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11回)
SIG-AM-11-01

従属クラスタ動的生成機構の導入による Must-Link 制約付き K-means 法の拡張に関する提案

Proposal of Must-Link Constrained K-means with Dynamic Generation of Subordinate Clusters

井本博之 [1]　高間康史 [1]

Hiroyuki Imoto[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: This paper proposes to extend must-link constrained K-means clustering by introducing dynamic generation of subordinate clusters. When clustering high-dimensional data there is a case where data which should belong to the same cluster form several distinct groups in a data space. In order to handle such a case without using distance metric learning, the proposed method generates subordinate clusters for each data group, which are merged after finishing K-means clustering. Result of a comparison experiment with a baseline method shows the effectiveness of the proposed method in terms of success rate and NMI (normalized mutual information).

1. はじめに

本稿では, Must-Link 制約を利用して従属クラスタを生成する機構を導入した制約付き K-means クラスタリング手法を提案する．クラスタリングはデータ群を類似した複数のグループに分ける操作であり，データ全体を俯瞰的にみる目的でデータマイニングの初期分析などによく用いられる．一般的なクラスタリングアルゴリズムは正解データを必要としない教師なし機械学習であるが，自動生成されるクラスタではユーザの要求する結果を得られない場合が多く存在する．そのため，近年ではユーザの意思をクラスタリング結果に反映させる目的で，ユーザフィードバックを利用して半教師あり機械学習を行う制約付きクラスタリングが研究されている．

制約付きクラスタリングで一般的に用いられる制約形式の 1 つに対制約があり，データ対が同一のクラスタに属すべきであるという Must-Link 制約と，データ対が異なるクラスタに属するべきであるという Cannot-Link 制約の 2 種類から構成される．対制約は様々なクラスタリング手法に適用可能[1]な他，インタラクティブに効率的な対制約付与を行うシステム[2][3]が提案されている．

対制約を利用した制約付きクラスタリングの手法としては, CCL (Constrained Complete-Link) [4]のような距離ベースのものと，COP K-means [5]のような制約ベースのものが提案されている．距離ベースの手法では, Must-Link 制約を付与されたデータ対は近くあるいはデータ間距離が 0，Cannot-Link 制約を付与されたデータ対は遠くあるいはデータ間距離が∞になるようなデータ空間に写像した後にクラスタリングを行う．距離ベースの手法は，K-means などの従来クラスタリング手法で制約を満たすクラスタを求めることができるが，元の距離空間とは異なる空間におけるクラスタリングとなるため結果の解釈が困難となる場合がある．結果のクラスタがどのような意味を持つかという解釈は実際にデータ分析を行う場合，非常に大きな意味を持つと考えられる．さらに，距離行列を計算するには，データ数 N に対し $O(N^2)$ の計算量が必要となるため大量データの分析には多大な時間的コストがかかることが問題となる．

一方，制約ベースの手法では，与えられた対制約をそのまま満たしながら目的関数を最適化してクラスタリングを行う．代表的手法である COP K-means は，単純かつ高速なクラスタリングアルゴリズムとして実用的に良く用いられる K-means に対制約を導入した手法であり，クラスタ割当て時に Must-Link 制約と Cannot-Link 制約の全てを満たすクラスタの中で，最も距離の近いクラスタにデータを割り当てる手法である．空間の写像などは行われないため，

クラスタリング結果についてデータそのものの属性に基づいた解釈が可能である．また，計算量の観点ではCOP K-meansの計算オーダが $O(N)$ となるため距離ベースの手法より優れている．しかしながら，地理的に離れた場所に同種データが存在する様なデータセットや，文書のように高次元のデータで同一クラスタにしたいデータが空間上の一カ所に集中していない場合などは，同一クラスタにまとめられるべきデータが異なる領域に分かれて存在することが考えられる．そのような場合に，データ空間内で離れた位置にあるデータ間に Must-Link 制約が付与された場合などはCOP K-meansでは良好な結果が得られないことがある．例えば，図 1 に示す様に，両端にあるデータグループ間に Must-Link 制約が付与された場合，COP K-means (K = 2) では図 2 に示した様なクラスタが得られてしまう．

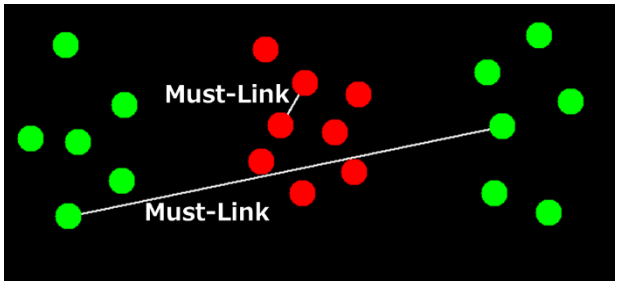


図 1．2 次元人工データセット

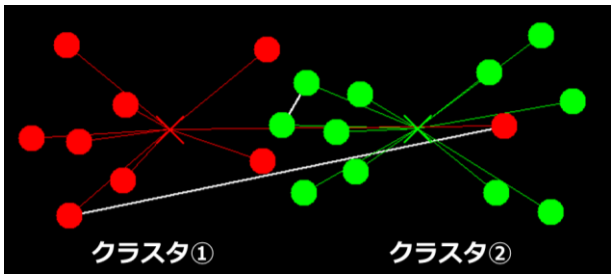


図 2．COP K-means の実行結果

提案手法ではCOP K-meansをベースとし，同一クラスタにまとめられるべきデータが複数領域に分かれて存在する場合には，それぞれの領域に対応した従属クラスタを動的に生成する．クラスタリング終了後，Must-Link によって繋がれた従属クラスタを統合することによって，距離学習せずにクラスタリング結果を求める．Must-Link 制約のみを含む人工データセットを用いて提案手法と COP K-means との比較実験を行った結果，NMI，クラスタリング成功率ともにCOP K-meansよりも良好な結果であることを示す．

2. 関連研究

本節では提案手法のベースとなるCOP K-meansについて述べ，関連研究として距離学習を用いている岡部ら[6]の制約付きグラフカットによる逐次クラスタリング手法を取り上げ，提案手法との違いについて述べる．

2.1. COP K-means クラスタリング

COP K-means は制約ベースの代表的手法であり，対制約をインスタンスレベルで K-means に組み込んでクラスタリングを行う．その基本アルゴリズムは以下の通りとなる．

① k 個のクラスタの初期値を設定する．
② データに付与された対制約を全て満たすクラスタのうち最も近いクラスタに各データを割当てる．
③ 各クラスタの重心位置を計算する．
④ ②，③を繰り返し，②の前後で所属クラスタに変更がなくなった時点で終了とする．

ただし②の処理の時，対制約のペアとなるデータのクラスタ再割り当てが行われていない場合はその制約を無視する．データに対制約が付与されていない場合は K-means と同様に，最も重心との距離が近いクラスタに割当てる．なお，全ての対制約を満たすクラスタが存在しなかった場合，強制終了となる．

COP K-means では制約数が多くなればなるほど計算量が大きくなるものの，計算オーダは $O(N)$であり高速なクラスタリングが期待できる．しかしながら，クラスタリングが正常に終了するかについては順序に大きく依存することや，正しい制約を付与したにも関わらずクラスタリング精度が落ちる場合がある[7]など，いくつか問題が指摘されている．

2.2. 制約付きグラフカットによる逐次クラスタリング

岡部らは，目的関数と制約条件を半正定値計画問題 (SDP：Semi-Definite Programming) で定式化し距離学習を行う手法を提案している[6]．アルゴリズムの概要を以下に示す．

① データ集合から非類似度に応じた重み付き隣接行列を作成する．
② 隣接行列からグラフラプラシアンを作成し，

最大グラフカット問題を定式化する．

③ 最大グラフカット問題を SDP による緩和問題として再定式化し，対制約条件を組み込む．

④ SDP を解いて得られた解行列を基にデータ集合を 2 分割する．

⑤ 生成されたクラスタのうち最大のデータ数を持つクラスタを選択し，②～④を繰り返して 2 分割操作を行う．

⑥ ⑤の操作をあらかじめ設定したクラスタ数になるまで繰り返す．

なお，岡部らの手法では Cannot-Link 制約は用いず，Must-Link 制約のみを用いている．これは，初回の 2 分割から Cannot-Link 制約も無理に満たそうとするため，Cannot-Link 制約が複数存在するとクラスタリングに悪影響を及ぼす可能性があるためである．Must-Link 制約のみを用いた COP K-means との比較実験を行った結果，NMI の値は複数のデータセットに対して互角もしくは優位な結果であったとしているが，計算時間は COP K-means が圧倒的に良い結果を示している．これは，①における隣接行列の計算に $O(N^2)$ かかることに加えて，SDP の処理にも多くの計算コストがかかるためである．

3. 従属クラスタ生成機構を持つ制約付きクラスタリング

3.1. クラスタリングアルゴリズムの概要

提案手法では COP K-means を拡張し，1 節で示した図 1，2 のように離れたデータ間に張られた Must-Link 制約によってデータが距離の遠いクラスタに割り当てられそうになった場合，動的に従属クラスタを生成する機構を導入する．さらに，クラスタリング終了後，Must-Link で繋がれたデータを含むクラスタ同士を 1 つに統合することにより，複数の領域に存在するクラスタを生成する．

提案手法のフローチャートを図 3 に示す．提案手法では K-means の初期値依存の影響を避けるため，1 回目のクラスタ割当てでは従属クラスタの生成は行わない．また，提案手法及び COP K-means では，ある領域にクラスタが集中した場合，Must-Link 制約により距離の近いクラスタ同士でデータの奪い合いが起こり，クラスタリングが収束しないことがあることを予備実験で確認したため，重心計算とクラスタ割当てのループ回数 *step* が閾値 *L* を超えた場合クラスタリングを終了とする．さらに，Must-Link 制約によるクラスタ統合のみではあらかじめ指定したクラスタ数とならない場合があり，その場合はクラスタ重心の距離が近いクラスタを統合する凝集型クラスタ統合を併用する．従属クラスタ生成機構，Must-Link クラスタ統合，凝集型クラスタ統合の詳細については 3.2，3.3，3.4 節にそれぞれまとめる．

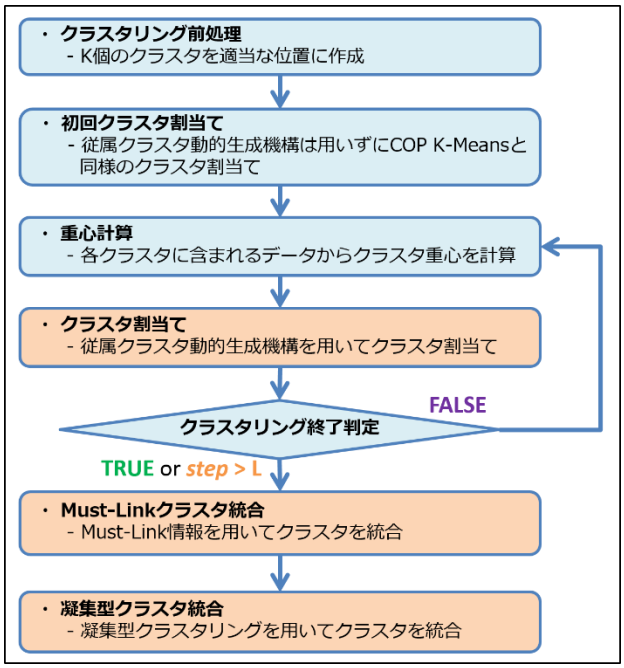


図 3. 提案手法のフローチャート

3.2. 従属クラスタ動的生成機構

提案手法では，図 4 のようにデータ x が Must-Link 制約によって距離の遠いクラスタ c_1 に割当てられそうになった場合，x の位置にクラスタ c_s を生成し，その後 x は c_s に固定割当てとする．距離の近い/遠いの判定には閾値 *th* を用い，距離が *th* よりも大きい場合にクラスタ生成を行う．ただし，同じ位置におけるクラスタ生成は行うべきではないという考えから，1 つのデータが行えるクラスタ生成は 1 回までに制限する．また，クラスタ生成を行うと次のクラスタ割当てによって各クラスタの位置が大きく変動することが考えられるため，1 回のクラスタ割当て時に行えるクラスタ生成も 1 回に制限する．

従属クラスタ動的生成機構を用いてデータ x のクラスタを決定する手続きを図 5 に示す．各データ x について，x に対する従属クラスタ生成が行われてなく，かつ Must-Link 制約が付与されている場合にクラスタ生成を行う（8 行目以降）．SEARCH_CANDIDATECLUSTER(x, C) により既存クラスタから割当て候補クラスタ集合 *CC* を求め（8

行目)，その中で最も近いクラスタに x を割り当てる（9 行目）．ただし，そのクラスタと x の距離が閾値以上の場合は初回クラスタ割当て時を除き従属クラスタを生成する (13 行目)．また，効率の良い探索を行うため，クラスタ割当て時に Must-Link 制約をクラスタに登録する (3，15 行目)．

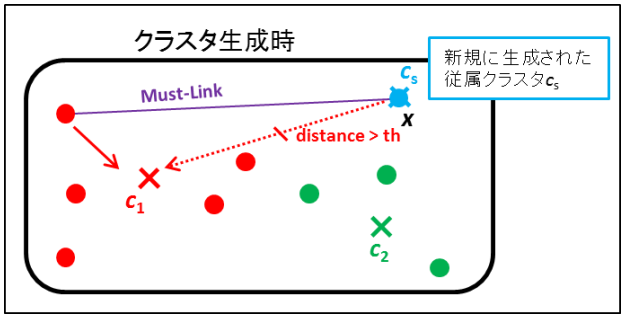


図 4．従属クラスタ生成時の様子

```
1  if x.fixed == TRUE {
2    x.curc = x.prec;
3    x.curc.REGISTER_MUSTLINK(x.mlg);
4  else{
5    if x.mlg = NULL{
6      x.curc = MOST_NEARCLUSTER(x, C);
7    }else{
8      CC = SEARCH_CANDIDATECLUSTER(x, C);
9      x.curc = MOST_NEARCLUSTER(x, CC);
10     if DIS(x, x.curc) > th
11       & step > 1
12       & crflg == FALSE{
13       x.curc = CREATE_CLUSTER(x);
14     }
15     x.curc.REGISTER_MUSTLINK(x.mlg);
16   }
17 }
```

図 5．従属クラスタ動的生成機構を用いたクラスタ割当て

SEARCH_CANDIDATECLUSTER(x, C) では，図 6 のようなクラスタ間の Must-Link による接続関係を利用して割当て候補クラスタ集合を求める．各 *mlg* は Must-Link 制約によって直接繋がったデータ集合を表しており，例えばデータ a と b の間に Must-Link 制約があり，a，b をそれぞれ含むクラスタ $c1$, $c2$ がある場合，a，b を含む *mlg* と $c1$，$c2$ が接続される．図 6 の例では，x の所属する *mlg*1 にはクラスタ $c1$, $c2$ に割り当てられたデータが所属しており，$c1$ に割り当てられたデータには *mlg*1，*mlg*2，*mlg*3 に所属するものが存在している．このような接続関係を x の所属する *mlg*1 を起点として全探索し，得られたクラスタ集合を候補クラスタ集合として出力する．

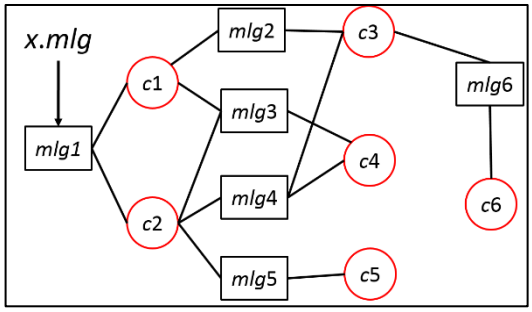
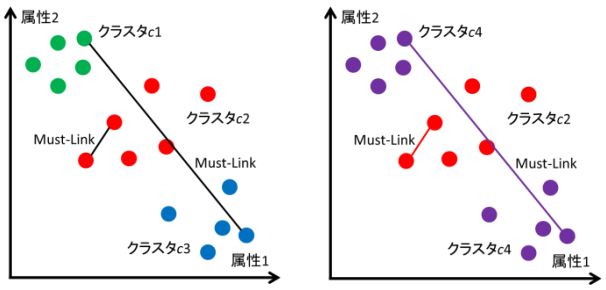


図 6．データ x と各クラスタとの接続関係

3.3. Must-Link クラスタ統合

3.1 節で述べたように提案手法では，クラスタリング終了後に Must-Link クラスタ統合を行う．クラスタ $c1$ 内のデータが他のクラスタ $c2$ 内のデータと Must-Link 制約で繋がっている場合，$c1$ と $c2$ を統合する．例えば，クラスタリング終了時の状態が図 7 (a) の様であった場合，クラスタ $c1$ とクラスタ $c3$ の間に Must-Link 制約が張られているため両者は統合され，図 7 (b)に示すクラスタ $c4$ が形成される．このクラスタは，「属性 1 あるいは属性 2 が排他的に大きい」という特徴を持つクラスタであると解釈できる．



(a) 統合前　　(b) 統合後

図 7．Must-Link クラスタ統合の例

なお，Must-Link クラスタ統合で統合されるクラスタ集合は，3.2 節の図 6 で示した様な接続関係にあるクラスタの集合であり，図 5 に示された SEARCH_CANDIDATECLUSTER() によって求められる．

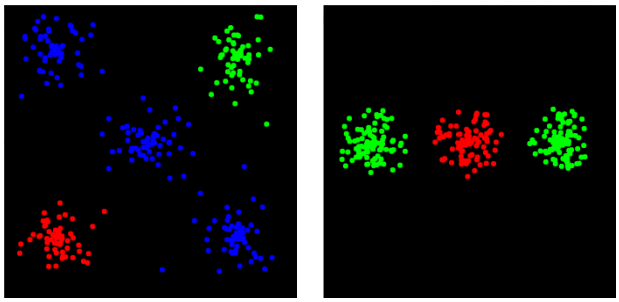
3.4. 凝集型クラスタ統合

3.1 節で述べたように，3.3 節の Must-Link クラスタ統合のみでは初期設定したクラスタ数 K 以上のクラスタ数となってしまう場合がある，これを補うた

め凝集型クラスタ統合を行う．凝集型クラスタ統合では Must-Link クラスタ統合後の各クラスタ中心をデータとし，凝集型クラスタリング (AHC) の最短距離法を適用してクラスタ数が K となるまで統合する[8]．凝集型クラスタ統合は，クラスタ生成過多により，本来は一つにまとめられるべきデータ集合が複数のクラスタに分割されてしまっている状態を修正することを目的とするため，鎖効果を期待して最短距離法を採用する．

4. 実験

図 8 に示すデータ数 300 のデータセット A，B に対して COP K-means 及び提案手法を用いてクラスタリングを行い，比較実験を行った．図において，同じクラスタに属するデータは同じ色としている．評価指標には正規化相互情報量 (NMI : normalized mutual information) を用い，NMI = 1.0 となる場合をクラスタリング成功とし, その割合を成功率とした．



a. データセット A　　b. データセット B

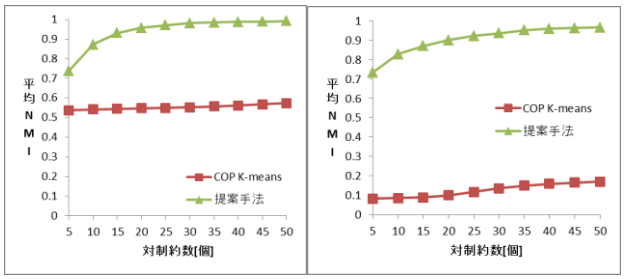
図 8. 実験に使用した 2 次元人工データセット

対制約は両手法ともに Must-Link 制約のみを用い，正解データペアに対して，5, 10, 15, 20, 25, 30, 35, 40, 45, 50 個をランダムに付与し，各 10000 回クラスタリングを行い平均値及び正解率を算出した．なお，データの範囲は各次元 [0，700] とし，提案手法における従属クラスタ動的生成機構の閾値 th は距離の 2 乗値に対して設定するため，データセット A に対しては th = 60000，データセット B に対しては th = 30000 とした．この値は予備実験の結果，各データセットにおいて良い結果が得られたものを選択している. なお, 全手法に対して最大ループ回数 L は 100 回と設定した．

提案手法と COP K-means について，図 9 に平均 NMI，図 10 に成功率，図 11 に平均実行時間を比較した結果をそれぞれ示す．また，図 12 に提案手法における平均最終クラスタ数の推移を示す．データセット A，B に対して，平均 NMI，成功率共に，COP K-means よりも提案手法の方が高い値を示している．特にデータセット B においてその傾向はより顕著であり，線形分離可能でない場合に有効性が期待できると考える．

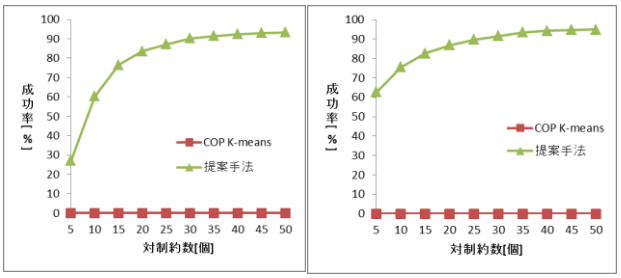
平均最終クラスタ数は対制約数の増加に伴い，データセット A，B ともにわずかな上昇を示しているが，収束の傾向もみられる．この現象に対する検証にはより大きなデータセットにおける実験が必要であり，また閾値の設定とも関連すると考える．

平均実行時間に関して，提案手法では対制約数の 2 乗オーダで増加している．このように計算量が増加するのは，対制約数の増加により生成される従属クラスタ数および mlg の数が増えた結果, 3.2 節に示した SEARCH_CANDIDATECLUSTER()などの計算時間が増加することが原因と考えられ，今後検証を行う予定である．しかしながら，制約付きクラスタリングにおける制約はユーザに付与されることが想定されており，大量に付与されるケースは少ないため，大きな問題とならないと考える．ただし，高間ら[3]のように複数の対制約を自動生成するインタフェースと組み合わせる場合には，制約生成数を抑制するなどの対策が必要と考える．



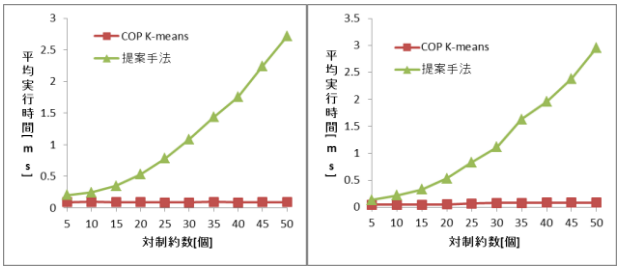
(a) データセット A　　(b) データセット B

図 9. 平均 NMI の推移



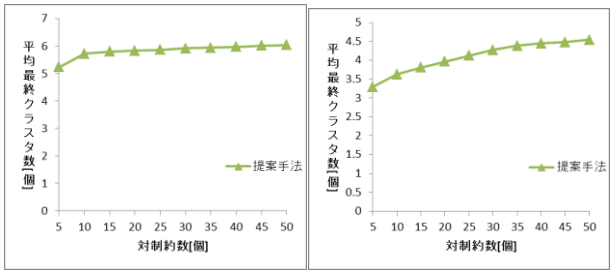
(a) データセット A　　(b) データセット B

図 10. 成功率の推移



(a) データセット A　　(b) データセット B

図 11. 平均実行時間の推移



(a) データセット A　　(b) データセット B

図 12. 平均最終クラスタ数の推移

5. まとめ

本稿では，同一クラスタにまとめられるべきデータが空間上の複数の領域に分かれて存在するケースにも対応できるように，COP K-means に対し Must-Link 制約を基にした従属クラスタ動的生成機構を導入した拡張手法を提案した．2 次元人工データを用いた比較実験により，同一クラスタに属するデータが平面上の異なる領域に分散して存在するような場合に，COP K-means よりも NMI, 成功率共に良好な結果が得られることを示した．また，計算量に関しても対制約数の 2 乗オーダで上昇してしまうものの，データ数 *N* に対しては K-means と同様であるため，距離学習を利用したクラスタリングなどに比べ，高速なクラスタリングが期待できる．距離学習を用いた場合とのクラスタリング精度の比較は今後行う予定である．また，提案手法の特徴は，得られたクラスタの解釈が元の空間上で行えることであり，その利点についてもユーザ実験により検証する予定である．

提案手法では従属クラスタ動的生成機構に対して閾値 *th* を指定する必要があり，適切な閾値をいかに決定するかが今後の課題となる．また，Cannot-Link 制約も利用可能なように拡張することも検討している．

参考文献

[1] 寺見明久，宮本定明：階層的クラスタリングにおける対制約の導入のための二つのアプローチ, FSS2010, MD2-4，2010.

[2] 山田誠二，水上淳貴，岡部正幸：インタラクティブ制約付きクラスタリングにおける制約選択を支援するインタラクションデザイン，人工知能学会論文誌 Vol.29 No.2，pp. 259-267，2014.

[3] 高間康史，三宅遼祐：グルーピング操作に基づくインタラクティブな対制約生成手法の考察，第 27 回人工知能学会全国大会，2F4-OS-04-31，2013.

[4] D.Klein ， S.D.Kamvar ， C.D.Manning ：“ From Instance-level Constraints to Space-level Constraints : Making the Most of Prior Knowledge in Data Clustering,” in Proc. International Conference on Machine Learning (ICML-2002)，pp. 307-314，2002.

[5] K.Wagstaf, C.Cardie, S.Rogers, S.Schroedl:“Constrained K-means Clustering with Background Knowledge,” in Proc． International Conference on Machine Learning (ICML-2001)，pp. 577–584，2001.

[6] 岡部正幸，山田誠二：制約付きグラフカットによる逐次クラスタリング，人工知能学会論文誌 Vol.27, No.3，pp. 193-203， 2012.

[7] I.Davidson ， K.Wagstaff ， S.Basu ：“ Measuring Constraint-Set Utility for Partitional Clustering Algorithms,” in Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD2015)，pp. 115–126，2006.

[8] 宮本定明：クラスター分析入門 ファジィクラスタリングの理論と応用，森北出版，pp. 88-105，1999.

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第11回)
SIG-AM-11-02

コンテクスト検索エンジンへの ランキング機能の導入に関する検討

Consideration of Introducting Ranking Function to Context Search Engine

手塚 拓哉[1]　山口 晃一[2]　諸 琰俊[2] 桑折 章吾[2] 高間 康史[2]

Takuya Tezuka[1], Koichi Yamaguchi[2], Yanjun Zhu[2], Shogo Kori[2], and Yasufumi Takama[2]

[1] 首都大学東京システムデザイン学部
[1] Faculty of System Design, Tokyo Metropolitan University

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

Abstract: This paper aims to introduce a ranking function to context search engine. Context search engine has been developed for answering trend-related queries. In order to achieve a more efficient search, ranking retrived results, which is one of important function of exsisting Web search engines, is expected to be effective also for the context search engine. This paper discusses several features that could be used for ranking function of context search engines, such as intensity and periodicity of temporal change. The result of ranking retrieved results with the intensity of temporal change is also shown.

1. はじめに

本稿では, 動向に関する問いにタスクを限定したコンテクスト検索エンジンにおいて, より効率的な検索を実現することを目標として, ランキング機能の導入について検討する.

Web 上には多種多様で膨大な量の情報が日々蓄積され続けられている. Web を利用することにより発見することのできる情報も増え, ユーザの求める情報も多様化している. その結果, ユーザの情報要求と既存検索エンジンの提供する基本検索機能の乖離が大きくなっている. この問題に対する解決策の一つとして, タスクを「動向に関する問い」に限定することにより, ドメインを限定せずに高度な検索機能を提供するコンテクスト検索エンジンが提案されている[1]. コンテクスト検索エンジンを利用することにより, 時間的変動の観点から関係のあるアイテムを発見するタスクなどにおいて, 有効性を確認している. 検索対象となる動向データとして Web 上のオープンデータを収集しており, 2015 年 7 月の時点で 27,848 アイテムが検索可能となっている[2]. 今後もデータベースを拡張していくことが予定されているが, 検索対象アイテム数の増加に伴い, 検索結果として返されるアイテム数も増加している. 現在のコンテクスト検索エンジンでは検索結果は順位づけられていないため, 結果の確認にかかるユーザの負担が課題となっている. そこで, より効率的な検索を実現するために, 本稿ではランキング機能の導入について検討する.

