりづらい可能性があると指摘された。

コンテンツ情報やメタ情報をトリガとした情報アクセスは、提案システムの主要な機能であるが、問題点がいくつか挙げられた。本機能は、ステップ 3 やステップ 5 に当たる。ステップ 3 は 5 人が操作しなかったため、全質問で Yes と回答した数が少なく、ステップ 5 は質問 (2) において Yes と回答した被験者が 1 人のみであった。ステップ 3 で操作を行わなかった理由に、本機能に気付かなかった、利用できることや使い方が分かりづらかったなどの意見が挙げられた。ステップ 5 の質問 (2) に対して 7 人が No と答えた理由として、クリックできることが分かりづらいという指摘があった。同様の指摘がステップ 4 でもあり、ステップ 4 は

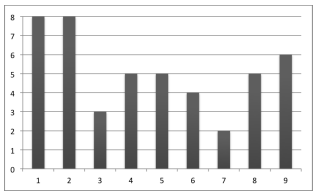

図 10: 質問 (1) に対して Yes と回答した被験者の数

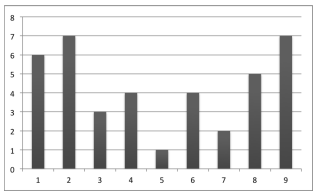

図 11: 質問 (2) に対して Yes と回答した被験者の数

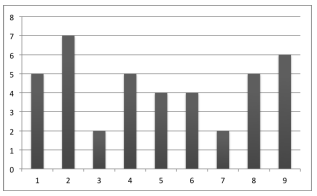

図 12: 質問 (3) に対して Yes と回答した被験者の数

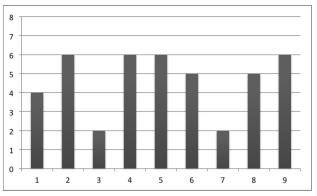

図 13: 質問 (4) に対して Yes と回答した被験者の数

8 人中 7 人が操作を行い、全ての質問に対する Yes の回答数は高い傾向にあったが、一覧表示された時事情報の見出しがクリックできると見た目だけでは分からなかったなどの意見が挙げられた。その原因として、他のテキストと同じフォント、文字色であるためという意見を得た。しかし、マウスカーソルを持っていくことでカーソルが指の形に変化し、文字色が赤色に変化したため、クリックできると気付いたようであった。

時系列情報の閲覧機能は、機能の理解が不十分であり、その点に対する指摘が挙げられた。本機能は、ステップ 6 、ステップ 7、ステップ 9 に当たる。ステップ 6 は 3 人が、ステップ 7 は 6 人が操作しなかったため、質問に対する Yes の回答数は低い傾向にあったが、ステップ 9 は全ての質問で Yes の回答数が高かった。ステップ 6 に対しては、その機能自体が理解しづらかったため、操作しなかった、何が起こるのか試しに操作したなどの意見を得た。ステップ 7 で操作を行わなかった理由として、実験で示した興味が「時事情報に関する理解を深めたい」であったため写真は関係ないと思ったという意見や、時系列表示で時事情報と写真を比較できたため操作しなかったという意見があった。

横断的な情報アクセスは、ステップ 3、ステップ 7、ステップ 8 に当たる。ステップ 8 は、 3 人の被験者が操作を行わなかったが、使用した被験者は全質問に Yes と答えた。使用しなかった理由に、気付かなかった、実験の環境上再生させるか迷ったなどの意見があった。

# 5 考察

## 5.1 提案システムの到達点

写真の閲覧機能では、写真のイベントごとの閲覧や拡大表示の機能を備えた。これらは図 10 に示すように、全被験者が使用を試みると回答し、実際に操作した。このことから、写真閲覧に最低限必要な機能を搭載できたと考えられる。

コンテンツ情報やメタ情報をトリガとした情報アクセスでは、写真の撮影日、時事情報の日付、トピック名を元にした関連情報へのアクセスを可能にした。図 11 に示すように、ステップ 3 は半分以上の被験者が操作を行わず、ステップ 5 はクリックできることが分かりづらいなど課題は残るが、本機能によって共通要素を持つ情報への円滑なアクセスを可能にした。

時系列情報の閲覧では、月ごとに写真と時事情報を表示させる機能を搭載した。事後アンケートによれば、写真と時事情報を対応させて閲覧できるため、自身の経験と比較して情報を理解できたと肯定的な意見が得られ、本機能が有用である可能性を見ることができた。