現在の Web 検索エンジンでは, PageRank スコアや文書適合度など多様な素性を用いて Web ページのスコアを計算する[3]. また, 各素性のスコアにおける重みはランキング学習[4]を用いて決定することが一般的になっている. 本稿でも, 同様の枠組みによりランキング機能を実装することを検討する.

上述の枠組みによるランキング機能の導入においては, スコア計算に用いる素性, およびランキング学習に用いる訓練データについて検討する必要があるが, 本稿では素性について検討する. 既存検索エンジンは, 基本検索機能としてキーワードベースのクエリを入力とし Web ページを検索結果として出力する. しかし, コンテクスト検索エンジンの検索対象は時系列データであり, クエリはアイテム名と期間, 変動タイプといった違いがある. このため, 既存検索エンジンと同様の素性を用いることができないため, 検索タスク・対象データに適した素性を新たに検討する必要がある.

本稿では, クエリ独立の素性として, 変動の激し

さや周期性などについて検討する．また，変動の激しさを利用したランキング機能を実装した結果を示し，その効果について考察する．

2. 関連研究

2.1. コンテクスト検索エンジン

現在，Web における情報アクセス手段として，キーワードを用いて Web ページを検索する検索エンジンが普及している．しかし，これら既存の検索エンジンには，提供する基本検索機能とユーザの情報要求との乖離が大きいことや，個々の情報要求に合わせ，適切なクエリに分解するのに要するユーザの負担が大きいことが問題として指摘されている．

この問題に対して，動向に関する問いに答える問いう幅広いドメインで見られるタスクに限定することにより，ドメインによらず利用可能という既存検索エンジンの利点を維持しつつ，より高度な検索機能をもつコンテクスト検索エンジンが提案されている[1].

コンテクスト検索エンジンでは，動向情報として「コンテンツとしての動向情報」と「ユーザ活動による動向情報」の 2 種類を Web から収集し，検索対象としている．前者の例として商品の価格や生産量，人口，後者の例として GoogleTrends やきざしランキングから得られるデータなどが収録されている．

それらの時系列データに対する基本検索機能は，Google を利用した検索作業において観測された検索意図[5]を元に決定されている．具体的には，指定したアイテムに関する動向が特徴的変動を示した期間の検索，指定した期間に特徴的変動を示したアイテムの検索，指定したアイテムに関する動向が特徴的変動を示したアイテムの検索の 3 つの基本検索機能が利用可能となっている．また，最大値や急上昇などの 6 種類の特徴的変動タイプを時系列データから抽出し，検索条件として指定可能である．

2.2. ランキング機能に用いる素性

既存のキーワードを用いて Web ページの検索を行う検索エンジンでは，Web ページの検索結果としての適合度を計算するために，多種多様な素性を用いている[3]．それらは，クエリ依存の素性，クエリ独立の素性に大別することができる．クエリ依存の素性とは，入力されたクエリと Web ページの関係からスコアを求めるものであり，BM25[7]や TF-IDF などクエリと Web ページの関連度に関する素性がよく用いられる[8].

これに対して，クエリ独立の素性とは，入力されたクエリに関係なく Web ページのスコアを決定するものであり，Web のリンク構造を利用した Web ページの重要度である PageRank[9]やアンカーテキストなどがある．

他にも Web 検索エンジンで利用されていると思われるランキングの素性として，Twitter や facebook などの SNS のアカウントに信頼度を付与し，投稿された短文のリンクに重みをつけるソーシャルシグナルや，検索履歴やクリックログなどのユーザシグナルを利用した素性がある[10].

ランキングの素性に利用されたものではないが，時系列データの特徴としては，その時間的変動に基づくものが考えられる．蓮井らは，言語表現を用いた時系列データ検索システムを提案している[6]．グラフの変動，変化の度合い，グラフの概形に着目し，グラフの始点と終点の範囲から「上昇」「下降」「安定」，傾きから「大きく」「小さく」「なだらかに」などの特徴を抽出している．

周期性の検出には周波数分析がよく用いられる．動向情報の周期性を判定する方法として，綱元らは web ページがブックマークされるタイミングの周期性を離散フーリエ変換とパワースペクトルを用いて判定し，ブックマークが周期的に利用されるページに関する調査を行っている[11].

3. 提案するランキング機能に用いる素性

本節ではランキングに用いるクエリ独立の素性として，変動の激しさ，周期性，増加/減少傾向の 3 つの素性について検討する．

3.1. 変動の激しさ

動向情報は外的要因の影響で激しい変動をすることがある．顕著な例として，2011 年 3 月の東日本大震災の前後で激しい変動をした動向情報は多く，震災に関連する動向情報として多くのユーザに有益である事が期待できる．動向情報の特徴的変動から関連アイテムや期間を検索するコンテクスト検索エンジンにとって，激しい変動を持つ動向情報の重要性は高いと考える．

本稿では，激しい変動とはデータ値が短期間に大きく変動することと定義する．激しい変動を行う期間と変動の大きさは重要な要素であると考えられる．激しい変動を行う期間を検出するために，コンテクスト検索エンジンで指定可能な変動タイプである急上昇と急降下を利用する．急上昇/急降下は，3 ヶ月

以内に、その動向情報の|最大値 − 最小値| の 1/5 以上の単調増加/減少が見られる期間として定義されている[1]. これを利用して, 急上昇と急降下が発生した期間を動向情報が短期間に大きな変動を行う期間と判断する.

変動の大きさに関して, 動向情報ごとに単位や平均値が異なるため, 固定的な閾値で判断することは現実的ではない. そこで, データ内において変動の占める割合を以下の式で定義する.

$$Intensity = \frac{|V_{start} - V_{end}|}{V_{max} - V_{min}} \qquad (1)$$

ここで, V_{start}, V_{end} は抽出された期間の開始時, 終了時のデータ値をそれぞれ示す. $V_{\max}$, $V_{\min}$ はその動向情報における最大値, 最小値である. この値を動向情報の変動の激しさの素性として扱い, 動向情報ごとに付与する. 複数の激しい変動がある場合は, その動向情報における最大の値を用いる.

3.2. 周期性

野菜の価格や自転車の販売量, Amazon の Google 検索数など周期性を持つ動向情報がある一方で, 乾電池の価格や内閣支持率など周期性の見られない動向情報がある. 周期性を持つ情報の特徴は, 気候,あるいは入学やクリスマスなどといった定期的な行事など周期性を持って発生する要因に影響を受けている点である. これらの要因について関心がある場合, 周期性を持つ動向情報は価値ある情報と考える.

本稿では, 自己相関を用いて動向情報の周期性を判定する. コンテクスト検索エンジンでは動向情報の粒度が月単位であるものがほとんどである[1]ため, 1 年周期の動向情報を主な対象とする. そのため, 動向情報のデータ点が 12 点以下であるか, データに欠損がある動向情報は除外した. 自己相関の計算とピーク値の推定には Matlab の xcorr 関数と findpeaks 関数を用いた. 推定したピークの中にはノイズによるものが含まれるため, 文献[12]を参考に,自己相関が 0.3 以上のピークに限定し, 時差なし以外に 1 つ以上主要な周期を含むものを周期性があると判断する. 周期性がある場合は 1, ない場合は 0 と設定しランキングの素性とする.

例として, 日本梨の価格の時系列データと自己相関のグラフを図 1, 2 に示す. また, ノートブックの価格の時系列データと自己相関のグラフを図 3, 4 に示す. 図 1 から日本梨の価格は毎年 6 月に周期的にピークを迎えていることがわかる. この時, 図 2 においても閾値を超えるピークが複数存在することがわかる. しかし, 図 3 のグラフではその様な傾向は読み取れず, 図 4 においても, 閾値を超えるピークが存在しないことがわかる.

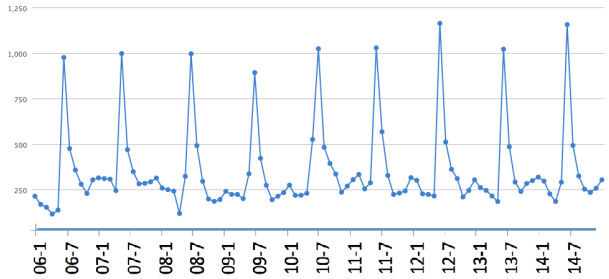


図 1. 日本梨の価格の時系列データ

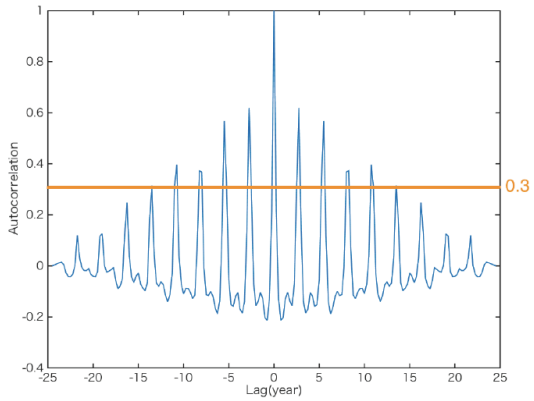


図 2. 日本梨の価格の自己相関

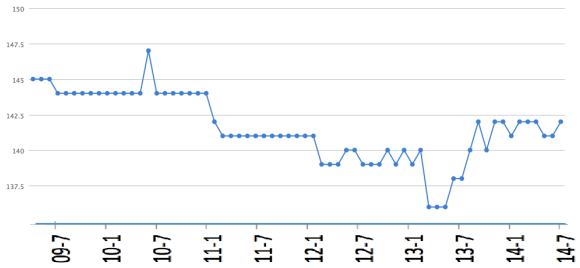


図 3. ノートブックの価格の時系列データ

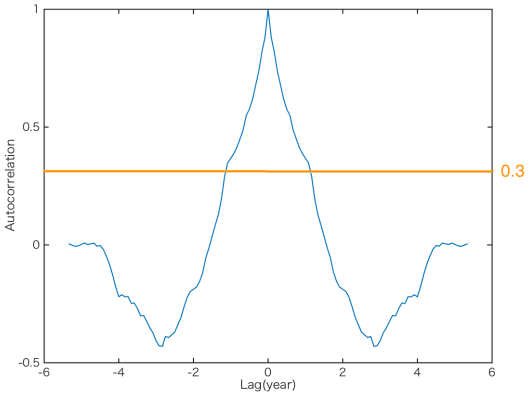


図 4. ノートブックの価格の自己相関

3.3. 増加・減少傾向

動向情報には，一時的な激しい変動や周期的な変動以外にも，右肩上がり/下がりといった傾向を示す変動がある．このような動向情報には，突発的な要因や周期的な要因とは異なる，長期的に普遍な要因の影響があると考える．変動の激しさや周期性とは異なる関連が期待できるため，検索者の目的によっては重要であると考える．

増加・減少傾向を判定するために，式(2)で定義されるピアソンの積率相関係数を用いる．

$$r(y) = \frac{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\bar{x})(y(n)-\bar{y})}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\bar{x})^2}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y(n)-\bar{y})^2}} \quad (2)$$

ここで，$y(n)$ は N 点からなる時系列データの n 番目の点であり，$x(n) = n\text{-}1$ とする．

例えば，何らかの要因によりあるアイテムの生産量が減少し，価格が高騰するなど，同じアイテムに関する動向情報が，同じ要因により反対の変動を示すことがある．このため，ランキングの素性とする場合，増加・減少傾向を区別する必要はないと考え，得られた相関係数の絶対値をランキングの素性とする．

4. ランキング機能の実装と考察

本節では，3 節で述べた変動に関する素性のうち，変動の激しさを用いたランキング機能を実装した結果を示し，その性質について考察する．

2008 年 1 月から 2011 年 12 月に最大値を示した動向情報を検索した結果を図 5 に示す．図 5 から，たちあがれ日本の政党支持率や東京電力の検索数(GoogleTrends) などが上位で検索されていることがわかる．これらの動向情報は 2011 年 3 月に発生した東日本大震災と関連が深いことから，それらがランキングの上位となっていることは，妥当であると考える．

次に，レギュラーガソリンの動向情報が最大値をとる時期に同様に最大値をとる動向情報を検索した結果を図 6 に示す．また，レギュラーガソリンの動向情報の一つである卸価格の動向を図 7 に示す．図 7 からレギュラーガソリンの卸価格は 2008 年 8 月をピークに急激に減少していることがわかる．また，図 6 からレギュラーガソリンと同時期に最大値を示した動向情報のランキング上位には，ストック(生花) や西洋梨など同時期に価格に大きな変動がある動向情報が検索されている．原油価格の高騰が，花しや果物の栽培用の燃料費の増加につながり，販売価格に影響を与えることが指摘されており[13]，これらが上位にランキングされていることは妥当であると考える．

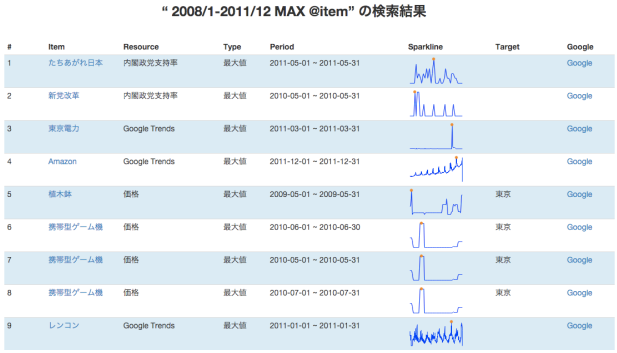


図 5. 2008 年 1 月から 2011 年 11 月に最大値を示した動向情報の検索結果

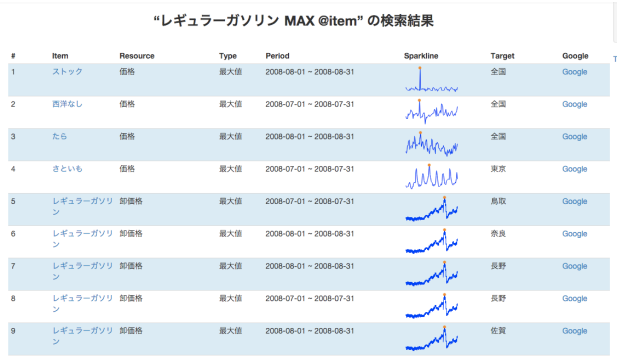


図 6. レギュラーガソリンの動向情報が最大値を示した期間に同様に最大値を示した動向情報の検索結果

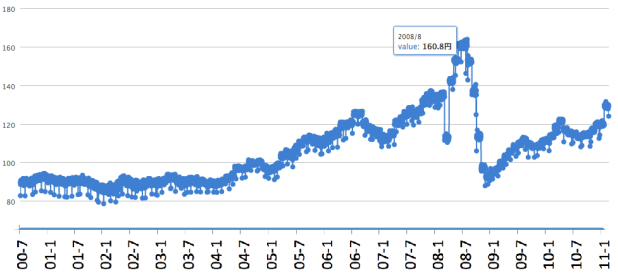


図 7. レギュラーガソリンの卸価格の時系列データ

5. おわりに

本稿では，コンテクスト検索エンジンにおいてより効率的な検索を実現するために，ランキング機能の導入を検討した．変動の激しさ，周期性，増加・減少傾向の 3 つの素性について，期待される役割や計

算方法を示した他, 変動の激しさを素性に用いたランキング機能を実装し, 検索結果の例を示した. 今後は, 本稿で検討したランキング素性を用いてランキング学習を行うために, クリックログやブックマークなどのユーザフィードバックを利用した訓練データの作成を予定している.

参考文献

[1] 高間, 加藤, 桑折, 石川: 動向に関する問いを対象とした検索エンジンの提案, 人工知能学会論文誌, Vol. 30, No. 1, pp. 138-147 (2015)

[2] 高間, Zhu, 桑折, 山口, 瀧口: 動向に関する問いに答える検索エンジンの開発, 人工知能学会第 10 回インタラクティブ情報アクセスと可視化マイニング研究会, pp. 9-15 (2015)

[3] 菱沼, 山口: 検索エンジン最適化の有効性に関する考察, 東京工科大学研究報告, pp. 3-13 (2008)

[4] H. LI: A Short Introduction to Learning to Rank, IEICE Transactions on Information and Systems, Vol. E94-D, No. 10, pp. 1854-1862 (2011)

[5] 桑折, 加藤, 高間: 検索エンジンを用いた情報検索におけるユーザ行動の分析, 人工知能学会第 4 回インタラクティブ情報アクセスと可視化マイニング研究会, pp. 9-14 (2013)

[6] 蓮井, 松下: 言語表現による時系列データ検索システムの提案, 人工知能学会第 3 回インタラクティブ情報アクセスと可視化マイニング研究会, pp. 58-62 (2013)

[7] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu: Okapi at TREC-3, 3rd Text REtrieval Conference, pp. 109-126 (1994)

[8] P. Matthew: Determining Relevance: How Similarity Is Scored, https://moz.com/blog/determining-relevance-how-similarity-is-scored

[9] S. Brin, L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, 7th International Conference World-Wide Web 7, pp. 107-117 (1998)

[10] M. Tober, L. Hennig, D. Furch: SEO Ranking Factors and Rank Correlation 2014 –Google U.S.-, searchmetrics Whitepaper (2015)

[11] 綱元, 亀井, 藤田: 周波数分析を利用した周期的にブックマークされる web ページの特定, 第 74 回情報処理学会全国大会講演論文集, Vol. 2012, No. 1, pp. 719-720 (2012)

[12] MathWorks: 自己相関を使用した周期性の検出, http://jp.mathworks.com/help/signal/ug/find-periodicity-using-autocorrelation.html

[13] 大山, 古在: 園芸用施設の暖房費および CO2 排出量削減(1), 農業および園芸, Vol. 83, No. 11, pp. 1157-1163 (2008)

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11回)
SIG-AM-11-03

コミック工学の これまで と これから

The Past and Future of Comic Computing

松下 光範 [1*]
Mitsunori Matsushita[1]

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

Abstract: The goal of our study is to establish a new research topic named "Comic computing." With the spread of small devices like tablet PC and smart phone, a market for e-books has been growing. In particular, expectation for digital comics is so huge that comics account for the largest portion in the sales amount. Under such circumstances, this paper presents a service concepts that can be realized when the comic contents become computable.

1 はじめに

タブレットやスマートフォン等，ディジタル端末で読むことのできる電子書籍が急速に普及しつつある．「電子書籍ビジネス調査報告書 2015」（インプレス社）によれば，2014 年度の電子書籍市場規模は 1266 億円と推計され，前年度に比べて 35.3% の増加となっている．このうち，ディジタルコミックの売上は約 8 割を占めるといわれており，電子書籍の普及に大きな役割を果たしている．

ディジタルコミックは，従来の紙媒体のコミックと異なり物理的な制約がないため，従来のコミックの枠に囚われない表現（e.g., 話の展開に応じて内容を切り替える，コマに動きを付与する）や自らの環境に最適化させた利用（e.g., 読み手の母語に応じて言語を切り替える，文字の大きさやフォントを変更する）が可能になる．しかし現状では，多くの作品は単に紙媒体のコンテンツをスキャナで取り込んでそのままディジタル化した静的なものであり，ディジタルコミックの可能性を十分に活かせる状況にはない．本研究の目的は，こうした状況を改善し，ディジタルコンテンツならではの利用を可能にすることである．

このような背景の下，本稿ではディジタルコミックをより活用するための技術やその応用について，これまで取り組まれている研究を概観しつつ，ディジタルコミックの可能性や課題について考察する．なお，本稿は文献 [22] をベースとして，その後に行われた研究を中心に加筆修正したものである．

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

2 論点 1: コミックのコード化

コミックコンテンツは，絵と文字が相補的かつ協調的に利用されているクロスモーダルなコンテンツである．そのため，これらを計算機で利用可能にするには，予めコミックの内容を解釈してコミックを構成する要素（i.e., キャラクターや吹き出し，コマ領域など）を抽出し，それらをコード化・構造化して蓄積しておく必要がある．コミックコンテンツは新聞記事などのテキストを主体とした媒体とは異なり，文字が絵のなかに配置され，その位置や字の形にも意味があるため，単純に画像の中から文字情報を抜き出すだけでは不十分であり，どのような形態で記述されているか（フォント情報や大きさ情報），どこに出現したか（位置情報），などの情報についてもコード化しなくてはならない．更に，コミックでは絵と文字が相補的かつ協調的に利用されているため，文字情報のみではなく，絵として描かれているキャラクターやオブジェクトの情報もコード化する対象に含めなくてはならない．

2.1 コミックの構成要素の抽出

現在，コミックは JPEG などの画像ファイルとしてページ単位で与えられているため，その画像の中からコミックを構成する要素を取り出す技術が必要になる．この取り出すべき要素は，主として線やドットで構成される二値の画像として表現されている．そのため，これらをコード化するためには，まず画像処理によって要素を同定する必要がある．

こうした要求に応える技術として，画像処理分野を中心に，コミックの画像ファイルを対象としたコマの識別やスクリーントーンの除去に関する研究，キャラ

クタ/吹き出しの抽出に関する研究などが様々に進められている [11]．以下に，コミックの要素毎に，現在進められている研究事例を示す．

コマの認識　一般的に，コミックではコマの連続によってストーリが展開していくため，コマはコミックの意味的な最小単位として扱われることも多い．このコマの領域同定に関しては，コミックの枠線を識別し，濃度勾配 (intensity gradient) の方向を利用してコマの分割線を同定する手法 [7, 43, 8] や，「コミックのコマは矩形であることが多い」という特徴を利用して，画像内から矩形領域を検出し，それを用いてコマを特定する手法 [6] などが提案されている．いずれの手法でも，概ね 80% を超える精度が報告されており，高い精度でのコマの同定が可能になっている．また，このようにして同定されたコマに対して，ヒューリスティクスを用いてコマを読み進める順序を決定する手法も提案されている [49].

スクリーントーンの除去　スクリーントーンは，コミック作成時に背景や陰影表現，心理的効果の付与を目的として貼付されるシールである．制作がディジタル媒体で行われる場合は，画像描画ツールのエフェクトを用いて付与されることが多い．登場人物(キャラクター)やオブジェクトの認識精度向上を目的として，このスクリーントーンを原画から除去する技術が検討されている．例えば，伊東らは，白黒のコミック画像を LoG (Laplacian of Gaussian) フィルタと FDoG (Flow-based Difference-of-Gaussian) フィルタを用いてスクリーントーン領域と線画領域を分離し，スクリーントーン領域を除去して線画を取り出す手法を提案している [12]．手法の精度は平均で 55% 程度でありまだ改善の余地は残るが，後段のキャラクター同定などの処理の精度向上に寄与する技術として期待される．

登場キャラクターの同定　コミックに登場するキャラクターの識別手法として，HOG 特徴量 (Histograms of Oriented Gradients) を手がかりにして画像内の顔候補を特定し [1]，その顔候補と予め作成したキャラクターの顔画像データベースとのマッチングを行い，顔候補画像がどのキャラクターであるかを識別する手法が提案されている [1, 10]．また，近年では画像の変形に対して頑健さを持つ Deformable Part Model をコミック画像に応用し，より高い精度での顔候補を特定する技術 [52] が提案されている．ただし，現状ではキャラクターによる精度のばらつきが大きく，安定した顔検出ができているとは言いがたい．その理由として (1) コミックに登場するキャラクターの顔は一般に線画で表現されており，実画像の顔認識に比べて識別に利用できる特徴量が限られている，(2) コミック特有の誇張表現ゆえに顔の輪郭や部品のばらつきが大きい，などが考えられる．この点について谷らは，(1) コミック内でのキャラクターの描き分けに髪の色の差異がよく利用される，(2) 連続した一連のコマには同じキャラクターが登場する可能性が高くなる，といったコミック特有のヒューリスティクスを用いて，キャラクター識別精度の向上を試みている [44].

吹き出しの分類　吹き出しの同定に関しては，田中らの手法や Rigaud らの手法が挙げられる．田中らの手法 [42] では，ページ内の文字領域を Ada Boost によって特定し，その領域をもとに吹き出し候補を検出する．また，SVM によって吹き出し形状分類（通信型，曲線型，折れ線型，四角型）を行う．この手法により，86% の吹き出しが同定されている．また，Rigaud らの手法 [35, 34] では，まずテキストの位置を特定してそれを手がかりに吹き出し領域を特定した後，その吹き出しの枠線の変位に着目し，吹き出し領域と枠線との距離を典型的変動パタン（e.g., zigzag, weavy, smooth）に照らして分類している．

これらを勘案すると，画像情報のコミックからそれを構成する要素を抽出したり構造を理解したりするための基礎的技術は，実用に向けて着実に進歩していると結論付けられる．

2.2　コミックコンテンツの構造理解

コミックは，コマを単位とし，それらの連続によって時間経過やストーリの展開を表現している．そのため，2.1 節で抽出された要素を利用するには，単にそれらを抽出するだけでは不十分であり，想定されるコマの順序や場面のセグメントなどを把握し，要素間の関係を構造化する必要がある．加えて，コミックは制作者のアイディアによって日々新しい表現技法が創出されているため，拡張性も担保しておく必要がある．

こうした問題に対して，Wikipedia に記載される項目や書誌情報の目録概念モデルである FRBR (Functional Requirements for Bibliographic Records) を利用してコミックから抽出すべきメタデータのモデル化する研究 [26, 5] や，それを考慮したメタデータ記述フレームワークの研究 [27, 23] が進められている．これらの研究では，メタデータの基盤となる語彙を (1) 知的内容，(2) 書誌記述，(3) 構造記述，(4) グラフィック要素の 4 つのカテゴリに分け，モデル化することにより，特定の利用に限定されない汎用的な知識構築を試みている．反対に，Rigaud らはコマの識別や吹き出しの識別

[1] 顔だけでなく，瞳などパーツ単位での位置情報取得やそれを利用したキャラクターの識別も試みられている [9, 11].

といった画像処理による低次の処理から，ドメインやコミックに関する事前知識を参照しつつボトムアップに構造を獲得していく方法について検討を進めている [33].

これらとは別のアプローチとして，コミックコンテンツに出現するキャラクターやオブジェクトに関する統計情報や台詞の談話構造を利用して，コンテンツの中に登場するキャラクター間の関係を特定する研究も進められている．例えば，Murakami らは「出現頻度が高いキャラクターと，共に出現する人物の間には関係がある」という仮説に基づき，コマ内に共起するキャラクターの頻度情報から，キャラクターの相関関係を推定する研究を行っている [28].

これまでテキストを対象としてそこからの知識獲得やコンテンツの再利用を行う研究が自然言語処理やデータベースなどの分野で進められてきたが，コミックのようなマルチモーダルコンテンツを対象とした研究はその需要にもかかわらずそれほど多くなかった．これは，コミックが ill-formed なコンテンツであるためと，分野を跨った技術が必要になるためである．これらの研究と 2.1 節で述べた研究とが連携することで，コミックコンテンツからの知識構築がより効率的かつ効果的に進められるようになると期待される．

3 論点 2: 獲得された知識の利用

2 章で述べた技術によってコード化されたコミックコンテンツを利用することで，様々な効果が期待される．この章では，コミック制作者の支援，コンテンツの再利用の観点から，獲得された知識の利用について述べる．

3.1 コミック制作者の支援

インターネットの普及や UGC (User Generated Content) 環境の充実に伴い，Blog やコンテンツ共有サービス (e.g., pixiv[2]) を利用して自らが描いたイラストやマンガを公開し，他者に閲覧・評価してもらうことができるようになってきている．こうした状況により，初心者であってもコミックを制作できるように支援する技術に注目が集まっている．既に，コミ Po![3] のような，事前に用意されたキャラクターや表情，オブジェクト等を組み合わることで，絵や図を描くことなくコミックを生成できる商用のコミック作成支援ツールが登場している．

また，POM [14] は，ユーザが自分の描きたい漫画のジャンルや作家名を入力すると，蓄積した過去の作品のデータから，ページ配分・コマ割・構図の候補をユーザに複数提案するシステムである．ユーザは提案された候補の中からイメージに合ったものを選んでカスタマイズし，それに絵と台詞を書き込むことで自分のコンテンツを作り上げることができる．

この他にも，読者の視点移動を考慮した初心者向けコミック作成支援システム [31] の提案や，オリジナルのコミックを作成する際のストーリ構築支援手法の提案 [13]，コミックの作成プロセスに基づき，メタデータを利用することで効率的にネーム（漫画の設計図）を作成・管理できるように支援するツールの提案 [24] が行われている．

これらのコミック作成者支援技術に共通するのは，過去に上梓されたコミックから取得した知識や事前に用意されたプリミティブ（キャラクターやオブジェクト，背景など）を利用している点にある．現状ではこうした知識・データは人手で抽出し作成しているため，制作やメインテナンスのコストがかかる．2 章で述べたようなコミックコンテンツのコード化が進展すれば，効果的かつ効率的に知識やデータを構築できるようになり，これらのシステムにも大きく寄与することが期待される．

3.2 コンテンツコンテンツへのアクセス

出版月報 (2014 年 2 月号) によると，2013 年度には 12,161 タイトルの新刊コミックが発売されている．こうしたコミックの増大に伴い，そのコンテンツに対する情報アクセスのニーズも多様化してきているが，それに応えるシステムは未だ十分とはいえない．例えば，一般的なディジタル書籍販売サイトでは，コミックの表題や著者名，出版社名といった書誌情報による検索は可能であるが，コミック中の特定のシーンを探したい，コミックの内容を手がかりにして表題や著者名を探したい，という要求には応えられない．こうした要求には「Yahoo! 知恵袋[4]」や「教えて goo[5]」などのインターネット上のサイトで質問することである程度解決可能であるが，回答を得るのに時間を要したり，回答が得られなかったりする場合も多い．また，長編コミックを短く要約して内容を短時間で把握したい，登場人物の出現頻度や発話数などの客観指標を知りたいといった要求の場合は，上記のような質問サイトでは要求に沿った回答を得ることが難しい．この問題を改善し，書誌情報だけでなくコンテンツをも対象にした柔軟な情報アクセスを可能にすることで，ディジタルコミックの利便性や有用性が高まると考えている．

Matsui らはスケッチされた画像に基いて，それに類似するコミックコンテンツの領域を検索する手法を提案

[2] http://www.pixiv.net/ (2013 年 4 月 18 日存在確認)
[3] http://www.comipo.com/ (2013 年 4 月 18 日存在確認)
[4] http://chiebukuro.yahoo.co.jp (2015 年 10 月 23 日存在確認)
[5] http://oshiete.goo.ne.jp (2013 年 10 月 23 日存在確認)

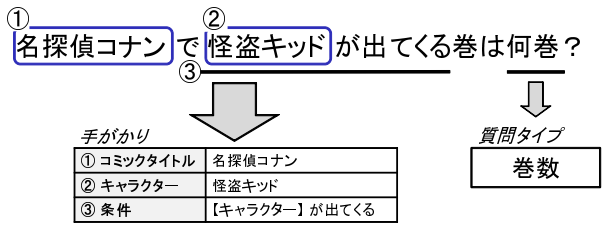


図 1: 手がかり要素の抽出例

している [18, 19]．この手法では，コミックコンテンツに適した画像特徴量として FMEOH (Fine Multi-scale Edge Orientation Histogram) 特徴量を提案・利用している．このシステムでは，検索対象のコミックから様々な大きさの窓を移動させて画像中のエッジ情報を抽出することで予め FMEOH 特徴量を抽出・蓄積しておく．ユーザが手書きで描いた図などをクエリとして蓄積されている FMEOH 特徴量を参照することで，それに類似した画像領域を高速かつ効率的に検索できる．同様に Sun らは，著作物保護のために違法コピーされたコミックを計算機を用いて発見することを目的として，画像全体の類似性とキャラクターの顔部分の類似性のふたつの指標に基いて同一のコミックコンテンツ箇所を検索・抽出する手法を提案している [38, 39]．

また，コミックを対象とした質問応答技術の研究も始まっている [4, 51]．質問応答は現在自然言語処理分野でテキストを対象として精力的に進められている研究の一つであり，コミック質問応答はこれをコミックコンテンツに拡張したものである．現在はその端緒として，ユーザから与えられる質問のタイプ分類方法の検討が進んでいる．図 1 に質問解釈の流れを示す．例えば，ユーザから「名探偵コナンで怪盗キッドが出てくる巻は何巻？」という質問が与えられた場合，この質問文の解釈にあたっては，まず ① がコミックの作品タイトルであるため，「コミックタイトル」として抽出する．次に ② が ① の作品中に登場する人物の名前の一つであるため，「キャラクター」とする．また，③ は②のキャラクターが出現する箇所を意味すると考え，「条件」として設定する．最後に，文末に記述されている文末表現に着目する．この例では，コミックの単位の一つである巻数を聞いているため，「位置に関する質問」に分類できる．これらを手がかりとして応答が生成される．

3.3 コンテンツの再利用

電子化されたコミックの利点の一つとして，再利用が容易である点が挙げられる．そのような再利用を促進する試みの一つとして，携帯電話端末のような表示領域が狭いデバイス上でコミックを閲覧しやすくなるように変換するシステムが提案されている [49]．このシステムでは，コミックのページ内のコマの順序を考慮して順に提示することにより，表示領域の狭さという問題の解消を試みている．また，こうした利用のために，重要な部分の歪みを抑えつつ，アスペクト比を変更する技術（内容に基づくリターゲティング）の研究も進められている [20]．小型の電子端末でコミックを読む際に重要部分だけでも理解できればその内容は概ね理解できるため，こうした技術にも期待が集まる．

コミックの分析は教育工学の分野でも盛んに進められている [41]．特に，コミックが学習や理解に及ぼす影響についての関心が高い．例えば，向後らは学習マンガを題材としてその利用の効果について実験を行なっている．実験の結果から，文章だけの表現に比べてマンガ表現を利用することが，学習内容に対する深い理解の促進や学習に対する関心の増大，長期の記憶保持に寄与する可能性が示唆されている [15]．また，谷本らはコミックの社会的意義を考察するために，コミックコンテンツから抽出した物語構造とアンケートによって獲得した読者世界との照応関係を把握し，コミックの社会的意義を明らかにしようと試みている [45]．コンテンツのコード化は，こうした分析を統制してより広範に行う上でも有用である．コード化されたコミックコンテンツを利用することで，これまでは定性的な分析にとどまっていたコミックコンテンツの分析を定量的な分析に拡張し，あらたなコミック分析の礎を提供できるようになる．

4 論点 3: コミックの表現技法の利用

日本語のコミックは，過去半世紀の間に独特な表現技法を産み出し，進化を遂げてきた [40]．例えば，効果線（流線）を用いることでスピート感を表現したり，コマ割りを工夫することで心理状態や時間経過を表現したりする，などがこれにあたる．更に，現在も日々新しい表現が産み出されている．こうした表現が，コンテンツの読みやすさや魅力の向上をもたらす要因のひとつになっている．このような，コミックの持つ特性を活かしエンタテインメントやプレゼンテーションなどにそれを利用する研究も進められている．本章では，ディジタルコミックを想定した新しい表現の産出に関する研究と，コミックの表現を利用したアプリケーションについて述べる．

4.1 新しい表現の創出

アニメーション等の映像媒体と異なり，従来のコミックでは声も音も文字として表現される．夏目は，コミッ

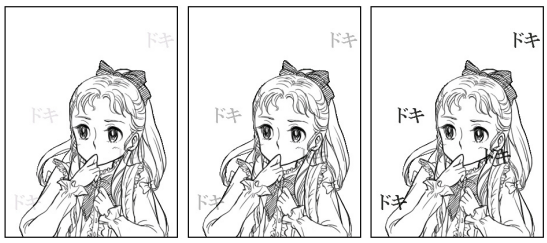


図 2: 音喩「ドキドキ」の動き (文字の点滅)

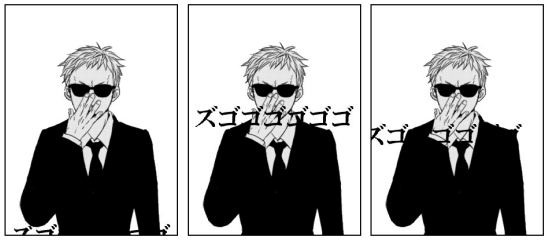


図 3: 音喩「ズゴゴゴゴゴゴ」の動き (振動 + 上昇)

ク中に出現する視覚化された音 (聴覚情報) には擬音語・擬態語の総称であるオノマトペの範疇に含めることが困難な表現が存在するという理由から，これらを「音喩」と呼んだ [30]．音喩はコミック中の環境音 (Sound Effects) を示したり，キャラクターの心理状態を表現したりすることを企図して付与されており，ストーリーに躍動感を与える効果を果たしている．今岡らはこの音喩に着目し，その効果の増大を目指して動きを伴う音喩表現を付与するためのシステムを提案している [21]．このシステムでは，ディジタルコミックの制作者が音喩のカテゴリからコマのシーンに合わせて音喩を選択し，速度や角度，向き等のパラメータを調整することでこの音喩に意図した動きを付与できる．図 2 が「ドキドキ」という音喩のアニメーション，図 3 が「ズゴゴゴゴゴゴ」という音喩のアニメーションである．辞書的意味と音象徴的意味に基づいて，音喩に応じて操作できるパラメータが設定されており，動作の細かい調整が可能になっている．また，これとは反対に音喩の動線を指定することで，その動線に適した音喩表現の候補を提示する研究も行われている [46]．

4.2 理解容易性の向上

コミック表現は，直観的な理解が容易であるという利点を持つ．ここではその利点を活用したアプリケーションとして，代表的なものを幾つか挙げる．

Comic Chat [17] は, Chat の内容をコミックの台詞として表示する．キャラクターの表情やジェスチャで感情を分かりやすく表現できる．また, ComicDiary [36, 37] は，体験の記録と共有のためにコミックの形式を使ったシステムであり，自らの体験を他者に伝えたり共有したりすることを企図している．

藤本らはマンガのコマ割り表現を用いたプレゼンテーションツールを提案している [2, 3]．従来，学会発表や打ち合わせなどで使用されるプレゼンテーションの資料は，Microsoft Powerpoint や Keynote などのツールを利用して作成されたスライドを用いる場合が多い．その場合，同じ形状・サイズの四角いスライドを 1 枚ずつ用いるため，全体の構成もメリハリのない均質なものとなりがちである．藤本らが提案したシステムはこの点の改善を狙ったもので，マンガのコマ割り技法に着目し，自由な形状とサイズのコマをレイアウトして, 時には複数の情報を同時に見せられるプレゼンテーションを作成することを試みている．

コミックの表現は，映像コンテンツの要約や閲覧にも利用されている．ぱらぱらマトリクスは，撮影した映像データを要約し，吹き出しやコマ割りなど，マンガの技法を用いて分かりやすく提示することを目的としたシステムである [16]．この他にも，マンガのコマ割りの概念を援用した表現として，静止画を用いたビデオの要約生成が提案されている [48, 47]．

コマ割り以外のコミックの特徴を利用した研究としては，アバターを介したコミュニケーションのための支援技術が提案されている．マンガのキャラクターを描き分ける際，髪型は重要な特徴である．吉澤らの研究ではこの点に着目し，コミックでの髪型表現の手法を援用して，アバターの表現のために特徴を強調した髪型を生成している [53]．

4.3 実世界インタフェースとの連携

コミック表現の利用は，実世界インタフェースを用いたエンタテインメントシステムでも利用されている．その一つが Manga Generator である．Manga Generator は自分がマンガのキャラクターになるシステムである [29]．深度センサ（KINECT）が設置されたブースで，提示されたストーリに応じて体験者がポーズを取ると，体験者の画像がマンガの中に取り込まれ，ポーズに基づいて決定されたエフェクトが追加されたマンガが出来上がる．図 4 に Manga Generator の出力例を示す．

「聖地巡礼 (Anime Pilgrimage)[6]」はアニメーション作品やコミック作品の新しい楽しみ方を提供するアプリケーションである．近年，作品の舞台となった場所や建物を訪れるツアーが一部の愛好者の間で流行している．例えば「けいおん！」というコミックでは舞台となった滋賀県犬上郡の豊郷小学校が，「忍たま乱太

[6] https://play.google.com/store/apps/details?id=com.animepilgrimage.android&hl=ja (2013 年 4 月 18 日存在確認)

図 4: MangaGenerator が生成した画像

郎」というコミックでは兵庫県尼崎市の七松八幡神社が愛好家の間で「聖地」と呼ばれ，そこを訪れるツアー（聖地巡礼と呼ばれる）が行われている．本アプリケーションはこういったツアーを気軽に楽しむために，コミックコンテンツと位置情報とを紐付けて提示する携帯端末型のアプリケーションである．類似したものとしては，コミックコンテンツに出現するレシピや名所などの実世界の情報にコンテンツ中のオブジェクトに紐付け，そのオブジェクトをクリックすることで実世界の情報に直接アクセスすることが可能になるコミックビューワアプリケーションが提案されている [25]．

これらのようなコミック表現を利用したシステムやサービスが広がることで，コミック工学の裾野が広がり，研究分野としての意義が高まると考えている．

5 おわりに

本稿では，コミック工学の可能性について，これまでに行われている研究を概観しつつ検討した．現在，漫画やアニメといったサブカルチャーコンテンツは日本発信の新しい文化として国内外で大きな注目を集めており，政府もクールジャパン政策 [7] のひとつとして後押しをしている．現状では，これらの作品制作は漫画家やアニメ制作者の経験や感性に基づいて職人的に進められているが，ICT 技術がそこに持ち込まれることによって，(1) 新しいクリエータが知識や経験を効率的・効果的に習得できるようになる，(2) これまでの利用形態を超え，より柔軟で意図に沿ったコンテンツアクセスが可能になる，という効果が期待される．また，新しい流れとして，コミュニケーションのプラットフォームにコミックを用いて，複数人で同じ漫画を批評したり鑑賞したりするソーシャルリーディングを可能にする研究が登場している [32, 50]．このような，これまでにないコミックの利用も，今後電子化されたコミックコンテンツやその閲覧プラットフォームが成熟するにつれ増えるものと予想される．

本論で概観したように，コミック工学に関わる研究は多岐にわたる．コミック工学が狙うのは分野を跨った研究の創出・連携の促進である．加えて，サブカルチャーに関する人文科学的研究に新しい分析手段やツールを提供し，当該分野の発展に寄与することも期待できると考えている．そのため，異なる専門性や研究分野の研究メンバが相互に共用可能なコミックコーパスを整備すると同時に，コミック自体だけでなく Wikipedia やブログなどコミックに関連する WEB 上の情報も利用して，コミックコンテンツを計算機で取り扱える知識の形式に変換し，それを蓄える知識ベース（レポジトリ）を構築する必要がある．今後，こうした連携や環境の整備を通じて，コミックに関わる研究が発展することを期待する．

[7] 内閣府知的財産戦略本部: クールジャパン推進に関するアクションプラン, http://www.kantei.go.jp/jp/singi/titeki2/kettei/cjap.pdf (2013 年 4 月 18 日存在確認)

謝辞

本研究は科学研究費補助金挑戦的萌芽研究 (課題番号:15K12103) の支援を受けた．記して謝意を表す．

参考文献

[1] 新井俊宏, 松井勇佑, 相澤清晴: 漫画画像からの顔検出, 電子情報通信学会総合大会, p. 161 (2012).

[2] 藤本雄太, 宮下芳明: プレゼンとプレゼンの場をマンガ表現するインタラクティブシステム, *WISS2010*, pp. 23–28 (2010).

[3] 藤本雄太, 宮下芳明: マンガのコマ割り表現を用いたプレゼンテーションツール, 情処研報, Vol. 2010-HCI-139, No. 11, pp. 1–7 (2010).

[4] 福田美沙紀, 白水菜々重, 松下光範: コミックを対象とした質問応答技術のための基礎検討, 人工知能学会ことば工学研究会, Vol. 40, pp. 57–62 (2012).

[5] 何雯凌, 三原鉄也, 永森光晴, 杉本重雄: Wikipedia を利用したマンガの書誌データからのストーリー単位の抽出, 情処研報, Vol. 2014-CH-101, No. 9, pp. 1–8 (2014).

[6] 野中俊一郎, 沢野拓也, 羽田典久: コミックスキャン画像からの自動コマ検出を可能とする画像処理技術「GT-Scan」の開発, *FUJIFILM RESERCH & DEVELOPMENT*, No. 57, pp. 46–49 (2012).

[7] 石井大祐, 河村圭, 渡辺裕: コミックのコマ分割処理に関する一検討, 信学論, Vol. J90-D, No. 7, pp. 1667–1670 (2007).

[8] 石井大祐, 河村圭, 帆足啓一郎, 瀧嶋康弘, 渡辺裕: コミック画像におけるコマの角検出に関する一検討, 情処研報, Vol. 2010-AVM-69, No. 7, pp. 1–6 (2010).

[9] 石井大祐, 渡辺裕: マンガからの自動キャラクター位置検出に関する検討, 情処研報, Vol. 2012-AVM-76, No. 1, pp. 1–5 (2012).

[10] 石井大祐, 山崎太一, 渡辺裕: マンガ上のキャラクター識別に関する一検討, 情報処理学会第 75 回全国大会 (分冊 2), pp. 71–72 (2013).

[11] 石井大祐, 柳澤秀彰, 三原鉄也, 渡辺裕: マンガ画像解析に関する取り組み, HCG シンポジウム 2014, pp. 290–293 (2014).

[12] 伊東浩太, 松井勇佑, 山崎俊彦, 相澤清晴: 漫画のスクリーントーン除去に関する検討, HCG シンポジウム 2014, pp. 280–285 (2014).

[13] 金剛元, 三上浩司, 近藤邦雄, 金子満: オリジナルマンガ制作のための段階的なストーリー構成手法, 情報処理学会第 75 回全国大会 (分冊 1), pp. 705–706 (2010).

[14] 小林由佳, 石若裕子: 漫画設計支援システム POM, コンピュータソフトウェア, Vol. 25, No. 1, pp. 82–88 (2008).

[15] 向後智子, 向後千春: マンガによる表現が学習内容の理解と保持に及ぼす効果, 日本教育工学会論文誌, Vol. 22, No. 2, pp. 87–94 (1998).

[16] 小関悠, 角康之, 西田豊明, 間瀬健二: ぱらぱらマトリクス: 漫画技法を用いた映像を要約するシステム, インタラクション 2005 論文集, pp. 177–178 (2005).

[17] Kurlander, D., Skelly, T. and Salesin, D.: Comic Chat, *Proc. SIGGRAPH1996*, pp. 225–236 (1996).

[18] 松井勇佑, 相澤清晴, Jing, Y.: スケッチ入力を用いた漫画画像検索, HCG シンポジウム 2014, pp. 341-346 (2014).

[19] Matsui, Y., Aizawa, K. and Jing, Y.: Challenge for Manga Processing: Sketch-based Manga Retrieval, *Proc. MM'15*, pp. 661–664 (2015).

[20] Matsui, Y., Yamasaki, T. and Aizawa, K.: Interactive Manga retargeting, *ACM SIGGRAPH 2011 Posters*, Article No. 35 (2011).

[21] 松下光範, 今岡夏海: ディジタルコミック制作のための動的な音喩表現生成システム, 2011 年度人工知能学会全国大会, 1C1–OS4a–3 (2011).

[22] 松下光範: コミック工学の可能性, 第 2 回 ARG WEB インテリジェンスとインタラクション研究会, pp. 63–68 (2013).

[23] 三原鉄也, 永森光晴, 杉本重雄: マンガメタデータフレームワークに基づくディジタルマンガのアクセスと制作の支援—ディジタル環境におけるマンガのメタデータの有効性の考察—, 信学論, Vol. J98-A, No. 1, pp. 29–40 (2015).

[24] Mihara, T., Hagiwara, A., Nagamori, M. and Sugimoto, S.: A Manga Creator Support Tool Based on a Manga Production Process Model — Improving Productivity by Metadata, *iConference 2014 Proceedings*, pp. 959–963 (2014).

[25] 盛山将広, 朝田貫太, 内藤貴史, 松下光範: コミック閲覧時のユーザの興味をトリガとした情報アクセス手法の検討, HCG シンポジウム 2014, pp. 352–356 (2014).

[26] 野村聡美, 両角彩子, 永森光晴, 杉本重雄: マンガのためのメタデータモデルを目指したマンガのアーキテクチャ分析, 第 36 回ディジタル図書館ワークショップ, pp. 3–14 (2009).

[27] Morozumi, A., Nomura, S., Nagamori, M. and Sugimoto, S.: Metadata Framework for Manga: A Multi-paradigm Metadata Description Frameworkfor Digital Comics, *Proc. DC-2009*, pp. 61–70 (2009).

[28] Murakami, H., Kyogoku, R. and Ueda, H.: Creating Character Connections from Manga, *Proc. ICAART2011*, Vol. 1, pp. 677–680 (2011).

[29] Nara, Y., Kunitomi, G., Koide, Y., Fujimura, W. and Shirai, A.: Manga Generator, *Laval Virtual VRIC 2013*, Article No. 29-7 (2013).

[30] 夏目房之介: マンガの力 成熟する戦後マンガ, 晶文社 (1999).

[31] 根来美貴, 曽我真人, 瀧寛和: 読者の視点移動を考慮した初心者向けマンガ作成支援システムの設計, インタラクション 2013, 2EXB–40 (2013).

[32] 落合香織, 三原鉄也, 永森光晴, 杉本重雄: マンガ Path 式を利用したソーシャル Web 上におけるディジタルマンガのアノテーション共有, 第 11 回情報科学技術フォーラム, pp. 327–330 (2012).

[33] Rigaud, C., Guérin, C., Karatzas, D., Burie, J. C. and Ogier, J. M.: Knowledge-Driven Understanding of Images in Comic Books, *IJDAR*, Vol. 18, Issue 3, pp. 199-221 (2015).

[34] Rigaud, C., Karatzas, D., Burie, J. C. and Ogier, J. M.: Speech Balloon Contour Classification in Comics, *Proc. GREC2013* (2013).

[35] Rigaud, C., Karatzas, D., de Weijer, J. V., Burie, J. C. and Ogier, J. M.: An Active Contour Model for Speech Balloon Detection in Comics, *Proc. ICDAR2013*, pp. 1240–1244 (2013).

[36] 坂本竜基, 角康之, 中尾恵子, 間瀬健二, 國藤進: コミックダイアリ: マンガ表現を利用した経験や興味の伝達支援, 情処論, Vol. 43, No. 12, pp. 3582–3595 (2002).

[37] Sumi, Y., Sakamoto, R., Nakao, K. and Mase, K.: ComicDiary : Representing Individual Experiences in a Comics Style, *Proc. UbiComp 2002*, pp. 16–32 (2002).

[38] Sun, W. and Kise, K.: Similar Manga Retrieval Using Visual Vocabulary Based on Regions of Interest, *Proc. ICDAR2011*, pp. 1075–1079 (2011).

[39] Sun, W. and Kise, K.: Detection of Extract and Similar Partial Copies for Copyright Protection of Manga, *IJDAR*, Vol. 16, Issue 4, pp. 331–349 (2013).

[40] 高月義照: マンガにおける表現技法の進化 –何がマンガを文芸に成長させたのか–, 東海大学紀要, Vol. 20, pp. 53–75 (2010).

[41] 玉田圭作: 教育とマンガに関する研究の全体像: 既存の研究と最近の動向から, 哲學, No. 123, pp. 207–228 (2010).

[42] 田中孝昌, 外山史, 宮道壽一, 東海林健二: マンガ画像の吹き出し検出と分類, 映像情報メディア学会誌, Vol. 64, No. 12, pp. 1933-1939 (2010).

[43] Tanaka, T., Shoji, K., Toyama, F. and Miyamichi, J.: Layout Analysis of Tree-Structured Scene Frames in Comic Images, *Proc. IJCAI'07*, pp. 2885–2890 (2007).

[44] 谷悠, 白水菜々重, 松下光範: コミックコンテンツにおける登場キャラクター抽出のための基礎検討, 情報処理学会第 75 回全国大会 (分冊 4), pp. 889–890 (2012).

[45] 谷本奈穂: 『スラムダンク』の「魅力」—読者解釈と構造分析, メディア文化を社会学する 歴史・ジェンダー・ナショナリティ (谷本奈穂, 高井昌吏 (編)), 世界思想社, pp. 266–292 (2009).

[46] 寺島亜耶香, 上間大生, 松下光範: 効果線の描画に着目した動的音喩の付与手法, 2012 年度人工知能学会全国大会, 1M2–OS–8b–3 (2012).

[47] 内橋真吾: ビデオ・マンガ要約を用いたインタラクティブなビデオ閲覧, インタラクション 2001, pp. 31–32 (2001).

[48] Uchihashi, S., Foote, J., Girgensohn, A. and Boreczky, J.: Video Manga: Generating Semantically Meaningful Video Summaries, *Proc. MM'99*, pp. 383–392 (1999).

[49] 山田雅之, 鈴木茂樹, ラフマットブディアルト, 遠藤守, 宮崎慎也: 携帯電話を利用したコミックの閲覧システムとその評価, 芸術科学会論文誌, Vol. 3, No. 2, pp. 149–158 (2004).

[50] 山西良典, 杉原健一郎, 井上林太郎, 松下光範: ソーシャルデータを用いたコミックからの感性的ハイライトの抽出, 日本感性工学会論文誌, Vol. 14, No. 1, pp. 155–162 (2015).

[51] 山下諒, 陸鑫一, 松下光範: コミックを対象とした質問応答システムのための質問タイプ分類の検討, 第 7 回インタラクティブ情報アクセスと可視化マイニング研究会, pp. 28–32 (2014).

[52] 柳澤秀彰, 石井大祐, 陳明, 渡辺裕: 漫画画像からの顔検出におけるパーツ特徴量の一検討, 映像情報メディア学会年次大会, 17-9 (2014).

[53] 吉澤勇気, 坂本雄児: 漫画的似顔絵における髪型の表現・協調についての考察, 信学技報, Vol. 105, No. 609, pp. 121–126 (2006).