横断的な情報アクセスでは、画像、テキスト、音楽情報の利用を想定して実装した。それにより、横断的なアクセスの足掛かりとして、画像からテキストや音楽、テキストや音楽から画像へのアクセスを実現した。

## 5.2 提案システムの課題

写真の閲覧機能では最低限の機能を実装したが、既存の写真閲覧ツールと比べて見劣りする部分が多く、機能が不十分であると感じた被験者がいた。今後は、写真閲覧の機能や使いやすさを改善する必要がある。

コンテンツ情報やメタ情報をトリガとした情報アクセスでは、写真の撮影場所や被写体、テキスト情報における文章などの要素を考慮しなかった。今後はこれらの要素も考慮し、実験で指摘された部分を改善する。

時系列情報の閲覧では、時系列表示が初めてのユーザには分かりづらいことが示唆された。それを踏まえ、

時系列表示がどのような機能、メリットがあるかを理解しやすく示すことが必要となると考えられる。

横断的な情報アクセスでは、今回用いた時事情報が主にテキスト情報であり、様々なモダリティの情報を十分に考慮できなかった。実験で、時事情報を音声や画像と共に閲覧したいという意見を得たため、今後はそれらを組み合わせた提示を実現したいと考えている。

また、今回の実験では被験者に対して興味を実験者側から提示し、加えて実験者側で用意した写真を被験者自身の写真であると想定してもらった。それによって、システムの改善点や問題点を収集できたが、実際に情報探索過程でユーザが抱く興味を十分に考慮できなかった。今後は実際のユーザが利用できるようにし、システムの評価実験を改めて行いたいと考えている。

## 6 むすび

本稿では、写真をトリガとした横断的な情報アクセスの支援に向けて、写真のコンテンツ情報やメタ情報に着目した探索を円滑にする手法を検討した。その実現に向け、時系列情報の探索を題材にシステム Phickle を実装し、実験を通じて改善点を整理した。今後画像認識技術が向上し、写真に写る人物やものが特定できれば、写真のコンテンツ情報をトリガとしたアクセスが実現できると考えられる。時系列表示では、文献 [13] などの人の記憶に関する研究を参考に、ユーザの体験や活動との紐付けを強化できると考えられる。また、Web のような膨大な情報を対象とするために、ユーザの興味に基づいて情報を編纂する技術を取り入れることも課題となる。

## 謝辞

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

## 参考文献

[1] 田中和広, 松下光範: 写真をトリガとした時系列情報へのアクセスを支援するシステム, 第 13 回 AI 若手の集い (2012).

[2] Marchionini, G.: Exploratory Search: From Finding To Understanding, *Communications of the ACM*, Vol. 49, No. 4, pp. 41-46 (2006)

[3] White, R. W.: *Exploratory Search: Beyond the Query-Response Paradigm*, Morgan andClaypool Publishers (2009)

[4] 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).

[5] 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, Vol. 141, No. 12, pp. 1-6 (2009)

[6] Crandall, D., Backstrom, L., Huttenlocher, D. and Kleinberg, J.: Mapping the World's Photos, In *Proc. of the 18th International Conference on World Wide Web*, pp. 761-770

[7] 小関基徳, 熊野雅仁, 亀井貴行, 小野景子, 木村昌弘: 写真属性と画像特徴を用いたホット撮影スポット・アノテーション, 第 2 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. 40-47 (2012)

[8] 捧隆二, 佃洸摂, 中村聡史, 田中克己: 時間・空間・人物情報に基づくインタラクションによるライフログ画像の探索手法の提案, *DEIM Forum 2012 D9-4* (2012)

[9] Yee, K. P., Swearingen, K., Li, K. and Hearst, M.: Faceted Metadata for Image Search and Browsing, In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 401-408 (2003)

[10] 田中和広, 松下光範: 可視化プラットフォームの実現に向けたグラフ描画セレクタの基礎検討, 第 3 回人工知能学会情報編纂研究会 (2010)

[11] 柳井啓司: 一般物体認識の現状と今後, 情報処理学会論文誌 : コンピュータビジョン・イメージメディア, Vol. 48, pp. 1-24 (2007)

[12] 堀雅洋, 加藤隆: HCI の拡張モデルに基づく認知的ウォークスルー法の改良 : Web ユーザ ビリティ評価における問題発見効率, 情報処理学会論文誌, Vol. 48, No. 3, pp. 1071-1084 (2007).

[13] Isola, P., Xiao, J., Torralba, A. and Oliva, A.: What makes an image memorable?, In *Proc. of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 145-152 (2011)