人工知能学会　インタラクティブ
情報アクセスと可視化マイニング研究会(第 11 回)
SIG-AM-11-04

ソフトバンクが提供する対話型 FAQ システム：APTWARE

石川　誠一[1]　尾曲　博隆[1]

[1] ソフトバンク株式会社 ボイスプラットフォーム開発部

Abstract: 対話型 FAQ システム「APTWARE」は，ソフトバンクが開発した一問一答型の FAQ システムである．ユーザーは自然な会話調の文章でシステムに質問することによって必要な情報を引き出せることが可能となる．ここでは APTWARE の特徴，検索の仕組みについて説明する．

はじめに

ユーザーが製品などに疑問を持った場合，Web サイトやマニュアルを確認し，FAQ 検索を経て，それでも解決しない場合はコンタクトセンターに問い合わせる流れが多い．そのため FAQ システムでユーザーの自己解決率を高めることで，顧客満足度を向上させることができる．

しかしコールセンター白書のレポート[１]によると，約 80%の人が，コールセンターに電話する前にサポート用のサイトを参照・検索したが「見たが解決しなかった」と回答している(図 1)．ここから，FAQ システムに必要な情報が登録・更新されていない，もしくは登録されていても，どこに情報が載っているか分からず，探し出せないことがわかる．

またコンタクトセンターに電話した場合の「問題や疑問が解決するまでのプロセス」として約 73%の人が「1 回の電話で解決した」と回答している．ここからコンタクトセンターは解決のための「回答」を持っており，対話からユーザーの聞きたい内容を特定していることがわかる．

以上から APTWARE の設計にあたって，1) FAQ 作成担当者が，オペレータの持つ知識をいかに容易に登録・更新できるか，2) 使い方を気にすることなく，ユーザーを自然な対話で回答へ誘導できるかの 2 点を重視し，開発を進めた．

APTWARE とは

APTWARE は，ソフトバンクが提供する一問一答の FAQ システムである[２]．ユーザーは Web ブラウザなどから自然な文章を入力するだけで，回答を得る．APTWARE はソフトバンク社内の ERP システムにおける「コンシェルジュ」機能として，2014 年 10 月から本格的に利用されている(図 2)．

APTWARE は，目的があいまいな質問に対して聞き返す回答を返し，一問一答をしながら，まるで対話をしているかのように回答候補を絞込む仕組みを持つ．これによりユーザーの目的を明確にし，満足度の高い回答を表示することが可能となる．

また技術者でない担当者でも QA データのメンテナンスが可能とする観点から，機械学習などの「AI 的な手法」を利用せず「キーワード・マッチング」で回答を引き当てるアプローチをとる (図 3)．検索がヒットしない理由が一目でわかることや，事前の学習データを準備しなくても FAQ システムを立ち上げられるようにするためである．そこでキーワード自動抽出や自動テストツールなど，回答を引き当てるためのチューニング作業を支援する機能を充実させている．

ここでは APTWARE の主な特徴，検索の仕組みについて説明する．

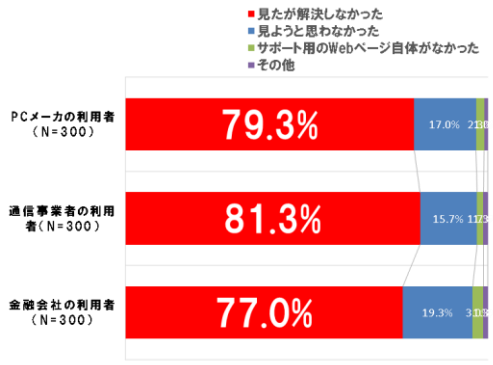


図 1 FAQ を事前に確認したか [1]

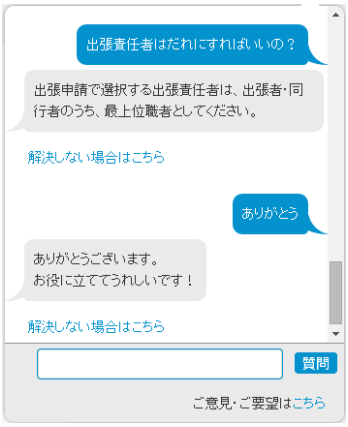


図 2 社内利用例

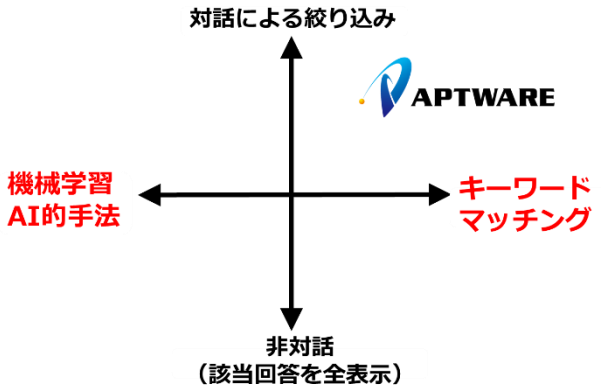


図 3　APTWARE のポジショニング

APTWARE の特徴

聞き返し処理と状態保持

例えば図 4 のように,「スマホについて教えて」という質問が来た場合,「スマホの何が知りたいですか？」と回答を返し，人間同士の自然な会話のように「聞き返し」を行う．この機能を使い，お客様の質問の真の意図を正確に判断することが可能とする．

また, APTWARE は直前の話題を記憶しておき「状態保持」を行うことができる. このように APTWARE は「スマホ」という話題を記憶できるため，「メールの操作について教えて」と問い合せた場合には,「スマホ」の「メールの操作について」を回答することができる.

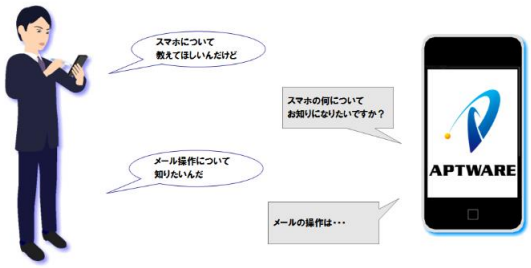


図 4　聞き返し機能のイメージ

大量 QA データに対する高い引当実績

APTWARE は，ソフトバンクの FAQ データ 4,045 件に対してテストを行った結果，他社を大きく上回る結果を記録した(図 5).

116 件の質問を 3 つのシステムに対して同様にテストを行ったところ, APTWARE 以外のシステムは，半分以上の質問が不正解だったことに対し，APTWARE はおよそ 85%の質問に対し正しい回答を得た*．このように APTWARE が持つ仕組みを利用することで, 4,000 件ものの大量 QA データに対しても，高い正答率を得ることができる.

*質問に対して模範回答を用意し, それらが検索上位 1 位になった割合を表す．また，このテストは 2013 年 3 月に実施した結果である.

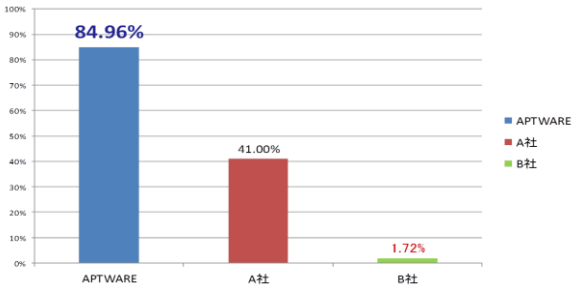


図 5 高い引当実績

APTWARE の機能

APTWARE の概要を図 6 に示す．FAQ 作成担当者は QA データおよび検索用キーワードを組み合わせて格納する．検索ユーザーが自然な文章で質問すると，APTWARE は質問に含まれる単語(もしくはフレーズ)を用いて回答を引き当てる．マルチテナント対応をしているため，社外向け FAQ や，社内の営業システム向け，経理システム向けなど複数の Web サイト用の FAQ を 1 つのシステムで対応できる.

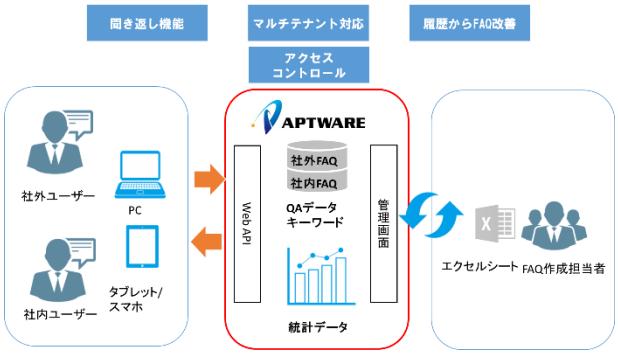


図 6　APTWARE 概要

図 7 機能概要

一般的に FAQ システムには，QA データの作成や管理を支援する機能，正答率の向上を支援する機能，および QA データの改善を支援する機能が必要となる(図 7).

以下では，他システムと比較して特徴的な機能について説明する.

QA データ作成支援

1.　データ作成 Excel シート

管理画面だけでなく，技術者ではない担当者でも普段から使い慣れた Excel を使って QA データやキーワードを登録することが可能である．既存の QA データをそのまま貼り付けてインポートすればそのまま使用できる．そのため QA データの追加／修正に IT 技術者が不要になり，より迅速な運用が可能になる.

2.　キーワード自動抽出

用意した QA データから未登録のキーワード候補を抽出・提示する．QA データ作成担当者は提示された候補を選択するだけでキーワードの登録ができる.

キーワードは，引き当てたい回答が一意に決まるような，回答を特徴づける単語(もしくはフレーズ)であることが望ましい．社内業務や商品の QA データを特徴づける言葉はその企業，業種特有の単語であり，名詞を合わせた複合名詞であることが多い[３].

例えば,「光回線開通工事の工事代金について」という回答を登録する場合，キーワードとして「光」,「回線」,「工事」のような一般的な単体の単語ではなく「光回線開通工事」が望ましい．より回答を特徴づけるキーワードとして「光回線開通工事の工事代金」のようなフレーズを使うと，さらに特徴づけができ、一意に特定しやすくなる.

APTWARE では「回線」のような単体名詞の抽出だけでなく,「光回線開通工事」のような複合名詞や,「光回線開通工事の工事代金」,「SB 光を利用」のようなフレーズをキーワード候補として提示できる.

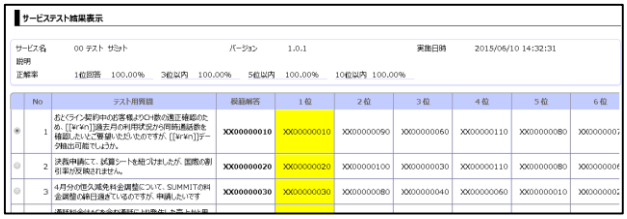


図 8 自動テストツール

3.　自動テストツール

QA データ作成担当者が事前に用意した複数の想定質問と模範回答の組み合わせを使い，APTWARE は正答率を一括で算出する．また正答率だけでなく，模範回答の順位，回答の引き当てに使われたキーワードなどを表示して，不正解となった理由を提示する(図 8).

4.　類義語・表現のゆらぎ対応

QA データには社内用語やサービス・商品名が頻繁に記述されているが，それらを検索ユーザーが正しく質問してくれるとは限らない．そのため，質問された単語そのものでは検索がうまくいかないことが多い.

例) 質問での表現　　：請求書の精算
　　QA データでの表現　：請求書払

APTWARE では，質問に含まれる単語を「請求書の精算=>請求書払」のように片方向に展開する「正規化辞書」を作成することで表現のゆらぎに対応する.

このような表記ゆれや類義語に関する課題に対して，類義語辞書を用いて対応を自動化している FAQ システムも存在する．これらは，検索がヒットしない「0 件ヒット」対策のために行われることが多い．類義語のいずれかが含まれる文章を検索して検索漏れをなくそうとするアプローチである.

しかし類義語辞典に登録している単語は「一般的な単語」であることが多く，社内用語やサービス名などの「専門用語」をそもそも含まない．類義語辞典を利用することで，専門用語に対してではなく，それ以外の「一般的な用語」を展開してしまい，検索結果が大量に抽出される可能性がある．そのため社内利用などの「限定された領域」での FAQ システムにおいて，検索結果を上位に上げるチューニングには適さないと筆者らは考えている.

5.　シミュレーション機能

正答率を高めるために「チューニング」が必要だが，これまでの FAQ システムはリアルタイムに成果率の変動を確認することが難しかった．APTWARE の「シミュレーション機能」は，チューニングをし

ながら，正答率を確認できる機能を提供する．精度を確認しながら作業ができるため，FAQ 作成担当者は手戻りなく，「自信」を持って作業が進めることが可能である(図 9).

図 9 シミュレーション機能

正答率向上支援

6. 聞き返しと状態保持

回答が特定できない曖昧な質問に対しては「何が知りたいですか？ 次の中から選んでください」のように聞き返す回答を返すことで APTWARE では対話を実現する．QA データをカテゴライズ(分類分け)して，質問に含まれるキーワードから条件分岐しながら，検索対象となる回答文を絞り込んでいく．この作業をすべて Excel ツールだけで実現できる．

7. 質問のサジェスト

ユーザーは，自身が欲しい回答を得るために，どのような質問をすればいいのか分からないことがある．そこで最近の多くの検索システムでは，質問をリアルタイムに補完してくれる機能を持つ(図 10).

APTWARE ではサジェスト用のデータを用意することなく，登録する QA データから質問の候補を提示する．APTWARE のサジェスト機能は次の特徴を持つ.

- ひらがな，カタカナ，漢字，ローマ字など入力される文字種に関わらず，候補を提示
- 入力した文字を，質問文の中でハイライト
- 質問の文字列のみだけでなく，その回答に紐づくキーワードも考慮して，候補を提示
- 先頭一致だけでなく，部分一致で候補を提示

なお，サジェスト機能の実装はアティリカ株式会社が開発したサジェストエンジン Akahai を用いた[4].

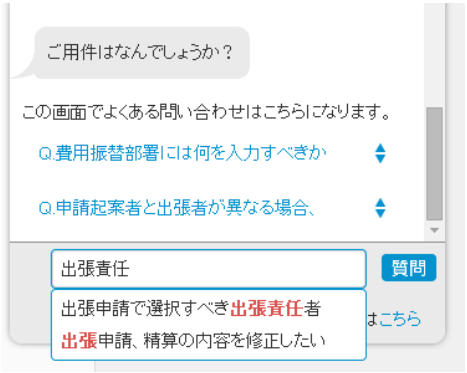


図 10 サジェスト機能

QA データ改善支援

8. レポート・統計情報

ユーザーのよく質問されるキーワードのランキングや，質問の履歴など利用状況をグラフで把握することができる．またそのデータを CSV ファイルで出力も可能である．公開中の回答が足りているか，よく質問されるキーワードなどから QA データの改善に反映させることができる(図 11).

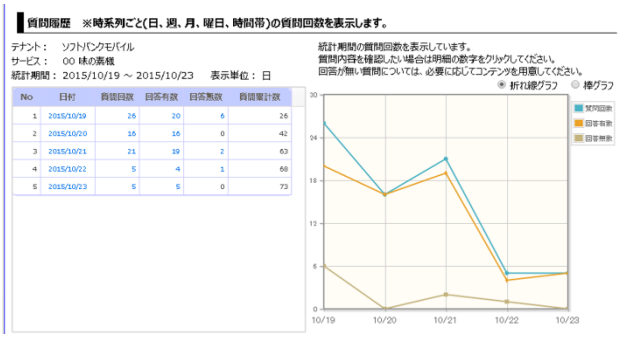


図 11 統計情報の一例

APTWARE の検索技術

APTWARE は検索の精度を向上させるために 2 種類の検索技術を使う．これらの実装には，検索のコアエンジンにオープンソースソフトウェアの Apache Solr（以下，Solr）を用いている． Solr は，実績のある全文検索ライブラリ Apache Lucene を用いた転置索引（インデックス）方式の全文検索エンジンである[５]．また形態素解析器としてアティリカ株式会社が開発した Kuromoji[６]を用いる．

図 12 に APTWARE の検索の流れを示す．大きく 3 つの部分で構成されており，それぞれ次の機能を持つ．

１．受付・前処理部
入力された質問を形態素解析し，同義語や表記の揺れに対して，キーワードの正規化を行う．また不要語(ストップワード)の削除など，検索前処理を行う．

２．検索部
登録済みの QA データに対して全文検索を行う「全文検索」と，QA データに紐づくキーワードを検索する「キーワード検索」を実行する．それぞれの検索手法を用いて，適合度(スコア)を算出して検索結果を出力する．なお，全文検索では Solr が算出したスコアをそのまま用いている．一方，キーワード検索では Solr のロジックの一部を改良して，APTWARE 独自のスコア計算を実装している．

３．評価・応答部
APTWARE 独自のスコア正規化手法を用いて，それぞれの検索結果のスコアを正規化し，順序付けする．

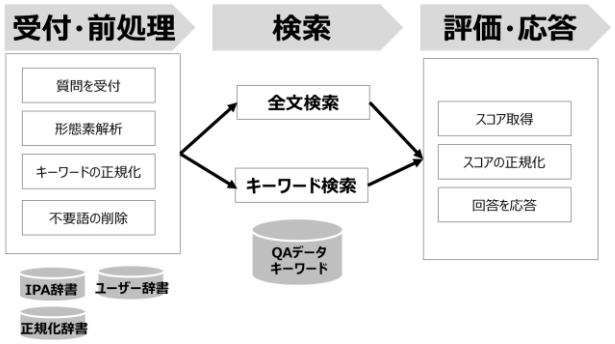


図 12 APTWARE 検索の概要

今後の展開

APTWARE は，社内 FAQ や商品サービスサイトでの活用のみならず，対話のシナリオを工夫することによって，デジタルサイネージやロボットなどの様々なクライアント・インタフェースでも活用できると考えている．近年の訪日外国人向けのインバウンド・ソリューションの注目の高さを受け，多言語対応にも取り組んでいくことを予定している．2015 年 11 月に米語対応 ベータ版をリリース予定であり，今後は北京語，韓国語の対応も検討していく．

参考文献

[１] コールセンター白書 2013 コールセンター利用者調査 コンピューターテレフォニー編集部・編
[２] https://rizbell.jp/
[３] 中川裕志, 森辰則, 湯本紘彰: "出現頻度と連接頻度に基づく専門用語抽出",自然言語処理, Vol.10 No.1, pp. 27-45, (2003)
[４] http://www.atilika.com/en/products/akahai.html
[５] http://lucene.apache.org/solr/
[６] http://www.atilika.com/en/products/kuromoji.html

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第11回)
SIG-AM-11-05

情報アクセスにおける受動性と能動性: 音声対話によるニュース記事アクセス

Intentionality in Information Access Behavior: A Spoken Dialogue System for Interactive Access to News Articles

林 良彦[1*] 藤江 真也[2,1] 福岡 維新[1] 高津 弘明[1] 小林 哲則[1]
Yoshihiko Hayashi[1] Shinya Fujie[2,1] Ishin Fukuoka[1] Hiroaki Takatsu[1] Tetsunori Kobayashi[1]

[1] 早稲田大学 [2] 千葉工業大学
[1] Waseda University [2] Chiba Institute of Technology

Abstract: Passive information consumption would an adequate type of information behavior for receiving the content of, for example, a news article. It may however be boring in many cases and even painful in some cases, especially when the information content is delivered by employing speech media. The user of a speech-based information delivery system, for example a text-to-speech system, usually cannot interrupt the ongoing information flow, inhibiting her/him to confirm some part of the content, or to pose an inquiry for further information seeking. We thus argue that spoken dialogue is a suitable media for enabling interactive information access that coordinate passive information consumption and active information seeking. This paper shows that a carefully designed spoken dialog system could remedy these undesirable situations, and further enables an enjoyable conversation with the users. The key technologies to realize such an attractive speech-based interactive information access system are: (1) pre-compilation of a dialog plan based on the analysis of a source content, and (2) the dynamic recognition of user's state of understanding and interests during the course of conversation. This paper illustrates technical views to implement these functionalities, and discusses a dialog example to exemplify our approach.

1 はじめに

人間の情報に関する行動 (information behavior) のうち，情報獲得・収集 (information acquisition) に関する行動は大きく，意図的な情報探索 (intentional information seeking) と，意図性のない受動的な情報行動 (unintentional passive information behaviors) に分けられるとされ [1]，情報学の分野では，主に前者を導く動機や状況に関するモデルの研究が行われてきた [3]．

コンピュータサイエンスの領域においても，その焦点はもちろん前者にあり，情報検索 (information retrieval)，あるいは少し広い概念としての情報アクセス (information access) のシステムについて，様々な観点からの研究開発が活発に行われてきた．

以上の研究状況の背景を推察するに，情報遭遇 (information encountering) などの受動的な情報行動は，主として偶発的な状況によることから，研究的な要素に乏しいと考えられてきたのではないかと考えられる．しかしながら，我々の日常の情報行動の実際をみれば，両者の区別は必ずしも明白ではなく，むしろ，これらの情報行動の状態を自由に遷移する過程であると考えるのが妥当であろう [2]．

さて本研究では，音声対話によるニュース記事アクセスシステムをとりあげる．ユーザ側からみれば，ニュースに関する情報を音声メディアを用いて獲得し，あるいは，消費する**情報アクセス**システムであるが，システム側の観点から言えば，音声メディアを用いて，ユーザに伝えたい・伝えるべきニュースを伝達するという**情報伝達**システムである．

システムが一方的に記事の内容を読み上げるとすれば，ユーザは読み上げ音声を黙って聞き続ける必要がある．ユーザにとって内容的に冗長である可能性もあるし，そもそも記事の内容に興味がないことに気づく場合もあるだろう．

このような情報提示システムの対極に，記事に関する簡単な内容 (例えば記事の見出し) を与えた後に，ユーザからの質問を一問一答形式で受け付けるモードに移行する質問応答型のシステムが考えられる．このよう

*連絡先: 早稲田大学理工学術院 実体情報学博士プログラム
〒 169-0072 新宿区大久保 2-4-12 ラムダックスビル 3F
E-mail: yshk.hayashi@aoni.waseda.jp

S_1: 羽生結弦選手が
U_1: うん
S_2: 国際大会を欠場することになったよ
U_2: え？
S_3: 欠場するんだ、腰の痛みのためだって
U_3: 腰の痛み... って？
S_4: 練習中に腰を痛めたということなんだ
...

図 1: 想定する対話の断片例．“S_n:” はシステム発話，“U_n:” はユーザ発話を表す．

なシステムのユーザは，適度な量の情報を得るまで，質問を発し続けることが必要になる．

先の情報獲得における意図性の議論からすれば，ユーザは両者のモードを，その「状況」に応じて，しかも簡単な手段によって，行き来できることが望まれる．そこで，本研究が想定するような対話を簡単化した断片の例を図 1 に示す．システムから伝達された情報に対してユーザは，必ずしも言語的ではない即応的な反応 (U_1=肯定的, U_2=疑問) によって理解状況を示したり，さらに対話の過程で生じた情報要求 (国際大会を欠場する理由) をある程度明確な言語表現を用いて示し (U_3=問い返し) たりする．システムは，必要に応じてこれらのユーザの状況を推定し，適切と思われる情報を付加しながら応答を返す (S_4)．

本研究の前提，あるいは，主張は，このような**受動的な情報獲得を主体としつつもインタラクティブ性を要する・有する情報行動の支援形態として，音声対話が適している**という点にある．

本稿の以下では，まず情報学における関連研究を参照しながら，上記の議論を補強し，本研究のスタンスを明確化する．次に，現在開発中のニュース記事を対象とする音声対話システムについて述べ，最後に今後の課題や研究の方向性について論じる．なお，音声対話システムに関する内容は，当研究グループにおける既発表 [17] の内容によっている．

2 情報行動における受動性と能動性

これまでに提案されている情報行動の分類には様々なものがあるが，Erdelez による分類 [2] を図 2 に示す．ここでは，非意図的な情報行動は機会主義的情報獲得 (Opportunistic Acquisition of Information: OAI) と呼ばれており，その主な下位分類として，情報遭遇 (information encountering) が位置づけられている．

Erdelez はさらに，(1) 気付き (noticing); (2) 停止 (stopping); (3) 検討 (examining); (4) 獲得 (capturing); (5) 復帰 (returning)．の各段階からなる情報遭遇における機能モデル (図 3) を提示した．

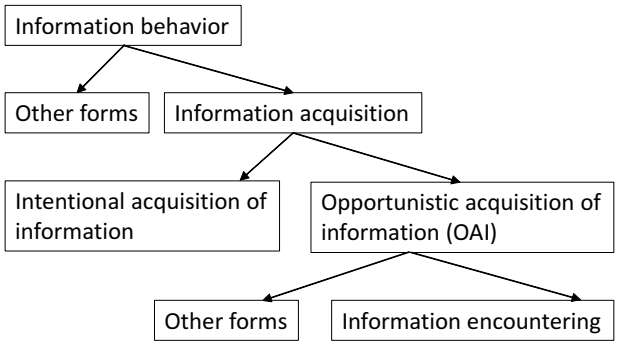


図 2: Erdelez による情報行動の分類 ([2] より作図).

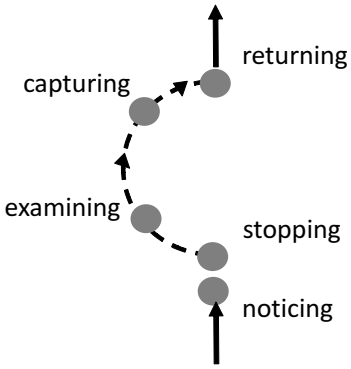


図 3: Erdelez による情報遭遇の機能モデル．実線部が foreground interest，点線部が background interest を表す．([2] より作図).

このモデルでは，ユーザは彼/彼女の主要な関心 (foreground interest) に関わる能動的な情報探索タスクを実行していることが仮定されているが，重要なことは，ユーザはこの過程の中で関連する関心 (background interest) に気づき，foreground の情報探索タスクを一旦停止したうえで，background に関する情報行動 (検討・獲得) を行い，その後に foreground の情報探索タスクに復帰するという点である．

以下，図 1 に示した対話例をこのモデルをと照らし合わせて考えてみる．この対話における「主要な関心」は，対象のニュース記事により定まる「羽生結弦選手の国際大会欠場」にある．これが主要な関心となる契機については問わない[1]が，システムが記事内容の伝達を行っている間は，ユーザは基本的には受動的な情報消費のモードにある．

本研究では，この対話でのユーザによる U_3 の発

[1]すなわち，システム側が勝手に見繕った記事 (受動的なあるいは機会主義的な情報獲得) かもしれないし，ユーザによるある種の情報検索の結果として選択されたもの (意図的な情報探索) であってもよい．

話 (「腰の痛み... って？」) の背景には，図 3 のモデルにおける noticing に相当する過程があると考える．すなわち，foreground に対する情報行動の過程の中で background に対する情報行動へのシフトが起こっている．ただし，図 3 の Erdelez のモデルでは，能動的で意図的な情報探索の過程において偶発的な情報遭遇が生じているのに対し，図 1 に示す対話においては，ある程度の明示性のある言語表現 (「腰の痛み... って？」) によって，モードのシフトが起こっている点が異なる．しかしながら，background に関する情報行動のモードが一段落した後は，foreground の情報行動のモードに復帰することは共通している．

以上にみたような，ユーザにとって自然で効率の良い情報行動は，受動的な情報消費から，意図的な情報要求による情報探索までの情報行動のモードをその「状況」に応じて，しかも簡単な手段によって，行き来することにより達成される．本研究では，このような情報アクセス・情報伝達を実現するために音声対話が自然で効率よいメディアであるという前提にたち，次節で述べるような音声対話によるニュース記事アクセスシステムを提案する．

3 音声対話によるニュース記事アクセスシステム

3.1 システムの要件

理想的な対話システムの実現へ向けては様々な課題が存在するが，少なくとも以下のような要件を考慮する必要がある．

伝えるべきニュース記事の選択: どのようなニューストピックを対象とするかの決定は，本研究の範囲外とする．すなわち,「本日の重大ニュース」でも,「本日のおすすめニュース」でも，ユーザによる情報検索の結果として選択されたものでも良い．

伝えるべき内容の選択と構成: システムは，対象とするニュース記事 (群) が与えられたとき，最低限どのような内容を伝えるべきかを決定する必要がある．ここでいう最低限伝えるべき内容とは，対話においてユーザが完全に受動的である場合においても，システム側がとにかく伝達しようとする骨格的な情報内容である．要約と言っても良いだろう．このような要約に相当する情報内容を補足する補助的な内容は，ユーザからの具体的な情報要求に基づいて提示することになる．

ユーザの状況の把握と対応: ユーザは，システムの発話に対する自身の理解状況や伝達内容への興味の状況を反映して，肯定的・否定的な短い即応的な発話 (以下，即応的情報反応と呼ぶ) や，もう少し明示的な情報要求を発することが想定・期待されている．したがって，システムはこれらのユーザの反応・発話からユーザの状況を適切に把握し，さらにはそれに応じた応答を返す必要がある．

リズムのある対話の実現: リズムのある対話を実現するためには，即応的情報反応を含むユーザの短い発話に対して，システムは素早く応答できることが望まれる．よって，対話の過程で提示されるユーザからの情報要求をある程度見越して,「こう聞かれたら，こう答える」ということを定めておくことが必要となる．

3.2 システムの構成

以上のような要件を (ある程度) 満たすものとして，我々が提案する音声対話システムの構成を図 4 に示す．システムは大きく分けて，ニュース記事をもとに対話に利用する発話計画を生成する事前処理部と，ユーザを相手に対話を行う対話システム部の二つから構成される．

事前処理部は，インターネットからの取得などによって与えられるニュース記事を解析して構造化する構造解析部と，その結果をもとにユーザの反応を織り込んだシステム発話計画を生成する計画生成部からなる．発話計画は，記事における主要な内容を伝達するための主計画と，それを補足する補助的な内容を伝達するための副計画からなる[2]．構造解析部については **3.4** で，計画生成部については **3.5** でそれぞれ詳細を述べる．生成された発話計画は，発話計画データベースに保存される．

対話システム部は，発話計画を読み込み，それに従って対話を進める．音声認識器は，ユーザの短い反応を認識する．対話制御部は，発話計画に従ってシステムの発話内容を含む発話文を音声合成器に出力する．また，システムの発話に対するユーザの反応に応じて，事前に生成した発話計画に従って発話内容の調整を行う．音声合成器は，対話制御部から生成された発話文を音声に変換してユーザに提示する．

本システムにおいて，音声認識器は ATR-Trek 製のものを使用している．また，音声合成器は Open JTalk[3] を基に開発したものを用いている．

3.3 発話計画とユーザの反応・応答

ここでは，システム側の視点から発話計画とユーザの応答の関係について述べる．本システムでは，想定

[2] これらはさらにネストしていてもよい．
[3] http://open-jtalk.sp.nitech.ac.jp/

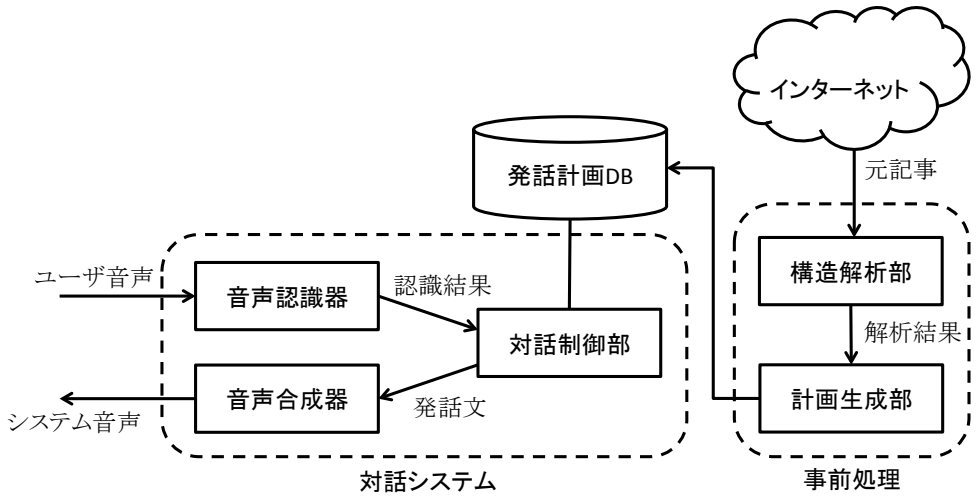


図 4: システム構成

されるユーザの反応を織り込んだ形で発話計画を事前に生成しておき，実際の対話時にはそれに従うことで効率的な情報伝達を実現する．ここではその発話計画がどのようなデータ構造を持つか，また，それに従って対話制御部がどのように対話を進めていくかについて述べる．

まず，発話計画のデータ構造を図 5 を用いて説明する．図に示した通り，発話計画は状態遷移構造として表される．各アークに示された $U_{i,j}$ はユーザ発話を表し，$S_{i,j}$ は，システム発話を表す．ユーザ発話は音声認識部へ送られ，システム発話はそのまま音声合成部に渡される．

ユーザ発話 $U_{i,j}$ は即応的情報反応を含む短い応答を想定しており，下記の 2 つのカテゴリのうちいずれかに分類される．

肯定的応答 (ACK) 「うん」「へー」といった相槌など，システムの発話進行に肯定的な態度を表す反応．システムの発話の一部を下がり口調のイントネーションにより反復する場合を含む

否定的応答 (NACK) 「え？」といった，システムの発話進行に否定的な態度を表す反応．上がり口調のイントネーションによるシステム発話の一部の反復を含む

すでに述べたように，システム発話の一部を反復することは，ある程度明確な情報要求がユーザにおいて発現している状況，すなわち，情報行動のモードにシフトが起こっている状況を示唆する．また，否定的な応答は，システム発話のいずれかの部分が聞き取れなかった状況，あるいは，発話された内容に対して明示的でない情報要求が生じている状況を表していると考えられる．

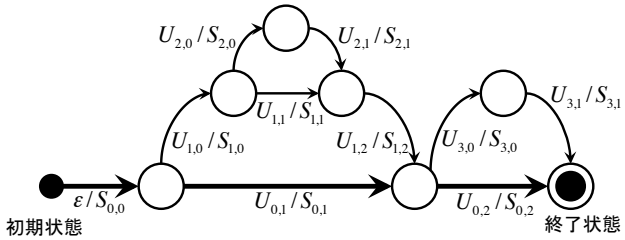


図 5: 発話計画の構造．$U_{i,j}$ はユーザの反応，$S_{i,j}$ はシステム発話を表す．

ϵ はユーザの反応を見ずに状態が遷移することを表す．図中で太いアークにより表されている部分を主計画と呼ぶ．主計画は，システムが最低限伝えるべき，記事の骨格をなす内容情報を表す．ユーザが受動的な態度 (すなわち，反応がないか，ACK のみを示す) を取り続ける限り，システムは主計画に従って淡々と，この内容情報の伝達を進めることになる．

本システムにおける対話制御は非常にシンプルで，アークに与えられたシステム発話 $S_{i,j}$ の内容を音声合成器に出力した上で状態を遷移させる．各状態ではユーザの反応を待ち，その内容に従って次の遷移を行う．例えば主計画上にいる際は，特にユーザの反応が得られなくてもそのまま発話を続けることが好ましいと思われるので，図中，$U_{0,1}$ や $U_{0,2}$ などは，特に反応が無くても時間経過 (例えば 0.6 秒無反応で経過) によって遷移させる．

否定的応答が認識された場合 (例えば $U_{3,0}$ が得られた場合) は，その直前のシステム発話 ($S_{0,1}$，あるいは $S_{1,2}$) を補足する副計画に従い，情報を提示する発話 ($S_{3,0}$) を生成する．

このように，想定されるユーザの反応を織り込んだ計画によって対話制御を行うことで，素早く効率的に

ソチオリンピック、／フィギュアスケート男子の／金メダリスト、／羽生結弦選手が／腰の／痛みの／ため、／今シーズンの／初戦と／して、／来月フィンランドで／出場を／予定していた／国際大会を／欠場することになりました。

図 6: 文節単位に分割された文

ユーザの理解や知識に見合った情報を伝達することが可能となる．

なお，ユーザの反応として，上記のカテゴリに含まれない，より明示的な質問により情報要求が提示された場合は，例外処理として質問応答型の対話制御に一時的に切り替えることを考えているが，その詳細は現在検討中である．

3.4 構造解析部

構造解析部では，対象とするニュース記事をもとに，発話計画を生成するために必要な情報を構造化する．すでに複数記事を対象とする場合の検討を進めているが，本稿では単一の記事を対象とする．

構造解析の目的は，ニュース記事が持つ情報をもとに発話計画を立てるための情報を抽出することにある．すでに示したように，発話計画はニュースの要点を伝える主計画とそれを補う副計画からなるので，構造解析部では主計画に含めるべき情報 (以降，主情報) と，周辺情報と主情報の関係性を抽出するという課題がある．

構造化: 構造解析は，文節単位の係り受け関係をもとに行う．そのため，まず文を文節単位に分割し，係り受け解析を行う．本研究では，形態素解析と係り受け解析に，Juman[4]，KNP[5] をそれぞれ利用した．

例として，ウェブニュース記事[6]の一文を文節に分割した例を図 6 に示す． さらに，係り受け解析に基づいて，文節をノードとする依存構造木を作成する．図 7 に，図 6 をもとに生成した依存構造木を示す．図中では省略しているが，各アークは係り受けの関係属性を保持し，各ノードは当該の文節の文法的情報を持つ．例えば,「羽生結弦選手が」に対応するノードは,「人物，主題」という情報を持つ．

主情報の抽出: 図 7 中，太枠で囲まれた文節がこの文における主情報であり，主発話計画を構成する．主情報として，まず，対象文の主辞となる文末の述語文節と，それに対する必須格要素となる文節を選択する．次に

[4]http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[5]http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[6]NHK NEWSWEB，http://www3.nhk.or.jp/news/

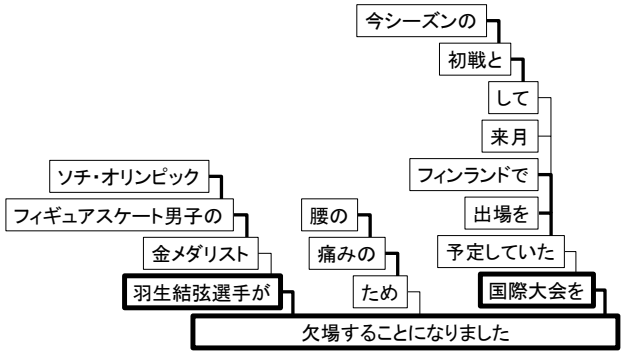


図 7: 図 6 の文に対する依存構造木

これらの文節に対して，次に説明する省略不可避性を有する文節を抽出する．図中では，太線によるエッジが省略不可避性を表す．

省略不可避性とは,「当該文節の係り先の文節が発話される場合は，当該文節が省略されてはならないこと」を表す．例えば「ため」という形式名詞からなる文節は，それだけでは特定の意味を持たないため,「痛みの」を省略できない．また,「痛みの」に対しての「腰の」が省略不可避であるかは微妙であるが，本研究では，野本による統計的な文圧縮の研究 [15] における依存構造木の「刈り込み」に準じた方法により，省略不可避性の判定を行う．主情報は主計画に組み込まれるため，主情報として選ばれた文節から文の最後の述語となる文節にたどり着くために通る経路上の文節の係り受け関係には，全て省略不可避性があるとする．

以上は，言わば非文法的な文の発話を回避するために必要な処理であるが，さらに，内容的に含めたほうが良いと思われる重要語も主情報に含める．このために，松尾らの手法 [18] をニュース記事の性質を踏まえて変更した重要語抽出処理を用いている [17].

3.5 計画生成部

計画生成部では，構造解析部で得られた結果を用いて，**3.3** で述べた発話計画を作成する．ここでは，**3.4** での例をもとに生成した，図 8 に示した発話計画の例をもとに，主計画，副計画の生成について説明する．

主計画の生成: 主計画は，前節で述べた主情報により構成する．例に挙げた文では,「羽生結弦選手が」,「国際大会を」,「欠場することになりました」という 3 つの文節が主情報であった．

選ばれた文節をもとに語順にしたがって配列して作成した文を適切な長さに分割する．これは，文節を連結した文をそのまま読み上げてしまうと，ユーザが短い反応を挟む間を奪ってしまう可能性があるためであ

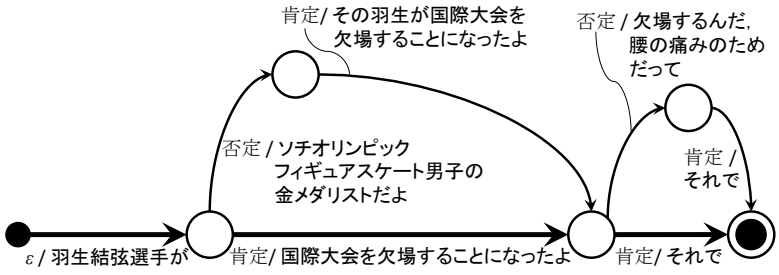


図 8: 発話計画の例

る．一方で，あまり短い単位毎に発話を区切ってユーザの反応をうかがうのは，ユーザに煩わしさを感じさせる可能性があり適切ではない．このため，簡単な文法的な規則を適用し，文節を連結することで適切な長さの発話を構成する．ここでは，海木らのポーズ挿入規則 [9] を参考に，次の条件を満たす箇所では分割し，それ以外では連結する．

- 当該文節の句が右枝分かれ，かつ，先行文節が左枝分かれ
- 当該文節が読点 (、) を含む

この規則を先の例に適用すると,「羽生結弦が」と「国際大会を欠場することになりました」という 2 つの発話となる．つまり，それぞれの区間をシステムが発話した後には一定のポーズがおかれるので，ユーザが反応を入れやすくなると考える．

副計画の生成　構造化された情報の中で，主計画に含まれる文節に係っている文節のうち，省略不可避性がない文節が副計画を構成する．さらに主計画の場合と同様に，各文節から省略不可避性を有する文節を再帰的に辿って取得した文節を計画に含める．図 7 の例で,「国際大会を」に対する副計画を生成するとき，まず「国際大会を」に係る文節である「予定していた」が選択され，それに対する省略不可避性をもとに「出場を」「フィンランドで」が選択される．全体として「フィンランドで」「出場を」「予定していた」という文節が「国際大会を」の副計画を構成することになる．ここで選択されなかった「来月」などの文節は,「予定していた」に対する副計画となる．このように発話計画は階層性を持つ．副計画も，主計画と同様の規則で文の分割を行う．

4　対話例と課題

提案システムとユーザの対話例を図 9 に示す．システムが主計画に沿って発話をすることで，ユーザは相槌をはさむ (肯定的応答 ACK)，あるいは黙っているだけでニュースの要点が得られる (受動的な情報消費)．また，ユーザの知らない言葉，興味を引いた単語，あるいは，理解できない表現が出て来たときには相槌 (否定的応答 NACK) や聞き返しを行うことが想定される．このときシステムは，対応した副計画に沿った発話を生成することで情報を補足する (background の情報行動) ．このように，ユーザの状況 (理解や興味の状態) に合わせながらニュース記事の内容を音声によって効率よく伝達することが実現できる．

S: 羽生結弦選手が	[主計画]
U: うん	[肯定応答]
S: 国際大会を欠場することになったよ	[主計画]
U: え？	[否定応答]
S: 欠場するんだ、腰の痛みのためだって	[副計画]
U: へー	[肯定応答]
S: それで、日本スケート連盟によると	[主計画]
U: うん	[肯定応答]
S: 羽生選手は全治 4 週間と診断されたんだって	[主計画]
U: そうなんだ	[肯定応答]
S: スケート連盟は	[主計画]
S: 『症状は重くない。練習が全くできない状況ではない。中国大会に向けて	[主計画]
U: 中国大会？	[反復応答]
S: うん、グランプリシリーズのだよ	[副計画]
U: ふーん	[肯定応答]
S: それで、万全の状態で望むためだ』としているよ	[主計画]
U: そうなんだ	[肯定応答]

図 9: 対話例．“S:” はシステム発話，“U:” はユーザ発話を表す．システム発話には主計画と副計画のどちらから生成されたか，ユーザ発話には反応がどの応答に分類されたか，をそれぞれ付与している．

むろん，課題も多く残されている．例えば,「国際大会を」と「欠場することになった」は，計画上では一つの発話としてまとめられている．二つの情報を一つ

にまとめた発話に対してユーザの聞き返しがあった場合，どちらに関する補足を行うかは自明ではない．また，一つの情報に複数の補足情報が存在することもある．従って，補足対象が定まったとしても，どの補足情報を伝えるべきかは別途決めなければならない．現状ではこれらの問題に対して,「発話中で最も後ろの情報に対して優先的に補足を行う」「時間や場所以外の補足情報を優先する」といった規則を適用しているが，より適切な情報提示のためには，どの補足情報が重要かといったことを考慮すべきである．また，用語や人物の説明といった補足情報は，それらの一般的な知名度や，個人の知識や嗜好などよって変化することを踏まえ，ユーザの反応に対する適切な補足情報の提示となるような発話計画を立てる手法を確立することが求められる．

5 関連研究

従来より，特定のタスク (交通案内や天気情報提供など) を対象として，ユーザからの発話に応じて情報を提示する質問応答型の対話システムが研究されてきた [8, 10, 14]．このようなシステムでは，ユーザの明示的な情報要求に応じて限定的な情報を確実に提供することに主眼が置かれていた．

近年，質問応答と組み合わせて，システム側からユーザに主体的に情報を提示する対話システムの研究も進み [19, 11, 12]，文書で表わされるような，まとまった量の情報提供を行う音声対話システムも提案されてきた [7, 20]．しかしながら，これらのシステムも基本的にはユーザの質問に対してシステムが回答を提示するという点では質問応答型の対話になっていると言ってよい．

ところで，ユーザにとって質問という行為は，システムの発話内容を理解した上で，問いかける内容を明示的に言語化する必要があるため，比較的負荷が高い．そのため，この種のシステムにおいてはユーザからの質問がなされにくく，システムは要約が提示することが主な機能となり，対話をとおして必要十分な情報を効率良く伝達することは困難であった．

これを解決するには，ユーザが質問を発しやすい状況を作り出すことが必要になる．音声対話の特性を考えれば，ここで言う質問には，言語表現を用いた明示的な情報要求だけでなく，相槌や聞き返し，相手の発話の一部の反復などの短い反応 (本稿では即応的情報反応と呼んだ) を含めて考えるべきであることはすでに論じたとおりである．これらの反応や情報要求を認識するには，発話される語句の音声認識が必要であることは当然であるが，状況や態度を表出する手段としてのパラ言語情報の識別が重要となる．さらには，身振り手振りや顔の動きなどのマルチモーダル情報も手がかりを与える．当研究グループでは，これまでにもユーザの短い反応を韻律情報などを利用して認識するシステムを提案してきている [16, 4, 6]．

6 議論

適用領域: まず，提案したような音声対話による情報アクセス・情報伝達システムの適用領域についてであるが，やはり，音声対話メディアの特性から，いわゆるユーザが「手が離せない」状況が考えられる．このような条件に合致した適用場面としては，機器の運転中や，料理などの手作業中などが考えられる．一方，システムに対して音声で話しかけたり反応を返したりが行いやすいかという点も問題になる．この問題には，もちろんユーザの嗜好や特性も影響するが，ある種の擬人化エージェント的なインタフェースが有用である可能性も考えられる．また，ユーザの反応を引き出しやすいような，システム発話の生成 [5] も有効な要素となろう．

情報アクセスシステムとしての位置づけ: 通常のテキストを中心とする視覚的メディアを用いた情報アクセスの研究は盛んに行われている．しかしながら，いわゆるサーチエンジンを超えるようなポピュラリティを得ているシステム・インタフェースはほとんどないと言える．一方で，先に指摘したように，音声メディアが適している，あるいは，音声メディアしか使えないような利用状況が考えられる．ただし，記事や文書のようなまとまった情報を音声で伝達したり，ブラウジングすることには困難がある．そのような意味で，音声対話によってもたらされるインタラクションを導入することにより,「基本は受動的だけど，能動的なつっこみもできる」情報アクセスシステムを実現しようとする本研究の方向性には，これまでにはなかった可能性があると考える．

対話システムとしての位置づけ: 対話システムの分類の軸として，対話の主導権をシステムが持つか，ユーザが持つか，または両者の混合かというものがある [13]．本研究の音声対話システムは，現在の範囲においては,「基本はシステム主導で，必要に応じてユーザ主導」になる．ただしこれは，微妙なコントロールを短い音声反応で行える範囲に限定しての話である．この成約は，対話制御部を非常にシンプルなものにするのに貢献するが，今後もっと明示的で複雑なユーザの情報要求を扱おうとする場合，情報要求をシステム内部の情報検索過程に対応付けるための対話や，答えられない要求に対する対話などの複雑な対話制御が必要となる．もっ

とも，サービス・機能的な側面から，どの程度のことまで行うべきか，行えるかを定めるための検討も必要となる．

7 おわりに

ニュース記事から，ユーザの反応を想定した発話計画を作成し，それに従って対話を行うことで，ユーザから見れば必要十分な情報アクセス，システムから見れば効率的な情報伝達が行える音声対話システムを提案した．提案システムのユーザは，音声対話において自然と考えられる反応を返したり，情報要求を提示することにより，過不足のない情報アクセスを実現することができる．

今後は提案システムの枠組みを発展させ，さらに効率的で豊かなインタラクションを実現することを目指す．そのための課題として，パラ言語を利用したユーザの反応の認識，複数のニュース記事群を対象とした，より適切な情報内容の選択と構成，ユーザの反応を引き出すような対話の展開，親しみやすい発話音声の生成など個々の技術における精度や使い勝手の向上が挙げられる．

本システムは明確なタスク達成指向のシステムではなく，また，話を継続することを目的とする雑談システムでもない．また，今後は楽しく役立つ情報対話といった要素も加味していきたいと考えている．その意味で，評価の観点や方法論が現状では未確立である．よって，構築したシステムを実際の対話で評価しながら，これらを確立することも課題である．

参考文献

[1] D.O. Case: *Looking for Information, A Survey of Research on Information Seeking, Needs, and Behavior, Second Edition*, Academic Press (2007)

[2] S. Erdelez: Information encountering, In [3], pp.179–184, (2005)

[3] K.E. Fisher, S. Erdelez, and L.E.F. EmKechinie (Eds), *Theories of Information Behavior*, Information Today, Inc. (2005)

[4] S. Fujie, R. Miyake, and T. Kobayashi: Spoken dialogue system using recognition of user's feedback for rhythmic dialogue, *Proc. Int. Conf. Speech Prosody*, OS2-4 (2006)

[5] K. Iwata and T. Kobayashi: Speaker's intentions conveyed to listeners by sentence-final particles and their intonations in Japanese conversational speech, *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 6895–6899 (2013)

[6] T. Kobayashi and S. Fujie: Conversational Robots: An Approach to conversation protocol issues that utilizes the paralinguistic information available in a robot-human setting, *Acoust.Sci. & Tech.*, Vol. 34, No. 2, pp. 64–72 (2013)

[7] Y. C. Pan，H. Y. Lee，L. S. Lee: Interactive spoken document retrieval with suggested key terms ranked by a Markov decision process，*IEEE Trans. Audio, Speech, and Language Processing*, Vol. 20，No. 2, pp. 632–645 (2012)

[8] S. Seneff, J. Polifroni: Dialogue management in the Mercury flight reservation system, *Proc. 2000 ANLP/NAACL Workshop on Conversational systems*, Vol. 3, pp. 11–16 (2000)

[9] 海木延佳，匂坂芳典: 局所的な句構造によるポーズ挿入規則化の検討，信学論 (D-II)，Vol. J79-D-II，No. 9，pp. 1455–1463 (1996)

[10] 駒谷和範，上野晋一，河原達也，奥乃　博: ユーザモデルを導入したバス運行情報案内システムの実験的評価，情処学研報，SLP，2003.75，pp. 59–64 (2003)

[11] 杉山　聡，堂坂浩二，川端　豪: 音声対話によるテキスト内容の伝達方法，情処学論，Vol. 41，No. 6，pp. 1883–1894 (2000)

[12] 杉山弘晃，南　泰浩: 情報提示対話を主導するシステムのためのユーザの潜在的情報要求の推定，信学論 (A)，Vol. J95-A，No. 1，pp. 74–84 (2012)

[13] 中野幹生，駒谷和範，船越孝太郎，中野有紀子 : 対話システム，コロナ社 (2015)

[14] 西村良太，北岡教英，中川聖一: 応答タイミングを考慮した雑談音声対話システム，人工知能学研資，言語・音声理解と対話処理研究会，Vol. 46，pp. 21–26 (2006)

[15] 野本忠司: 係り受け構造の刈り込みと CRF による文の要約，言語処理学会年次大会，pp. 488–491 (2008)

[16] 藤江真也，江尻　康，菊池英明，小林哲則: 肯定的/否定的発話態度の認識とその音声対話システムへの応用，信学論 (D-II)，Vol. J88-D-II，No. 3，pp. 489–498 (2005)

[17] 藤江真也，福岡維新，麥田愛純，高津弘明，林 良彦，小林哲則: 効率的な情報伝達を志向した音声対話システムの提案，人工知能学会 第 74 回 言語・音声理解と対話処理研究会，SIG-SLUD-B501-02. (2015)

[18] 松尾　豊，石塚　満: 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム，人工知能学論，Vol. 17，No. 3，pp. 217–223 (2002)

[19] 翠　輝久，河原達也，正司哲朗，美濃導彦: 質問応答・情報推薦機能を備えた音声による情報案内システム，情処理学論，Vol. 48，No. 12，pp. 3602–3611 (2007)

[20] 吉野幸一郎，河原達也: ユーザの焦点に適応的な雑談型音声情報案内システム，人工知能学研資，言語・音声理解と対話処理研究会，Vol. 70，pp. 53–58 (2014)

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第11回)
SIG-AM-11-06

ソーシャルシェアデータを用いた観光エリア推薦システム

Sightseeing-Area Recommendation Based on Social Data

加藤風太 [1*]　熊野雅仁 [2]　木村昌弘 [2]
Futa Kato[1]　Masahito Kumano[2]　Masahiro Kimura[2]

[1] 龍谷大学大学院理工学研究科電子情報学専攻
[1] Division of Electronics and Informatics, Ryukoku University
[2] 龍谷大学理工学部電子情報学科
[2] Department of Electronics and Informatics, Ryukoku University

Abstract: 近年，Web 空間に共有化された Geo-tag 付ビッグデータを観光に応用する研究が注目されている．従来研究では，主に観光スポットが着目されてきた．我々は，ユーザの実世界における行動履歴情報を集合知的観点から集約することで，個人化推薦システムの構築を目指している．本研究では，訪れる地域が指定されたとき，大量の Geo-tag 付き写真に基づいて複数の観光スポットを含む地域を効果的に可視化・推薦する観光エリア推薦システムを提案する．

1　はじめに

近年，ソーシャルメディアの発展により，ユーザが日常行動に応じて感じた思いや気持ちを言語化したり，評点，賛意の表明など，様々なアクティビティを通じて Web 空間に公開し，シェアする時代が到来しており，人々の多様な意見や嗜好情報を含んだソーシャルシェアデータがビッグデータとして注目されている．

一方，人々の観光行動においても，観光先で，実際に，何にどのような魅力を感じたかについての情報が Web 空間で公開され，シェアされ始めているため，エビデンス（実行動）に裏づけられた，人々の観光に関する潜在的な嗜好を抽出し，有益な情報を旅行者に還元できる新たなサービス創出の可能性が高まっている．

これまで，観光産業では，独自の調査と専門的な知識に基づいて観光案内情報を収集・集約し，魅力的な観光先の推薦を行ってきた．しかし，人々の多様で潜在的な嗜好を捉えることができれば，これまでの観光産業による観光先や，多くの人に好まれる観光先に加え，観光の専門家が気づかなかった，もしくは注目しなかった観光先を嗜好に応じて適切に推薦できる可能性があるため，個人化され，より洗練された推薦システムを構築できる可能性が期待される．また，これまで，観光対象を検索するシステムでは，主に個々の観光スポットが独立に扱われるため，例えば，検索結果として得られた観光スポット上位二つが，実空間上，距離的に離れているため，ユーザの持ち時間では二つを訪れることができない問題などが生じる場合も考えられる．つまり，観光先に関する推薦では，個々の観光スポットを推薦するよりも，観光スポットが複数含まれる，実空間上のエリアを推薦することが望ましいと思われる．さらに，例えば「お寺好き」「街歩き好き」「お酒好き」など，ユーザは一般に多重嗜好を持つので，ユーザの多重嗜好を効率良く反映できる推薦システムが望まれる．

本研究では，ユーザが訪れる都市を指定したとき，ソーシャルメディアデータとしての何年にも渡る大量の Geo-tag 付き写真データから集合知的観点に基づいて，多くの人々に好まれる観光先だけでなく，行動パターンが近い撮影者達の訪問先も強調して複数同時に可視化することで，ユーザの多重嗜好に応じた観光スポットを地図システム上に可視化する観光エリア推薦システムを提案する．2010 年から 2014 年までの日本で撮影された Geo-tag 付写真による実データを用いて，評価実験を行い，提案する観光エリア推薦システムの有効性を示す．

2　観光エリアの個人化推薦

個人の嗜好に応じた適切な推薦を行う上で，推薦システムを使うユーザの嗜好データがない場合や，推薦する対象に嗜好データがない場合，ユーザへの推薦がうまくいかないというコールドスタート問題がある [1]．コールドスタート問題の解決は重要な課題となっている．

また，例えば，訪れる都市が決まっているものの，その都市や周辺地域に存在する施設や観光先を知らない場合を考える．予め，ユーザに関する嗜好情報が得られ

*連絡先：龍谷大学
滋賀県大津市 瀬田大江町横谷 1-5
E-mail:t14m008@mail.ryukoku.ac.jp

ていない場合，ユーザ自身の好みが明瞭な場合は，ユーザに，好みを限定するためのカテゴリー選択やキーワードの入力を要求する方法が考えられる．しかし，現地に何が存在するか知らない場合や，ユーザ自身が自分の嗜好をうまく把握できていない場合など，カテゴリーの選択が難しく，検索用のキーワードが思いつかない場合さえある．したがって，多くのユーザにとっては明瞭なキーワードやカテゴリーを特定して観光先を絞ることが一般に困難である．

一方，近年，Trip Advisor や Yelp, Foursquare など，施設への評価を投稿する施設共有サイトが注目されている．ユーザがこれらのサイトにおいて，過去に訪れた施設への評価を登録していれば，嗜好情報として推薦システムに与える方法を考えることができる．ただし，例えば，近年の訪日外国人が興味を示す観光スポットとして，東京・渋谷のスクランブル交差点や，新宿ゴールデン街の街並みなど，明瞭な施設ではない意外な場所が観光先となっていたり，京都・伏見稲荷大社の千本鳥居など，大きな敷地に存在する施設の一部に人気が集中する場合もあるため，きめ細かな嗜好を捉えるには，施設という単位に依存しない柔軟な観光先の捉え方と，その観光先に関する嗜好を捉え得る手法が望まれる．

我々は，人々が多重の嗜好を持つ傾向があると考えており，個人向けのきめ細かい推薦を行うためには，指定された都市や周辺地域について，明瞭なキーワードやカテゴリーを特定して観光先を絞り単一の観光スポットを探し当てるのではなく，嗜好の近さを可視化しつつ多様な観光スポットを同時に提示することが望ましいと考えている．

これらの観点に対して，本研究では，人々の写真撮影行動に着目する．近年，GPS に基づく情報（Geo-tag）を写真に付与できるカメラやスマートフォンの普及とともに，Flickr などの写真共有サイトが普及することで，一般の旅行者達が観光した際に撮影した大量の写真が，撮影場所の緯度経度，撮影日時，焦点距離などの撮影条件，付与されたコメント，撮影者のプロフィールやソーシャルネットワーク情報などとともに，Web 空間に蓄積され続けており，世界的に大規模なソーシャルシェアデータとなっている．旅行者は，視覚的に興味を抱いた対象に出くわすと写真を撮影する傾向があり，厳選した写真を Web 空間に公開する傾向があると考えられるため，質の高い嗜好情報が得られる可能性がある．そのため，写真の撮影行動を集約し，人気スポットを抽出する研究 [2] や，観光へ応用する研究が注目されている [3]，[4]，[5]，[6], [7]．つまり，Geo-tag を用いた集合知的観点から観光スポットを抽出するアプローチを採用し，嗜好が似た撮影者群を捉えることで，施設名を持たない意外な観光スポットも抽出できる可能性が期待される．また，ユーザのカメラやスマートフォンに蓄積された過去の写真群や，写真共有サイトに公開した写真と付随する情報をユーザの嗜好データとして推薦システムに与えれば，撮影者が自らの嗜好をうまく言語化できない場合でも，嗜好に応じた観光先を推薦することができる可能性がある．さらに，ユーザの嗜好に合う多様な観光スポットを嗜好の近さが判別できるよう地図上に同時に可視化すれば，観光において，現実的に使用できる制限された時間帯で，どのエリアを観光するかについて，効率的な選定を可能にするシステムの実現が期待される．本研究では，これらの観点に基づいた観光エリア推薦システムを提案する．

3 提案システム

3.1 システム概要

ユーザが訪れる都市の観光エリアを推薦する問題に対して，我々は，まず，明確な領域を持つ施設単位で観光スポットを捉えるのではなく，過去において，人々の実際の行動に裏づけられた未知数の観光スポットを含む観光エリアを自動的に抽出するため，大量の Geo-tag 付写真データから観光エリアを抽出する．ここで，既存の施設は土地区画上の領域を持つが，提案法での観光エリアは実行動に基づいて定まる領域であることに注意しておく．抽出された多数の観光エリアのうち，撮影者の人数が多いほど，人気エリアであると見なせる．人気エリアは，どのような嗜好のユーザに対しても，基本的な推薦対象と考えられる．また，全く嗜好情報が得られないユーザに対して人気エリアを推薦すれば，ユーザに関するコールドスタート状態でも，推薦先が無い状況を避けられる点に注意しておく．

次に，過去の撮影行動情報が得られるユーザに対し，観光エリアの個人化推薦問題におけるユーザの多重嗜好に配慮した推薦手法の第一歩として，本研究では，協調フィルタリングの観点に着目する．例えば，京都を訪れる予定のユーザ u が過去に北海道で写真を撮影した観光エリアにおいて，同様に撮影を行った他のユーザ w が既に京都に訪れ写真を撮影した観光エリアが存在した場合を考える．このとき，協調フィルタリングでは，ユーザ u とユーザ w の過去の撮影行動が似ているほど，類似度が高い関係と見なし，ユーザ w が過去に訪れた観光エリアを推薦する．ここで，ユーザ u が多様な撮影行動をしている場合，ユーザ u の異なる嗜好ごとに類似性の高いユーザ w, w' が，過去に京都を訪れていれば，ユーザ u が訪れる京都で，多様な嗜好に基づくエリアを類似性が高い観光エリアとして同時に推薦できる可能性がある．

本研究では，人気エリアと多重嗜好を考慮したエリアの両方を重視することから，両者を統合して観光エリアとしてユーザに提示する推薦手法を提案する．

3.2 観光エリア抽出

施設単位の観光スポットではなく，過去の撮影行動に基づいて観光エリアを抽出するため，ここでは MeanShift 法 [8] に基づく観光エリア抽出法を適用する．また，ユーザが指定した都市に対し，観光エリアを推薦するための，各種推薦手法と提案推薦法について述べる．

3.2.1 入力データ

正の整数 T に対して，T 日の期間 [1,T] 内に撮影された写真データ全体の集合を $\mathcal{D}_0$ とする．本研究では，集合知的観点から個人差を吸収し，撮影者の人数を重視するため，緯度と経度 2 の 2 次元平面を離散化し，最小矩形領域ごとに，1 人の撮影者による写真を 1 枚抽出する正規化を行う．このとき，最小矩形領域内の写真数は，その領域内で撮影を行った人数に相当することを注意しておく．この離散正規化を $\mathcal{D}_0$ に適用して得られる写真集合を，

$$\mathcal{D}_1 = \{d_n; n = 1, \cdots, N\}$$

とする．また，$\mathcal{D}_1$ の写真を撮影したユーザ集合を

$$U_1 = \{u_i; i = 1, \cdots, M\}$$

とする．各写真データ d_n には位置情報 x_n，時間情報 t_n，ユーザ情報 u_m とが付随しており，

$$d_n = (x_n, t_n, u_i),\ (n = 1, \cdots, N,\ i = 1, \cdots, M)$$

と記述する．ただし，$x_n = (x_{n,1}, x_{n,2})$ であり，$x_{n,1}$ と $x_{n,2}$ はそれぞれ緯度と経度，t_n は d_n が撮影された日，N 写真データ総数，M はユーザ総数である．

3.2.2 Meanshift 法に基づく観光エリア抽出法

d 次元 Euclid 空間 $\mathbf{R}^d$ 上の点群 $S = \{s_n\}_{n=1}^N$ がある確率分布に従う標本集合であるとき，任意の点 $s \in \mathbf{R}^d$ における確率密度関数を，ノンパラメトリックアプローチであるカーネル密度推定

$$\hat{p}(s) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^d}G\left(\parallel (s - s_n)/h \parallel^2\right), \left(s \in \mathbf{R}^d\right)$$

により推定することを考える．ここに，$\parallel\ \parallel$ は $\mathbf{R}^u$ の Euclid ノルム，$G(z)$ はカーネル関数である．ここで，$G(z)$ は，Epanechnikov カーネルを利用する．また，h (> 0) は，データの各位置 s_n ごとに存在する確率密度関数のバンド幅を規定するパラメータである．Crandall ら [2] は，写真データ集合 $\mathcal{D}_1$ から主要な撮影地域を抽出する手法として，緯度と経度に基づく d=2 次元の点群 $\{x_n\}$ を対象として MeanShift クラスタリングを適用している．

ところで，主要な撮影地域は，街角レベルから都道府県，州，国規模など，様々なスケールが考えられるため，最適なサイズを容易に決定できない問題がある．本研究では，徒歩圏内の多くの撮影者が集まる地域を観光エリア A_k と呼び，Crandall らの Metropolitan-scale (h=100m) として Epanechnikov カーネルを用いた Meanshift 法 [8] によって観光エリア A_k の抽出を行う．ここで，観光エリア A_k は，その領域に含まれる写真群の撮影位置 $x_n \in A_k$ の集合とするが，撮影位置の集合に基づく地理空間上の領域を示しているとする．また，Δt_0 年内（期間 I_0 と呼ぶ）に撮影された写真群を $\mathcal{D}_2 = \{d_n \in \mathcal{D}_1; t_n \in I_0\}$ とするとき，$\mathcal{D}_2$ を対象に Meanshift 法で抽出された観光エリア集合を $\mathcal{A}_0 = \{A_k; k = 1, \cdots, K'\}$ とする．また，A_k に含まれる写真集合を $D_k = \{d_n \in \mathcal{D}_2; x_n \in A_k\}$ としたとき，$|D_k| > \mu_0$ を満たす観光エリアを

$$\mathcal{A}_1 = \{A_k; k = 1, \cdots, K\}$$

と記述する．ただし，K は抽出された観光エリアの総数である．

3.2.3 新規撮影地点の観光エリア配属法

ユーザ u_i が撮影した観光エリアと，他の撮影者 u_j, $(i \neq j)$ が撮影した観光エリア $\mathcal{A}_1$ が一致するかを調べる方法を述べる．本研究では，提案法を評価する際，ユーザ u_i が，新たな期間 I_1, $(I_0 \cap I_1 = \emptyset)$ に訪問する観光エリア A_k を予測することで，推薦システムの性能評価を行う．ただし，期間 I_0 で抽出される観光エリアと，期間 I_1 で抽出される観光エリアは同じ Meansift 法を適用しても，全く同じになるとは考えにくい．期間 I_0 には全く撮影が行われていなかったエリアが I_1 で撮影が行われるようになったり，ほぼ同じエリアでも，期間の違いにより，写真数や撮影位置が変化し得るため，完全に領域が一致しないものがほとんどであると予想される．このとき，ユーザ u_i が，期間 I_1 に撮影を行った位置と，他のユーザ u_j が撮影を行った位置が同じ撮影エリアであるかを定める上で，曖昧性が生じる．そこで，本研究では，期間 I_1 に撮影された写真がどの観光エリアに帰属するかは，過去の期間 I_0 で抽出された観光エリア $\mathcal{A}_1$ に帰属するかを調べるという方法を採用する．これは，期間 I_0 のデータで生成した確率密度関数を用いて新規撮影地点の収束先を求め，確率密度関数の極値近傍に収束したか否かを判別する方法となる．つまり，いずれかの極値の近傍と判断すれば新規撮影地点の配属先が定まる．ただし，いずれの極値近傍でもないと判断されれば，配属先がないことになるが，これは，新しく表れた観光エリアである可

能性があるものの，本研究における予測実験においては対象外と見なす．

3.3 観光エリア推薦法

ユーザ u_i から訪れる都市 c が指定されたとき，本研究では，$\mathcal{A}_1$ のうち，c に含まれる観光エリア集合を $\mathcal{A}_1^c$，c 以外の都市 c', $(c \neq c')$ に含まれる観光エリア集合を $\mathcal{A}_1^{c'}$ とする．このとき，各観光エリア $A_k \in \mathcal{A}_1^c$ に優先順位を与え，ランキング形式で観光エリアを推薦する手法を採用する．ここでは，優先順位を与える方法として，人気エリア法，二つの協調フィルタリング法，混合法について述べる．

3.3.1 人気エリア法

抽出された観光エリア A_k に含まれる写真群の数 $|D_k|$ が多いほど人気度の高い観光エリアであると言える．つまり，推薦する観光エリア A_k のスコア AS_k を

$$AS_k = |D_k|$$

とし，人気度の高さに応じて AS_k をランキングする手法である．この手法では，たとえユーザ u_i に関して，過去の撮影行動データが無いコールドスタート問題が生じる場合でも，多くの人々に人気のある観光エリアを推薦することができる．ただし，すべてのユーザに対して同じ推薦結果となることに注意しておく．

3.3.2 協調フィルタリング法

ユーザ u_α へ観光エリアを推薦する際，ユーザ u_α と行動類似性の高い，他のユーザ u_β が過去に撮影した観光エリアを推薦する手法である．ユーザ u_α と，ユーザ u_α 以外のユーザ u_β, $(\alpha \neq \beta)$ の行動類似性を $sim(u_\alpha, u_\beta)$ で表す．ユーザ u_β が都市 c でこれまでに撮影したことがある観光エリアの全体の集合を $RA_{u_\beta}^c$ とする．このとき，推薦する観光エリア A_k のスコア AS_k を

$$AS_k = \sum_{u_\beta \in U_1} \chi_{k.\beta} sim(u_\alpha, u_\beta)$$

で定義する．ここに

$$\chi_{k.\beta} = \begin{cases} 1 & (A_k \in RA_{u_\beta}^c) \\ 0 & (A_k \notin RA_{u_\beta}^c) \end{cases}$$

とする．これは，投票形式で観光エリア $A_k \in RA_{u_\beta}^c$ にスコアを加える方法となるが，ユーザ u_α と行動類似性の高い他のユーザが撮影した観光エリアほど，高い値が加算される点で嗜好が考慮されていくことにな

図 1: 観光エリア A_k のスコア AS_k と色の対応

る．つまり，スコア AS_k の大きさに基づいて観光エリアのランキングを行い，推薦を行う手法が本研究における協調フィルタリング法である．また，ユーザ u_β ごとに，異なる嗜好に基づいた類似度がスコア AS_k へ加算されることから，多重嗜好を反映した複数の観光エリアを推薦できる可能性があることに注意しておく．

jaccard 係数

ユーザ u_α が期間 I_0 で撮影経験のある観光エリアの集合を $PA(u_\alpha)$，ユーザ u_α 以外のユーザ u_β, $(\alpha \neq \beta)$ が撮影経験のある観光エリアの集合を $PA(u_\beta)$ とするとき，$PA(u_\alpha)$ と $PA(u_\beta)$ の共起性の度合いを表す jaccard 係数を用いて $sim(u_\alpha, u_\beta)$ を表す方法である．

$$sim(u_\alpha, u_\beta) = \frac{|PA(u_\alpha) \bigcap PA(u_\beta)|}{|PA(u_\alpha) \bigcup PA(u_\beta)|} \quad (1)$$

3.3.3 混合法

本研究では，ユーザの過去の行動履歴が無い場合や，多くの人に好まれる観光エリアが推薦できる人気エリア法に加え，ユーザの過去の行動履歴がある場合には，個人の嗜好に寄り添った推薦を行う方法を実現するため，人気エリア法と協調フィルタリング法の混合法を提案する．A_k に対し，正規化された人気エリア法のスコアを ξ_k，正規化された協調フィルタリング法のスコアを ϕ_k とし，$0 \leq w \leq 1$ としたとき，次のようにする．

$$AS_k = (w-1)\xi + w\phi$$

3.4 観光エリアの可視化法

提案する観光エリアの可視化法としては，ユーザ u_α から指定された都市 c に対し，都市 c に含まれる抽出された観光エリア A_k を可視化する．ただし，観光エリア A_k は，撮影地点の集合であり，抽出された A_k は領域を持っている．そこで，提案法では，観光エリアを A_k の代表地点で表現するのではなく，撮影地点を包含する円領域として可視化することにより観光エリアが地図上において占める領域の相対的な大きさを容易に視認できる可視化法を採用する．円領域の決定法としては，観光エリア A_k 内の撮影地点群 D_k の中心点を定め，その中心から最も遠い撮影位置を半径とした円の内側を観光エリア A_k の領域として可視化に用いる．

また，A_k の領域と同時に，ユーザ u_α の嗜好に合う観光エリア A_k を強調的に可視化し，特徴的なお勧め観

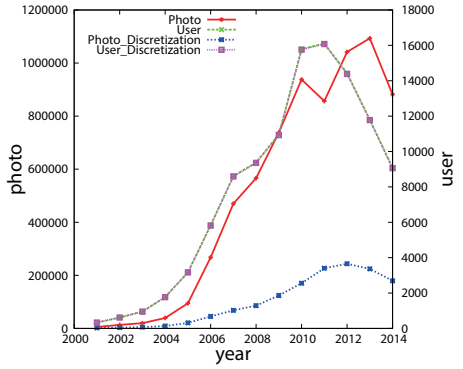


図 2: 写真共有サイト Flickr に登録された日本における Geo-tag 付写真数とユーザ数の変遷

表 1: 初期データセット

DataSet	DataSet'	訓練&学習データ	予測データ（京都）
Dataset1	Dataset1'	2010 年，2011 年	2012 年
Dataset2	Dataset2'	2011 年，2012 年	2013 年
Dataset3	Dataset3'	2012 年，2013 年	2014 年

光エリアの発見を促す可視化システムを提案する．観光エリア A_k を嗜好に応じて強調的に可視化する方法として，ユーザ u_α の嗜好との類似性がスコア AS_k として算出されているため，本研究では，図 1 に示した色の対応を用い，スコア AS_k が低いほど青く高いほど観光エリア A_k の領域を赤く提示することで観光エリア推薦度を強調的に可視化する．これにより，都市 c において，一般的な観光エリアとともに，ユーザ u_α の嗜好に合う複数の観光エリア候補から，現実的な持ち時間や移動距離を考慮して訪問先を吟味することができる．

4 エリア抽出法及び推薦法評価実験

4.1 データセット

日本を対象に提案システムの検証実験を行うため，写真共有サイト Flickr から，日本の WoID（23424859）を持ち，日本国内で位置情報を持ったデータの収集した．図 2 は収集した写真データについて，2000 年から 2014 年までの位置情報付写真数とユーザ数の変遷である．また，離散化を行った結果の写真数とユーザ数の変遷も同様に図 2 に示す．図 2 より 2010 年から 2014 年において写真数が増加しているが，本論文では，投稿が盛んな期間を対象として実験データセットの構築を行う．また，ユーザが指定する都市としては，数多くの多様な観光エリアが存在する京都を対象として実験を行う．実験では 2 年間を訓練データとして，翌年

表 2: 観光エリア抽出後の最終データセット

Dataset	観光エリア	訓練ユーザ	予測ユーザ数
Dataset1	2548	850	620
Dataset2	3122	1050	620
Dataset3	3440	1399	637
Dataset1'	2548	850	560
Dataset2'	3122	1050	582
Dataset3'	3440	1399	604

表 3: 各データセットの観光エリア $|A_k|$ 抽出数と配属率

Dataset	観光エリア $\|A_k\|$	配属率 (%)
Dataset1	108547	69.0
Dataset2	117329	70.1
Dataset3	107779	67.9
Dataset1'	108547	74.0
Dataset2'	117329	76.1
Dataset3'	107779	74.5

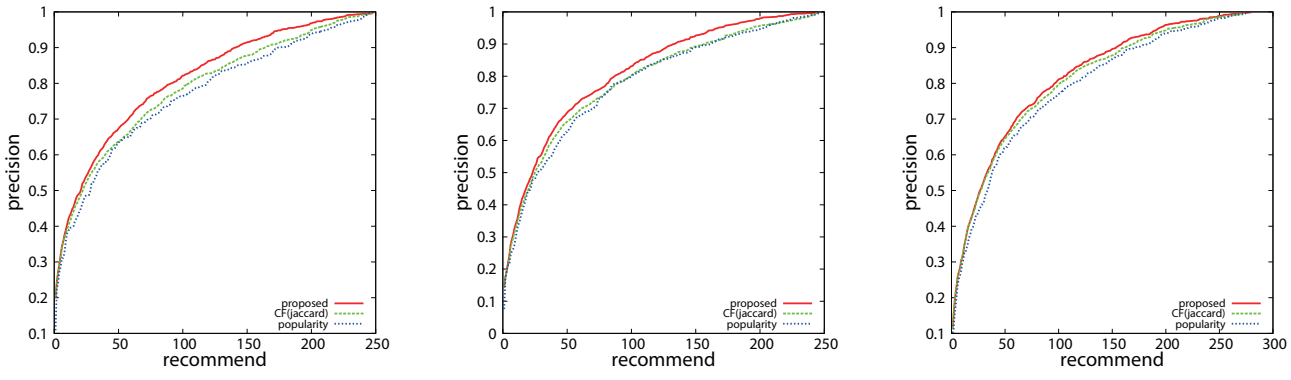
に京都に訪れているユーザを対象に実験を行うためのデータセットを構築した．また，学習データにおいてユーザが京都を訪れている，つまり過去に京都の観光エリアを訪れている場合に推薦において影響がある可能性を考慮し，その情報を削除した Dataset' を構築した．データセットの詳細を表 1 に示す．ただし，抽出された観光エリアを用いて観光エリアの推薦を行う上で，本研究では観光エリアが多数の嗜好を反映している必要を考え，$\mu_0 \geq 10$ を満たす観光エリアを A_k とした．また，協調フィルタリングを行う上で推薦を行うためには訓練ユーザが予測対象都市 c とその他に少なくとも一つの観光エリアを訪れている必要がある．以上の観点から最終的なデータセットは表 2 のようになった．

4.2 観光エリア抽出結果と配属率の結果

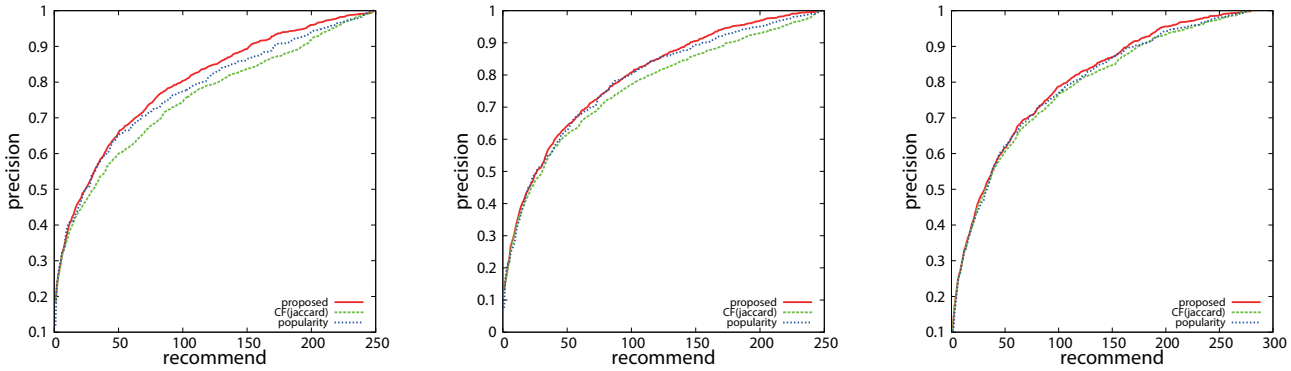
ここで，表 1 の各データセットについて観光エリア A_k の抽出を行った結果を表 3 に示す．また，ユーザ $u \in U_1$ を対象として，予測期間にユーザ u が撮影した地点が抽出されたいずれかの観光エリア A_k に配属される配属率を求めた．各データセットに関する配属率の結果も表 3 に示す．どのデータセットにおいても 7 割前後と比較的高い水準で配属されていることがわかる．

4.3 評価手法

本研究では，予測期間に京都を訪れているユーザが実際に撮影した観光エリアを隠蔽し，ユーザへ推薦する観光先をスコア AS_k のランキング上位から順に推薦したとき，ユーザが実際に訪れた観光エリアを正解データとして，推薦先と一致するかという観点から，適合率（precision）に着目して評価を行った．また，提案



(a) Dataset1　(b) Dataset2　(c) Dataset3



(d) Dataset1'　(e) Dataset2'　(f) Dataset3'

図 3: 各 Dataset における観光エリア推薦法適用結果の precision

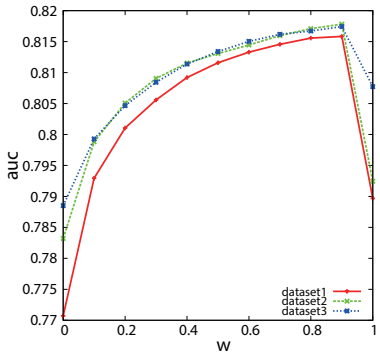


図 4: Dataset におけるパラメータ w と AUC 値の関係

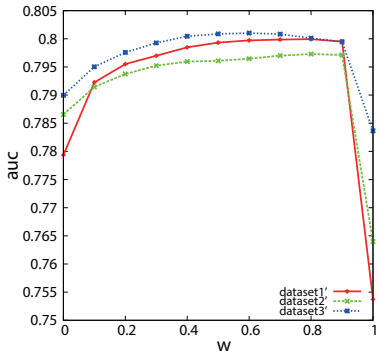


図 5: Dataset' におけるパラメータ w と AUC 値の関係

する混合法において，AUC 値が最も高い値を示す重み w を検証した．

4.4　実験結果

図 3 は各データセットに対し評価実験を行った際の precision の結果である．図 3 において，どのデータセットにおいても協調フィルタリングと人気エリア法が共に高い結果を示している．特に図 3(a)，(b)，(c) においては協調フィルタリングを用いた値が高く，個人化推薦の有用性が示唆されるが，図 3(d)，(e)，(f) においては人気エリア法に比べ協調フィルタリングがわずかに低い値を示している．次に提案混合法を用い，協調フィルタリング法 (jaccard 係数) に人気エリア法を混合することによりアクティビティの少ないユーザへの推薦を補完し，コールドスタート問題に対応した推薦を行う．そのために混合法における最適な重み w を決定すべく実験を行った．その結果を図 4 と図 5 に示す．実験結果，図 4，図 5 ともに混合を行っていない $w = 0.0$(人気エリア法) や $w = 1.0$(協調フィルタリング) より混合を行った結果が，いずれのデータセットにおいても AUC 値が高く性能が良い結果となっている．さらに，重み w の変化における性能検証の結果，いずれのデータセットにおいても混合するとさらに性能が

上がり，重みが $w = 0.9$ の時に AUC の値がほとんどのデータセットにおいて最大になることがわかる．ところで，$w = 0.9$ とした混合法の precision 値が図 3 の赤線であり，どのデータセットにおいても他手法より高い値を示しているため，混合法を用いた推薦の有効性が示唆されていると思われる．

図 3(d)，(e)，(f) において人気エリア法に比べて協調フィルタリングが低い値を示すことについては，ユーザが過去に京都観光しているという京都観光でのリピート性が高い事が嗜好に関係し，予測精度に影響を及ぼしたという仮説が考えられる．また，学習データにおける京都での撮影結果を削除した影響により，学習データ不足によって推薦する観光エリア A_k のスコア AS_k への評価が減少した可能性も考えられる．これらは，今後，対象都市を変えたより詳しい検証が必要であると考える．

5　可視化システムの評価実験

5.1　実験設定

混合法における w について，観光エリア推薦を行った結果から，$w = 0.9$ が最も推薦結果を高くすることが示された．人気エリア法よりも，嗜好を強く反映させた推薦を行うことが効果的であることを示唆していると思われる．そこで，提案可視化システムにおいて，人気エリア法による可視化結果と，混合法による可視化結果を比較検証するため，一例として，Dataset1 を対象とし，人気エリア法による可視化結果と $w = 0.9$ とした混合法による可視化結果を比較する．

5.2　実験結果

人気エリア法による可視化結果を図 6 に示す．図 6 より，京都駅が最も赤く示されている．しかし，それ以下の観光エリアはあまり目立っていなことがわかる．これは，多くの人々が共通して京都駅を撮影する傾向があるためであると考えられる．一方，提案する混合法の可視化結果を図 7 と図 8 に示す．図 7 では複数の個所が赤く表示されていることがわかる．その観光エリアは図 7 において，(a) 金閣寺，(b) 二条城，(c) 銀閣寺，(d) 清水寺，であった．これらは京都における有名な寺社仏閣であり，伝統的な施設を好む嗜好が可視化に反映されていると思われる．次に図 8 では特定の有名な施設を含む観光エリアではなく，図 8 において，(a) 三条通や (c) 四条通といった店や，古来の街並みを残した通り，(d) 錦市場といった商店街が推薦されており，街歩きや食べ歩きを好む多重嗜好が可視化に反映されていると思われる．以上から，混合法は，ユーザの多重嗜好に応じた推薦を行える可能性がある点で提案法の有効性が示唆される．

6　まとめ

大量のメタ情報付写真群を用いて実行動に基づく観光エリアの推薦システムを構築した．提案可視化システムの評価実験により，個人化推薦の観点で有効性を示した．今後は，より多くの都市を対象に評価実験を行い，より洗練された個人化推薦を構築するための探究を行う予定である．

参考文献

[1] 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，vol.23，no.1，pp.89–103，2008.

[2] D.J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, “Mapping the world’s photos,” Proceedings of the 18th International Conference on World Wide Web, pp.761–770, 2009.

[3] S. Kisilevich, F. Mansmann, and D.A. Keim, “P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos,” 1st International Conference on Computing for Geospatial Research & Application, pp.38:1–38:4, 2010.

[4] S. Kisilevich, F. Mansmann, P. Bak, D.A. Keim, and A. Tchaikin, “Where Would You Go on Your Next Vacation? - A Framework for Visual Exploration of Attractive Places,” GeoProcessing 2010, pp.21–26, Feb. 2010.

[5] 王　佳な，野田雅文，高橋友和，出口大輔，井手一郎，村瀬　洋，“Web 上の大量の写真に対する画像分類による観光マップの作成,” 情報処理学会論文誌，vol.52，no.12，pp.3588–3592，2011.

[6] 熊野雅仁，小関基徳，小野景子，木村昌弘，“地理および時間情報をもつ写真データに基づいたホット撮影スポットの抽出,” 情報処理学会論文誌，vol.5，no.3，pp.41–53，Sept. 2012.

[7] 熊野雅仁，岩渕聡，小関基徳，小野景子，木村昌弘，“集合知に基づいたポピュラー撮影スポットに関する旬シーズンの可視化,” 芸術科学会論文誌，vol.13，no.4，pp.218–228，2014.

[8] D. Comaniciu and P. Meer, “Mean shift: a robust approach toward feature space analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.24, no.5, pp.603–619, 2002.

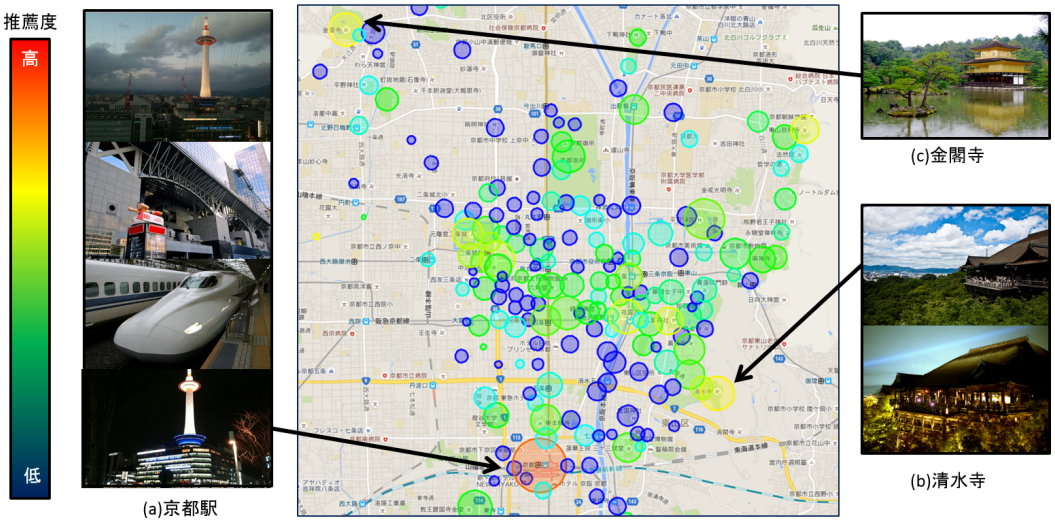


図 6: 人気エリア法を用いた可視化結果

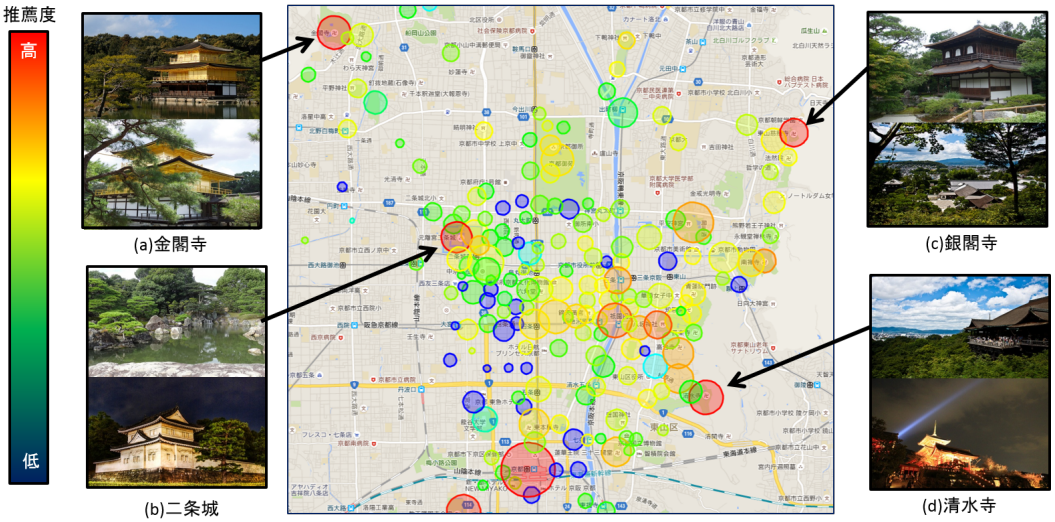


図 7: 提案法を用いた可視化結果 1

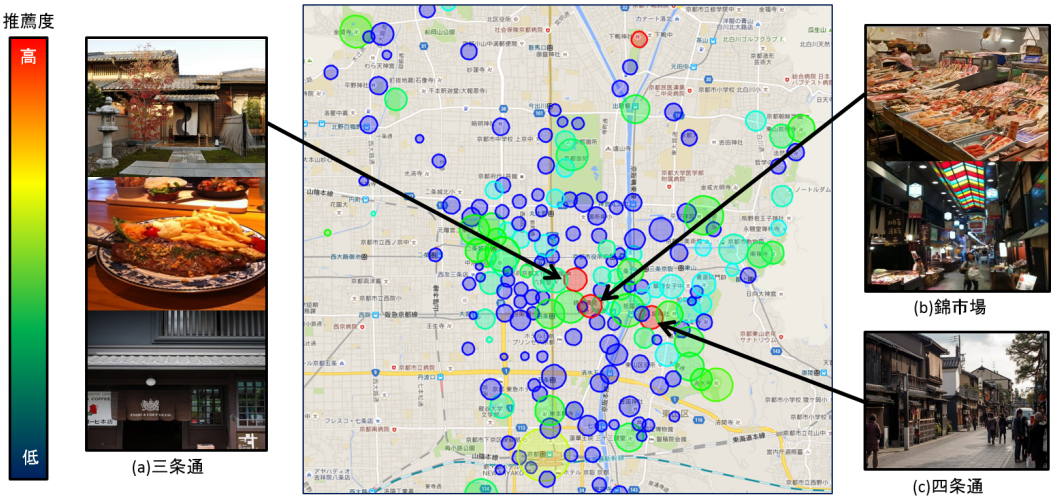


図 8: 提案法を用いた可視化結果 2

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11回)
SIG-AM-11-07

ストリームデータモニタリングにおける確認タイミングの判断支援に関する予備的検討

Preliminary Study on Support of Determining Timing to Monitor Stream Data

吉田和人* 高間康史

Kazuhito Yoshida, Yasufumi Takama

首都大学東京システムデザイン学部
Faculty of System Design, Tokyo Metropolitan University

Abstract: This paper reports the preliminary study on the development of monitoring support system for stream data. It is supposed that stream data such as online news has to be monitored during break of user's primary job. If a user check a stream data at wrong time, the efficiency of his/her primary job would go down. In order to help a user to monitor stream data, we are developing a system that gives a user a clue for determining the timing to monitor with using a dynamic bar chart. This paper reports the result of preliminary experiment, in which the effect of color of bars and individual differences on the timing decision is investigated.

1. はじめに

本稿では，オンラインニュース等のストリームデータを確認するタイミングの判断を支援するシステムの構築に向けて，棒グラフ形式のメータを利用したユーザ実験を行った結果について報告し，棒グラフの色や個人差の影響について考察する．

近年, ストリームデータの量は膨大になっており，ヤフー株式会社の提供するニュースポータルサイトYahoo!ニュース[1]から提供されている「主要」カテゴリのニュースだけでも一日当たり70-100件程度配信されている．また，Twitter[2]では一日当たり5億件のツイートが発生している[3].

これらのストリームデータは有益な情報も多く含んでおり，定期的に観覧しているユーザは多数存在するが，個人が全ての情報を常時確認することは通常不可能である．従って，情報のモニタリング支援を行う必要があると考える．

モニタリング支援に対するアプローチの一つとしてニュースキュレーションサイトが挙げられる．Gunosy[4]では, ユーザの趣向から自動的にニュースを選別し，提供している．これらのサービスは膨大なストリームデータの中から関心のある情報を効率的に発見する作業を支援している．しかし，それらのニュースを確認するタイミングついては，朝刊や夕刊などの定時配信程度であり，個人及びその状況に基づく適切な確認タイミングなどは考慮されていない．

ストリームデータの確認は，休憩時間などの本務の合間に行われる作業とみなすことができる．したがって，モニタリングを頻繁に行えばその都度本務が中断されることになる．反対に，モニタリングの間隔を大きくすれば，ストリームデータの確認に時間を要する可能性があり，本務の中断時間が長くなる．一般に，作業に割り込みが発生すると，知的生産性が低下することが指摘されている．中断後に再開したタスクは中断されない場合の二倍時間がかかること[1], 中断されたタスクの40%は再開されないこと[2]などが指摘されている．従って，適切なタイミングでモニタリングを行うことは，効率的なストリームデータの確認のためだけではなく，円滑な本務遂行の上でも重要と考える．

本稿では，ストリームデータを確認すべきタイミングをユーザが適切に判断することを支援するシステムの構築を目的とする．本務を中断可能なタイミングを計算機が推定する割り込み可能性推定とは異

[1] http://news.yahoo.co.jp/
[2] https://twitter.com/
[3] https://about.twitter.com/ja/company
[4] https://gunosy.com/

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

なり，ユーザ自身がタイミングを判断することを想定している．支援システムでは，ストリームデータの蓄積量を可視化により提示することを想定している．この時，ユーザが適切なタイミングで判断できるように，データ量と視覚的変数のマッピングをユーザごとに調整する必要があると考える．そのための予備実験として，棒グラフ形式のメータを利用したユーザ実験を行った．その結果に基づき，棒グラフの色や個人差の影響について考察する．

2. 関連研究

2.1. 割り込み可能性推定

スマートフォン等の普及に伴い，人とのコンタクトをいつでも取ることができるようになった．しかし，仕事などの別の作業に集中しているときにこのようなコンタクトがあると作業を中断しなければならない．再び作業に戻る際に先ほどまで何をしていたか思い返す必要があるため，1 節で述べたように作業効率の低下につながる. このような問題に対し，作業への割り込みをいつ行うと本務への影響が少なくなるかを推定する研究が行われている[1,2,3,4,5].

ユーザの作業を推定するには，ユーザの状態を観測する必要があり，様々な手法が提案されている．卓ら[3]は推定可能な作業の汎用性を考慮してリストバンド型センサの 3 軸加速度データを用いてユーザの状態を取得している．谷ら[4]はセンサを身体につける煩わしさの観点から机上にかかる圧力のデータを用いてユーザの状態を取得している．田中ら[5]は，検知の容易さ，PC を用いた作業との親和性の観点から，利用アプリケーションの切り替えデータを取得して推定に用いている．

2.2. 情報可視化システム

ストリームデータを可視化することによるモニタリング支援システムが研究されている[6,7,8].

沼野ら[6]は定期的なオンラインニュースのモニタリング作業を支援するインタフェースを提案している．ニュース記事の文章クラスタリングに基づく話題の検出，追跡を行い，それらを新着記事，話題記事ごとの確認を行うリストモード，関心のある話題の新着記事数を確認できる続報記事確認モードの二つのモードを用いて可視化を行っている．続報記事確認モードでは続報記事数を黄色い四角形の数で可視化している．

奥村ら[7]は定期的な BBS (電子掲示板) のスレッドのモニタリング作業を支援するインタフェースを提案している. キーワードベースの可視化を採用し，現在関心のある話題に関する投稿を追跡可能であるほか, 新たな話題の発見も可能となっている. また，特定のスレッドから抽出したキーワード，複数スレッドから抽出したキーワードをそれぞれ別のビューで提示する，Overview + details を採用している．

黒澤ら[8]は OSS (オープンソースソフトウェア) の複数バグ管理システムから継続的に配信されるバグ更新情報のモニタリング作業を支援するインタフェースを提案している．報告されたバグはノードとして可視化され，ノードの大きさによりバグの修正に向けた進展や議論の進捗，色によりバグの修正状態を表現している．また，前回確認時から変化していない部分の縮小表示や軌跡の描画によって変化部分の確認を容易にすることで，効率的なモニタリングを支援している．

3. 予備実験用インタフェース

本稿では，モニタリングすべきタイミングを判断する手がかりをユーザに提示する汎用的な手段を検討する．テキストストリームデータの種類によらず必要な手がかりとして，モニタリングしていない間に到着したストリームデータの蓄積量を提示することを考える．本稿では，ストリームデータの蓄積量を棒グラフにより可視化する．データ量にマッピングすべき視覚的変素として，棒グラフの高さだけでなく色も併用することで，ユーザが視認しやすくなることが期待できるが，明度や色相など，視覚的変数として利用可能な色の属性は複数存在し，またユーザにとっても視認しやすさが異なる可能性がある. そこで，本稿では色やユーザ毎の特性の違いについて考察するために予備実験を行う．

予備実験に用いたインタフェースのスクリーンショットを図 1 に示す．4 節に後述するように，予備実験は二種類行っているが，図 1 右の確認ボタンの有無の違いと，本務に相当するアプリケーションと同一の PC 上で動作するかの違いがある他は，両実験で用いたインタフェースは同様のものである．

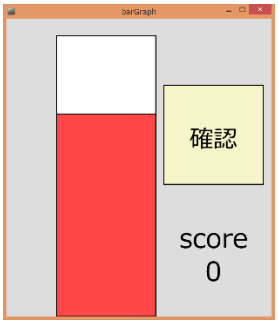


図 1. 実験に用いたインタフェース

図 1 の棒グラフはデータの蓄積量を可視化している．あらかじめデータ数の上限を設定しておき，上限の値で棒グラフの量が最大となるようにマッピングしている．インタフェース上の確認ボタンをクリックするか， Space キーまたは Enter キーを押すことで，データの確認を行ったとみなし，蓄積量を 0 に戻す．本稿での予備実験では，実際のストリームデータではなく人工的なデータの蓄積量を反映した．詳細については 4 節にて後述する．

本稿の実験で視覚的変数として用いている色のパターンは図 2 に示す 5 種類である．棒グラフの量の増加に従い，棒全体の色が変化する．

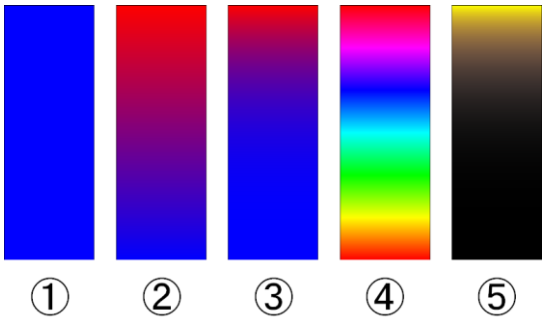


図 2. 実験で視覚的変数として用いた
棒グラフの色パターン

①は青単色であり，色が変化しない場合のサンプルとして利用する．②は青から赤へ等比率で変化する．③は②と同様に青から赤への変化であるが，棒グラフ上方部で大きく変化するように調整している．④は色相環を一周するパターンである．⑤は明度の変化である．③と同様に，棒グラフ上方部で大きく変化するように調整している．②と③では，青と赤は信号など一般的に利用される色であるため利用した．⑤の明度は大きさを表現するのに適している[9]ことから利用した．また，今回の実験で行うタスクにおいては，棒グラフの高さを判断の手がかりにするだけでなく，特定の色を手がかりとする可能性も存在するため，違いを認識しやすい色相を④で利用している．

本稿での実験では，実験協力者にデータの最大値になるべく近い所でリセットしてもらうための動機づけとして，スコアを導入している．スコアはリセットを行うたびに加算される．リセット時の棒グラフの高さを，最大値に対する割合 (%) で求め，その値を二乗したものをスコアとする．例として，棒グラフの高さが 50%の時にリセットを行った場合 $50^2 = 2500$ 点が加算される．ただし，リセット前に最大値を超えた場合は強制的にリセットされ，スコアとして-10000 が加算される．

4. 予備実験

4.1. 実験概要

20 代の工学系大学生，大学院生を対象に，前節で述べたインタフェースを利用した実験を実験 (1)，(2) の 2 回に分けて行った．実験 (2) では実験 (1) で得られた知見や反省に基づきタスクや設定などを変更している．

本実験の目的は，確認タイミングの判断支援のために用いる棒グラフのような，直接数値によっての表現ではなく量や色を用いた表現に対する人の認知特性の違い，および色に関するマッピングが判断に与える影響について調査することである．

本実験では，本務に相当する作業としてタイピングゲームを用いる．また，棒グラフにマッピングする人工的データの増加パターンとして，以下の 4 種類を用いる．データ数の上限は 10000 とする．

P1: 一定の速度で上昇変化する．
P2: 一定の速度で上昇変化する．P1 よりも高速．
P3: ランダムな速度で上昇変化する．
P4: ランダムな速度で上昇変化する．P3 よりも高速．

4.2. 実験 (1) 概要

本実験は，20 代の工学系大学生，大学院生 12 人を対象に行った．

実験の概要は以下のとおりである．

I. タイピングゲーム『寿司打 [5]』を難易度『高級コース : 練習』で行ってもらい，結果を記録する．
II. 図 1 に示す棒グラフ形式のメータを起動する．タイピングゲーム『寿司打』難易度『お手軽コース : 練習』を行いながら，棒グラフの挙動，リセット動作の確認を一度だけ行ってもらう．

[5]http://typing.sakura.ne.jp/sushida/

III. タイピングゲーム『寿司打』難易度『高級コース:練習』を行いながら，棒グラフ形式のメータのリセットを行ってもらい，その結果を記録する．
IV. 実験中，リセットの際に意識していたことについてのアンケートを実施する．

タイピングゲームの実行時間は一回あたり約 2 分である．実験に用いたデータ増加パターン (P1～P4) ，色パターン (図 2 の①～⑤) の組み合わせを表 1 に示す．実験協力者がリセット動作を行うたびに，実行順序に示す順番で組み合わせが変化する．

表 1. データ増加パターン，色パターンの組み合わせと実行順序

		実験協力者 (A～L)			
		A,E,I	B,F,J	C,G,K	D,H,L
実行順序	1	P1, ①	P2, ②	P3, ③	P4, ④
	2	P3, ②	P4, ①	P1, ④	P2, ③
	3	P4, ③	P3, ④	P2, ①	P1, ②
	4	P2, ④	P1, ③	P4, ②	P3, ①

4.3. 実験 (1) 結果

表 2 は，各色パターンでのデータの蓄積量を実験協力者毎に示している．4.1 節で述べたとおり，データ数の上限は 10000 であるため，データ蓄積量の最大値も 10000 となる．表 2 において 10000 と示されているものは，リセットをせずに最大値を超えてしまったものである．

表 2. 実験 (1) 結果

		色パターン			
		①	②	③	④
実験協力者	A	9150	9170	9450	8908
	B	8884	8550	8901	9000
	C	9300	9658	9650	9314
	D	9917	9600	9718	9750
	E	9100	10000	8700	8796
	F	8281	7950	8741	8600
	G	9600	9755	9150	8880
	H	9313	9500	9421	8850
	I	8800	7944	8100	7914
	J	10000	9450	9201	8450
	K	9150	7976	8900	9610
	L	8941	9400	8378	8700

データ蓄積量を実験協力者，色パターン毎に箱ひげ図で表した結果を図 3，図 4 に示す．

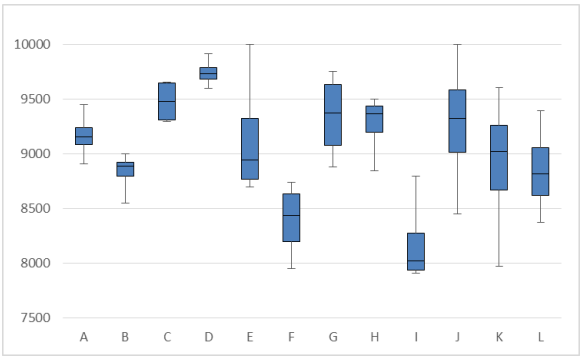


図 3. 実験協力者毎のデータ蓄積量の分布

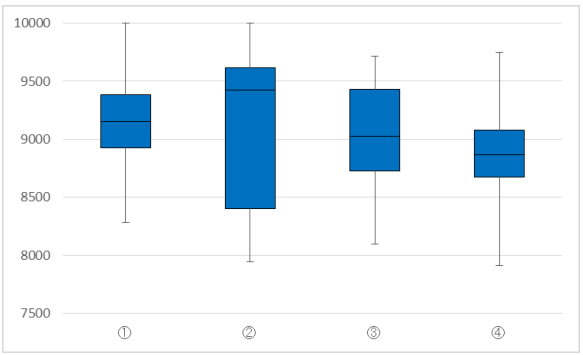


図 4. 色パターン毎のデータ蓄積量の分布

実験協力者間のデータ蓄積量の比較では，優位水準 5%で有意な差が見られたが，色パターン間では有意な差は確認されなかった．

アンケートの結果，棒グラフの色については全く意識していない実験協力者が多数であった．色パターンをデータ量の変化の確認に用いている実験協力者も存在したが，色パターンだけを基準にリセットを行う実験協力者は存在しなかった．

以上の結果より，データ蓄積量に対する確認タイミングはユーザによって異なると考える．

4.4. 実験 (2) 概要

実験 (1) では，色パターン毎に一回ずつ実験を行ってもらっていた．これに対し，実験 (2) では，実験協力者毎に同じ色パターンで複数回実験を行ってもらい，リセット時データ蓄積量の平均やばらつきについて検証する．20 代の工学系大学生，大学院生 8 人を対象に行った．

実験の概要は以下のとおりである．

I. タイピングゲーム『寿司打』を難易度『高級コース:練習』で行ってもらい，結果を記録する．
II. 図 1 に示す棒グラフ形式のメータを起動する．タイピングゲーム『寿司打』難易度『お手軽コース：練習』を行いながら，棒グラフの挙動，

リセット動作の確認を一度だけ行ってもらう.

III. タイピングゲーム『寿司打』難易度『高級コース：練習』を行いながら，棒グラフ形式のメータのリセットを行ってもらい，その結果を記録する．この手順は色パターン①で行う.

IV. 色パターンを④あるいは⑤に変更し，手順Ⅲと同様の実験を行ってもらう.

V. 実験中リセットの際に意識していたこと，色に対して意識していたことについてのアンケートを実施する.

ⅢおよびⅣに関して，実験に用いたデータ増加パターン，色パターンの組み合わせを表 3 に示す. 実験協力者がリセット動作を行うたびに，実行順序に示す順番でデータ増加パターンが変化する.

表 3. データ増加パターン，色パターンの組み合わせと実行順序

		実験協力者 (A～H)		
		全員 (Ⅲ)	A,B,E,F (Ⅳ)	C,D,G,H (Ⅳ)
実行順序	1	P1, ①	P1, ④	P1, ⑤
	2	P3, ①	P3, ④	P3, ⑤
	3	P2, ①	P2, ④	P2, ⑤
	4	P4, ①	P4, ④	P4, ⑤

4.5. 実験 (2) 結果

一回のタイピングゲームの間に，すべての実験協力者は 4 回のリセットを行った. 実験協力者 A,B,E,F の色パターン毎のリセット時のデータ蓄積量の平均，標準偏差を表 4 に示す．また，実験協力者 C,D,G,H の実験結果を同様に表 5 に示す.

表 4. 実験 (2) 実験協力者 A,B,E,F の各色パターンでの平均，標準偏差

			色パターン	
			①	④
実験協力者	A	平均	9459.25	9756.25
		標準偏差	147.14	197.09
	B	平均	9922.50	9875.00
		標準偏差	106.18	107.61
	E	平均	9645.75	9397.25
		標準偏差	264.31	208.30
	F	平均	9422.25	9644.75
		標準偏差	424.99	62.57

表 5，実験 (2) 実験協力者 C,D,G,H の各色パターンでの平均，標準偏差

			色パターン	
			①	⑤
実験協力者	C	平均	9448.00	9407.75
		標準偏差	209.12	212.75
	D	平均	9776.25	9816.00
		標準偏差	105.86	36.18
	G	平均	9908.75	9835.75
		標準偏差	106.96	200.75
	H	平均	8201.50	8939.00
		標準偏差	412.51	242.34

実験協力者間でのリセット時のデータの蓄積量に関して，どの色パターンでも優位水準 5%で統計的に有意な差が見られた．色パターン間の比較では，実験協力者毎の平均，および標準偏差いずれについても有意な差は確認されなかった.

アンケートの結果，④の色相環のように色が頻繁に変化すること，①の青色のように元々の彩度が高いものが動くことでタイピングゲーム中に頻繁に棒グラフに注意が向かってしまい，集中できなかったという実験協力者が存在した.

以上 2 つの実験結果より，ユーザ毎に棒グラフに対する確認タイミングの意識は異なり，ユーザ毎に可視化の調整を行う必要があると考える.

一方，色の変化は確認タイミングの変動の要因になりうる可能性があるが，ユーザによって反応や好みが異なり，過度な色の変化は本務に悪影響を与えてしまう可能性があると考える.

5. おわりに

本稿では，オンラインニュース等のストリームデータを確認するタイミングの判断を支援するシステムの構築に向けて，棒グラフ形式のメータを利用したユーザ実験を行った結果について報告し，棒グラフの色や個人差の影響について考察した.

ユーザ実験により，ユーザによって棒グラフの量に対する認知特性に有意な差がある結果が得られた. そのため，ユーザ毎に可視化の調整を行う必要があると考える.

一方，色の違いは確認タイミングの手がかりになるとは限らず，過度な色の変化，彩度の高い物体の変化は注意を集め，本務の集中を妨げる可能性があることがわかった．今後は，ユーザ毎の確認タイミングの差を調整するための手法の考案，可視化に用いる色の再検討を行い，インタフェースの開発を進める予定である.

参考文献

[1] M. Czerwinski, E. Horvitz, and S. Wilhite: A diary study of task switching and interruptions, 2004 Conference on Human Factors in Computing Systems (CHI'04) , pp. 175-182 (2004)

[2] B. O'Conaill, D. Frohilch: Timespace in the workplace: dealing with interruptions, 1995 Conference on Human Factors in Computing Systems (CHI'95), pp. 262-263 (1995)

[3] 卓　璐, 王　琛, 浅井　洋樹, 山名　早人 : 3 軸加速度を用いたデスクワーク中の割り込み可能性の推定, DEIM Forum 2015, E1-5 (2015)

[4] 谷　尭尚, 山田　誠二 : 机上にかかる圧力を用いたユーザの割り込み可能性推定, 人工知能学会論文誌, vol.29, No1, pp. 129-136 (2014)

[5] 田中　貴紘, 藤田　欣也: オフィスワーカーの状況推定―割り込み拒否度を中心に―, 電子情報通信学会誌, Vol. 95, No. 5, pp. 457-460 (2012)

[6] 沼野　航希, 高間　康史 : オンラインニュースを対象としたモニタリングシステムの提案, 第 8 回インタラクティブ情報アクセスと可視化マイニング研究会, pp.18-23 (2014)

[7] Y. Takama, M.Okumura: Interactive Visualization System for Monitoring Support Targeting Multiple BBS Threads, International Journal on Intelligent Decision Technologies, (DOI) 10.3233/IDT-140232 (2014)

[8] Y.Takama, T.Kurosawa: Visualization System for Monitoring Bug Update Information, Trans.IEICE, Vol.E97-D, No.4, pp. 654-662 (2014)

[9] R. Mazza（著), 加藤　諒（編集), 中本　浩（翻訳）, 情報を見える形にする技術, pp. 45-47 (2011)

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11回)
SIG-AM-11-08

文書ストリームにおけるトピックダイナミクスの階層化ビジュアリゼーション

Hierarchical Visualization System of Topic Dynamics in Document Stream

澤井裕介 [1*]　熊野雅仁 [2]　木村昌弘 [2]
Yusuke Sawai[1]　Masahito Kumano[2]　Madahiro Kimura[2]

[1] 龍谷大学大学院理工学研究科電子情報学専攻
[1] Division of Electronics and Informatics, Ryukoku University
[2] 龍谷大学理工学部電子情報学科
[2] Department of Electronics and Informatics, Ryukoku University

Abstract: ソーシャルメディアの発達により WEB 上に大規模文書ストリームが多数出現しており，それらをわかりやすく整理，説明することが強く求められている．近年，文書ストリームにおけるトピックの融合・分離に着目したトピックダイナミクスが注目されている．しかし，従来は，トピックの活性度が考慮されていなかった．本研究では，新聞データを用いて日々のトピック間の関係や活性度の変化を視覚的に分析できる TimeLine を用いた階層的可視化システムを提案する．

1 はじめに

日々刻々と変化する世界の情勢を把握することは，社会を生きる人々の重要な関心の対象といえる．これまで，世界の情勢は，新聞，ラジオ，テレビなどの主要メディアに携わる人々によって，限られた紙面，時間制限の中で情報の取捨選択せざるを得ない状況があった．しかし，近年，WEB 世界の発展により，世界情勢を伝える発信源は無尽蔵に増え続け，主要メディアで取り上げられなかった情報はもとより，ソーシャルメディアの発達により，これまで主要メディアの情報を知る側であった人々が意見を述べ，情報の発信源としても成長していることから，WEB 空間に文書ストリームが無数に出現している．また，ソーシャルメディアが報じる無数の情報は，主要メディアの報じる内容にまで影響を及ぼし始めているため，主要メディアやソーシャルメディアで生み出される無数のトピック間相互に，複雑に影響し合う関係が成立しており，ダイナミクスが存在し得ると予想される．

そのようなトピックのダイナミクスを捉えるためには，観測できる対象として，複数の文書ストリーム間の関係を捉え，それらをわかりやすく整理，分析できる環境を構築することが望ましいと思われる．ただし，文書ストリームにおけるトピックは一定ではなく，日々，変容しており，ひとつのトピックが分離して発展・活性化したり，複数のトピックが融合して活性化するなど，様々な変化・変動が起きている可能性がある．近年，トピックの動的な変化を扱う研究 [1] や，時間軸上で前後に位置するトピック同士の依存関係響に着目して時間展開を捉える研究 [2]，トピックの発生と消滅を捉える研究 [3] や，さらにはトピックの分離・融合を扱う研究 [4] など，トピックの時間依存性やトピック相互の関係に着目した研究が注目されている．

ただし，生活や文化，政治や経済などの一般的なトピックには，政治の場合，例えば税金，外交，防衛など，大局的に安定して存在し続けるトピックが存在する．これらのトピックは，完全に消滅するとは考えにくいと思われる．一方，沖縄問題，尖閣諸島，オスプレイなどは，異なるトピックと言えるが，日々，活発に話題になったり，沈静化したりと，変動する傾向があるだけでなく，安定して存在する防衛トピックや，外交トピックのいずれとも関係がある．ここで，一年間で安定して存在する話題を年間主要トピック，日々，変動の大きい話題をデイリートピックとしたとき，トピックを階層的に捉えつつ，時間変化を捉える方法が考えられる．その観点において，デイリートピックは，無から発生したり，完全に消滅するのではなく，安定した複数の年間主要トピックと影響し合いながら，活性度が高まったり，沈静化しているにすぎないと考えることができる．また，デイリートピック同士も，相互に影響を及ぼし合いながら，分離して発展したり，相

*連絡先： 滋賀県大津市 瀬田大江町横谷 1-5
龍谷大学大学院理工学研究科
E-mail:t14m009@mail.ryukoku.ac.jp

互の依存関係に応じて融合するようなトピックであると考える．我々は，複数存在する文書ストリームに対して，安定した年間主要トピックと，変化・変動が起きやすいデイリートピックの関係を，包括的に捉えることでトピックダイナミクスを分析できる可能性に期待している．

そこで本研究では，生活や文化，政治や経済などの主要なトピックが含まれる新聞データを用いて，トピックを階層的に捉え，年間主要トピックとデイリートピックとの関係を可視化しつつ，デイリートピック間の関係や活性度の変化を時間軸に沿って視覚的に分析できる TimeLine を用いた階層的可視化ビジュアリゼーション法を提案する．本稿では，本研究の第一歩として，主要メディアのトピックを階層的に Timeline 上で捉えた際の視覚的分析に関する可能性を探るため，毎日新聞データセットを用いた実データによる実験で，提案システムの有効性を示す．

ただし，この方法では，トピックが分離や融合して話題が活性化したのか，それとも沈静化したのかなどを分析することができない．しかし，もし分離・融合とともに活性度がわかれば，より詳細に変動の様子を分析できる可能性がある．また，年間を通して安定して存在する年間主要トピックでも，活性度は変化している可能性があり，デイリートピックとの関係度の強さも変動している可能性がある．このため，年間主要トピックがどの日に活性化しているかや，年間主要トピックとデイリートピックの関係を可視化できれば，さらに詳細に変動分析が可能になることが期待される．本研究では，ダイナミクスの一面として，トピックの分離・融合だけでなく，活性度を可視化し，さらに年間主要トピックとデイリートピックとの関係度を可視化することで，文書ストリームのダイナミクスにおいて，複数の面から変動を分析することができるトピックダイナミクスの階層化ビジュアリゼーション法を提案する．

2 提案法

本稿では、文書群の時系列データ（文書ストリーム）として 1 年間の新聞記事群を考え、そのトピックダイナミクスの可視化法を提案する．

2.1 入力データ

ある年の新聞記事の全体（文書ストリーム）

$$\mathcal{D} = \bigcup_{t=1}^{T} D_t$$

を入力データとする．ここに、T は文書ストリームの総日数 (365 または 366) であり、D_t は第 t 日における記事全体の集合

$$D_t = \{d_{t,n} \,|\, n = 1, \cdots, N_t\} \ (t = 1, \cdots, T)$$

である．ただし、$d_{t,n}$ は第 t 日の第 n 記事であり、N_t は第 t 日における記事の総数である．各記事 $d_{t,n}$ は、形態素解析を行い、単語頻度ベクトル

$$\boldsymbol{x}_{t,n} = (x_{t,n,1}, \cdots, x_{t,n,V})$$

により BoW(bag-of-words) 表現する．ここに、各 $x_{t,n,i}$ は、想定する語彙集合 $\{voc_1, \cdots, voc_V\}$ に対し、文書 $d_{t,n}$ における語彙 voc_i の出現回数である．V は、想定する語彙の総数である．

2.2 年間主要トピックスの抽出

文書ストリーム $\mathcal{D}$ における年間レベルでの主要トピックの出現と消滅のダイナミクスを調べるために、文書群データ $\{D_t \,|\, t = 1, \cdots, T\}$ を多重トピックを考慮した文書の確率的生成モデルである HDP-LDA(hierachical Dirchlet Process - Latent Dirichlet Allocatio)[5][1] によりモデル化する．

各 t に対して、第 t 日の記事群 D_t を、それに属するすべての記事を単純につなぎ合わせて一つの長い文書と考え、単語頻度ベクトル

$$\boldsymbol{X}_t = (X_{t,1}, \cdots, X_{t,V})$$

により BOW 表現する．ここに、各 $X_{t,i}$ は

$$X_{t,i} = \sum_{n=1}^{N_t} x_{t,n,i}$$

である．そして HDP-LDA モデルに基づいて、観測データ $\{\boldsymbol{X}_t \,|\, t = 1, \cdots, T\}$ に対する、潜在トピック集合

$$Y = \{y_1, \cdots, y_L\}$$

および、各潜在トピック $y \in Y$ の下での単語生成ベクトル

$$\boldsymbol{\theta}_y = (\theta_{y,1}, \cdots, \theta_{y,V})$$

を、それぞれ推定する．我々は、各 $y \in Y$ を文書ストリーム $\mathcal{D}$ の**年間主要トピック**と呼び、それら年間主要トピックスを抽出し分析する．

まず、文書ストリーム $\mathcal{D}$ の年間主要トピック $y \in Y$ が、第 t 日にどのくらい活発であったかを、事後確率を用いて、

$$f_y(t) = P(y \,|\, \boldsymbol{X}_t)$$

により測定する．我々は、$f_y(t)$ を年間主要トピック y の第 t 日における**活性度**と呼ぶ．t に関する $f_y(t)$ の変

動を調べ、年間主要トピック y のダイナミクスを分析する.

また、各 $y \in Y$ に対し、潜在トピック y の下での単語生成ベクトル $\boldsymbol{\theta}_y$ においてランキングを行うことによって y と関係がより深い単語を抽出することにより、年間主要トピック y を説明する.

2.3 デイリートピックスの抽出

文書ストリーム $\mathcal{D}$ における日レベルでのトピックについて調べるために、各 t に対して、第 t 日の文書群 $D_t = \{d_{t,k} \,|\, k = 1, \cdots, K_t\}$ を多重トピックを考慮した文書の確率的生成モデルである HDP-LDA によりモデル化する. そして、HDP-LDA モデルに基づいて、観測データ $\{\boldsymbol{x}_{t,k} \,|\, k = 1, \cdots, K_t\}$ に対する、潜在トピック集合

$$Z_t = \{z_{t,1}, \cdots, z_{t,K_t}\}$$

および、各潜在トピック $z \in Z_t$ の下での単語生成ベクトル

$$\boldsymbol{\phi}_{t,z} = (\phi_{t,z,1}, \cdots, \phi_{t,z,V})$$

を、それぞれ推定する. 我々は、各 $z \in Z_t$ を文書ストリーム $\mathcal{D}$ における第 t 日の**デイリートピック**と呼び、それらデイリートピックスを抽出し分析する.

まず、任意の t に対して、第 t 日の各デイリートピック $z \in Z_t$ を次の 2 つのやり方で説明する.

1. 事後確率 $P(z \,|\, \boldsymbol{x}_{t,k})$ が高い記事 $d_{t,k}$ を抽出する.
2. 単語生成ベクトル $\boldsymbol{\phi}_{t,z} = (\phi_{t,z,1}, \cdots, \phi_{t,z,V})$ において要素の値が大きい語彙 voc_i を抽出する.

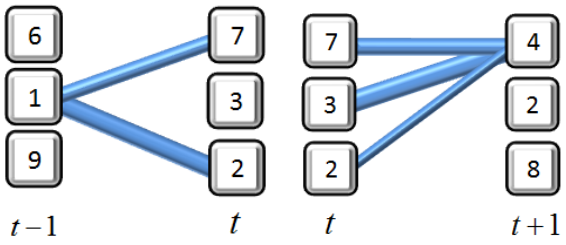
また、デイリートピック $z \in Z_t$ がどのくらい活発であったかを、事後確率を用いて、

$$f_z(t) = \sum_{n=1}^{N_t} P(z \,|\, \boldsymbol{x}_{t,n})$$

により測定する. 我々は、$f_z(t)$ をデイリートピック $z \in Z_t$ の**活性度**と呼び、デイリートピックスの活性度を分析する.

次に、第 t 日のデイリートピック $z \in Z_t$ がどの年間主要トピック $y \in Y$ と関係しているかを、単語生成ベクトル $\boldsymbol{\phi}_{t,z}$ と $\boldsymbol{\theta}_y$ のコサイン類似度で測定する. 我々は、その値をデイリートピック $z \in Z_t$ と年間主要トピック $y \in Y$ の**関係度**と呼び、デイリートピックスと年間主要トピックスの関係度を分析する.

次に、第 t 日のデイリートピック $z \in Z_t$ が、翌日である第 $t+1$ 日のデイリートピック $z' \in Z_{t+1}$ とどのように関係しているかを、単語生成ベクトル $\boldsymbol{\phi}_{t,z}$ と $\boldsymbol{\phi}_{t+1,z'}$ のコサイン類似度で測定する. 我々は、その値をデイリートピック $z \in Z_t$ とデイリートピック $z' \in Z_{t+1}$ の**関係度**と呼び、隣接する日々のデイリートピックス間の関係度を分析する.



(a) 分離　　(b) 融合

図 1: デイリートピックの分離と融合

2.4 トピックのダイナミクス

1 年間の文書ストリーム全体に対し，HDP-LDA[5][1] を適用すると，年間を通じた潜在トピックが得られる. ただし，この年間を通じた潜在トピック（年間主要トピック）は，年間を通じて安定して存在するトピックの一面を捉えていると考えられるものの，日々の変動を捉えてはおらず，この年間主要トピックだけに着目してもダイナミクスを捉えることはできない.

一方，一日ごとに多数の情報源（新聞であれば多数の記事）に対して HDP-LDA を適用すると，一日ごとの潜在トピック（デイリートピック）が得られる. デイリートピックは，日々変動する話題内容の変容を捉えている可能性があるだけでなく，日ごとに変化し得る潜在トピック数の変動も捉える可能性がある. 図 1 は，ある日に得られたデイリートピックを番号で表し，次の日に得られたデイリートピックとの関係を線で結んだ様子を示しているが，線の太さに関係の強さを割り当てている例である. このような潜在トピックの可視化を実現すれば，図 1(a) のように，第 $t-1$ 日のデイリートピック $z_{t-1,1}$ が，第 t 日のデイリートピック $z_{t,7}$ と $z_{t,2}$ に分かれるようなトピックの分離を捉え得る可能性がある. また，図 1(a) のように，第 t 日のデイリートピック $z_{t,7}$, $z_{t,3}$, $z_{t,2}$ が，第 $t+1$ 日のデイリートピック $z_{t+1,4}$ に集中するようなトピックの融合を捉える可能性もあるため，デイリートピック間のダイナミクスを捉える可能性がある.

3 提案システムデザイン

提案システムの概観を図 2 に示す. 図 2 のように，提案システムは，四つの View で構成される. 基本となる View A は，年間主要トピック $y \in Y$ の全体をタイムライン上に可視化するものである. View B は，View

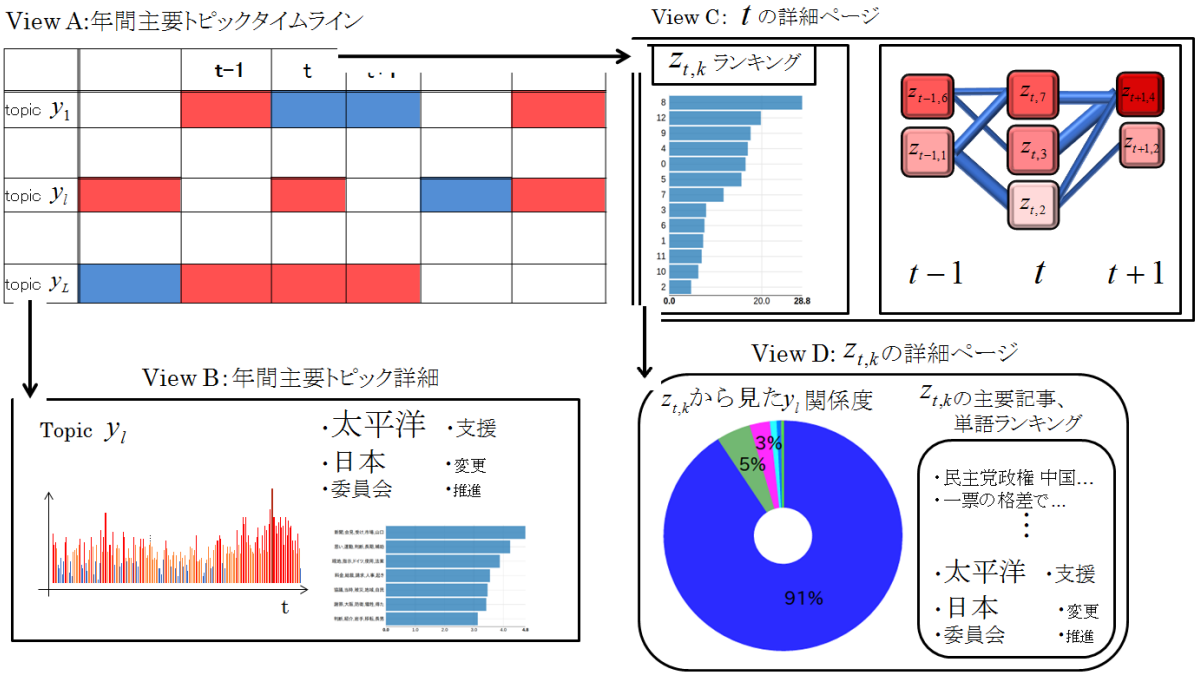


図 2: 四つの View で構成される提案システムの概観

A に表示された一つの年間主要トピック $y \in Y$ の詳細を可視化するものである．View C は，View A に表示されたある日 t に対応する活性度の高いデイリートピックと，日 $t-1$ と日 $t+1$ に対応する活性度の高いデイリートピックとの関係度の強さを可視化するものである．View D は，個々のデイリートピックで出現確率の高い単語のランキング上位や，デイリートピック $z \in Z_t$ と年間主要トピック $y \in Y$ の関係度の強さを可視化するものである．次に，これら四つの View について，個々に詳細を説明する．

3.1 年間主要トピックタイムライン:View A

提案法では，年間主要トピック $y \in Y$ に関して，日単位で活性度 $f_y(t)$ を算出することができる．活性度は，高い状態を赤，中間を白，低い状態を青で表す．図 2 の View A は，その可視化の様子を示したものである．この View A により，ユーザは，どの年間主要トピックがいつ活性化しているかを確認することができる．

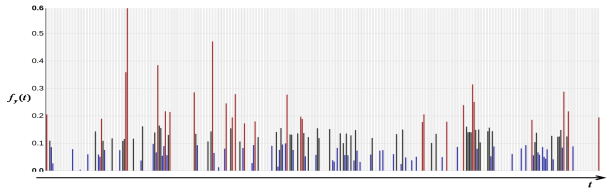


図 3: 年間主要トピック $y \in Y$ における各日の活性度

また，View A 上の年間主要トピック名を押すと，View B により，年間主要トピックの詳細を確認することができ，View A 上の日名を押すと，日 t に含まれるデイリートピックに関する詳細を知ることができる．

3.2 年間主要トピックの詳細：View B

一つの年間主要トピックに焦点を当て，より詳しい情報を提示する View である．より詳しい情報として，横軸を日とした一年，縦軸を活性度とした図 3 のグラフを通じて，活性度のより細かい変化を確認することができる．

また，図 2 の View B において，右上部にフォントサイズの異なる単語を確認することができる．これは，年間主要トピック $y \in Y$ において，Bag of words 表現された単語の出現確率の高さをフォントサイズの大きさで表現したものである．これにより，年間主要トピック $y \in Y$ と関連の高い単語のランキング上位を確認することができる．

また，図 2 の View B において，右下部にある棒グラフを拡大したものが図 4 である．これは，年間主要トピック $y \in Y$ とある日 t のデイリートピック $z \in Z_t$ との関係度ランキングを表しており，ランキングの上位が降順に整列されている．これにより，View B を用いて注目している年間主要トピック $y \in Y$ が，ある日 t のデイリートピック $z \in Z_t$ とどのような関係にあるかを確認することができる．また，図 4 のように，各

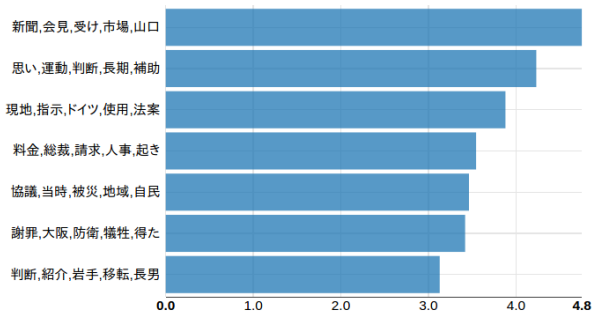


図 4: 年間主要トピック $y \in Y$ からみたデイリートピックス Z_t の関係度ランキング

デイリートピック $z \in Z_t$ の棒グラフの左側には単語生成ベクトル $\boldsymbol{\phi}_{t,z} = (\phi_{t,z,1}, \cdots, \phi_{t,z,V})$ において要素の値が大きい語彙 voc_i の Top 5 が表示されている．これにより，活性度の高い $z \in Z_t$ が，どのような単語に強い関係を示すかがわかり，その $z \in Z_t$ が表す潜在的トピックの内容を調べる足がかりとなる．

3.3 第 t 日の詳細：View C

図 2 の View C では，View A に表示された第 t 日のデイリートピックス Z_t に関する詳細が確認できる．図 2 の View C の左側にある棒グラフを拡大した一例が図 5 である．これは，第 t 日に抽出されたデイリートピック $z \in Z_t$ を活性度 $f_z(t)$ によってランキングを行い，降順に可視化したものである．これにより，各デイリートピック $z \in Z_t$ の活性度の値がわかり，活性度のランキングもわかるため，どのデイリートピックがどの程度活性化しているかを確認することができる．

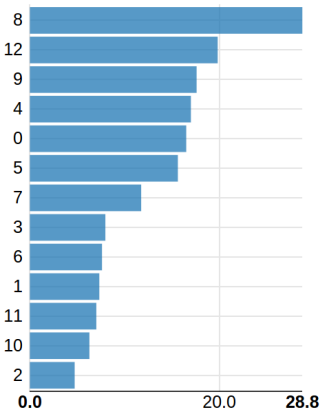


図 5: デイリートピック $z \in Z_t$ の活性度ランキング

また，図 2 の View C の右側では，第 t 日だけでなく，第 $t-1$ から第 $t+1$ 日までのデイリートピック間の関係度の変化がわかるだけでなく，日ごとにデイリートピック Z_t の活性度に関するランキング情報が可視化されるため，デイリートピック間の関係度と活性度の変化を同時に分析することができる．この可視化については，3.5 で説明する階層的トピックダイナミクス分析と関連が深いため，3.5 で詳細に説明を行う．

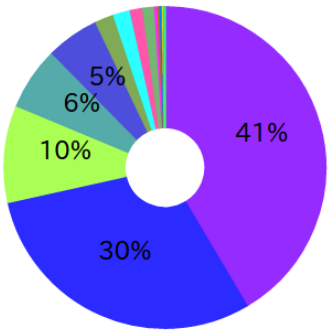


図 6: デイリートピック $z \in Z_t$ と年間主要トピックス Y との関係度

3.4 各デイリートピックの詳細：View D

図 2 の View D は，View C において，一つのデイリートピックの詳細情報を見るため，注目するデイリートピック $z \in Z_t$ を押した際に表示されるものである．図 2 の View D の右側には，注目するデイリートピック $z \in Z_t$ での voc_i を確認することができる．これにより，デイリートピックの内容を調べる足がかりとなる．また，図 2 の View D の左側にある円グラフを拡大したものが図 6 である．図 6 は，各デイリートピック $z \in Z_t$ と各年間主要トピック $y \in Y$ の関係度を円グラフで表現したものである．これにより，デイリートピックがどの年間主要トピックと関係が強いかを知ることができる．また，異なるデイリートピックの円グラフを図 2 の View C の右側の部分で同時に表示することができる．この機能を用いれば，図 7 のように，表示することで，数日間のデイリートピック間の関係度の変化や活性度の変化を同時に分析できるだけでなく，さらに，個々のデイリートピック $z \in Z_t$ と年間主要トピック $y \in Y$ との関係度を円グラフを通じて確認することができるため，デイリートピックの分離・融合に関するダイナミクスだけでなく，階層的に活性度や関係度のダイナミクスを分析できる可能性がある．

3.5 階層的トピックダイナミクス分析

図 7 は，図 2 の View C 右側にある数日間のデイリートピックタイムラインにおいて，いくつかのデイリートピックに関する年間主要トピックとの関係率を示す円グラフを同時に可視化した例である．

この可視化により確認できることは，まず，第 $t-1$ 日から第 $t+1$ 日までの各日に抽出されたデイリートピック $z \in Z_t$ の個数が変化する様子である．図 7 にお

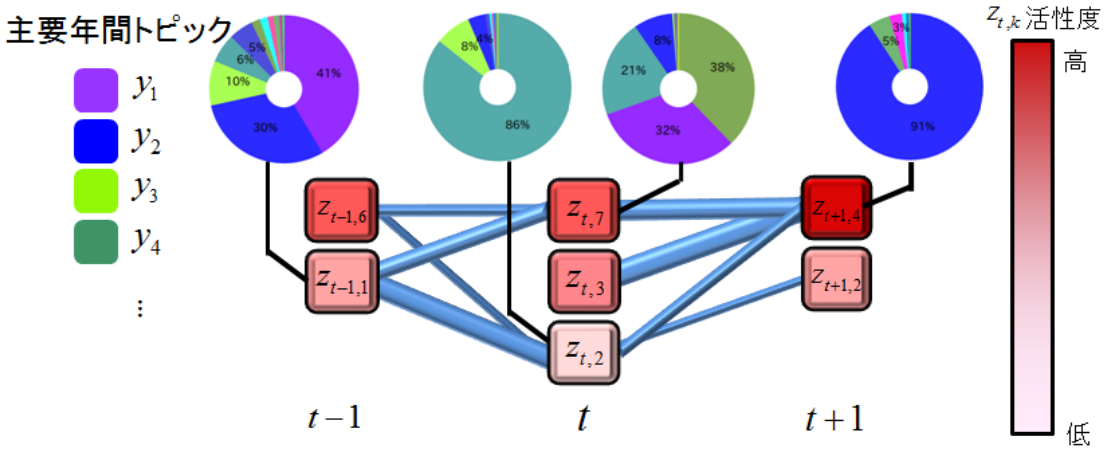


図 7: デイリートピックのタイムラインを用いた階層的トピックダイナミクス分析

いて，第 $t-1$ 日には二つのデイリートピック z があり，第 t 日には三つのデイリートピック，第 $t+1$ 日には二つのデイリートピックがあることが見てとれる．また，各デイリートピックは，日ごとに活性度ランキングに基づいて縦方向に降順で整列されているため，各日の最も上に位置する $z \in Z_t$ が，その日に活性度が最も高いデイリートピックとなる．さらに，各 $f_z(t)$ は，図 7 右にあるような白から赤の色で着色されているが，これは，活性度の値を示している．これにより，たとえ，ある日の活性度が最も高いデイリートピックであっても，他の日と比較すると，活性度に差があることを視認できる．例えば，第 $t-1$ 日では，$z_{t-1,6}$ が最も活性度が高く，第 t 日では，$z_{t,7}$ 第 $t+1$ 日では，$z_{t+1,4}$ が最も活性度が高いデイリートピックであることが見てとれるが，第 $t-1$ 日の $z_{t-1,6}$ や第 t_2 日の $z_{t,7}$ よりも，第 $t+1$ 日では，$z_{t+1,4}$ の赤色の彩度が最も高いことが見てとれる．つまり，活性度の変化をより詳しく視認することができる．

次に，第 $t-1$ 日のデイリートピック $z_{t-1,1}$ は，第 t 日で活性度が最も高い $z_{t,7}$ と，活性度が低い $z_{t,2}$ に分離している可能性があることや，第 t 日の特に $z_{t,7}$ と $z_{t,3}$ が，第 $t+1$ 日では $z_{t+1,4}$ に融合していることがわかる．$z_{t+1,4}$ は，第 $t-1$ 日から第 $t+1$ 日の中で，最も彩度の高い赤を示していることから，何かが起きている可能性を期待させる．このように，活性度を用いると，分離や融合の観点だけでなく，より詳しくデイリートピックの変化を捉えることができる可能性がある．

さらに，図 7 では，デイリートピックの分離・融合に関係していると見なした第 $t-1$ 日の $z_{t-1,1}$，第 t 日の $z_{t,7}$ と $z_{t,2}$，第 $t+1$ 日の $z_{t+1,4}$ に関するデイリートピックと年間主要トピックとの関係率を示す円グラフを選択的に同時表示した様子を示している．円グラフの色は，図 7 左端にあるように，年間主要トピック

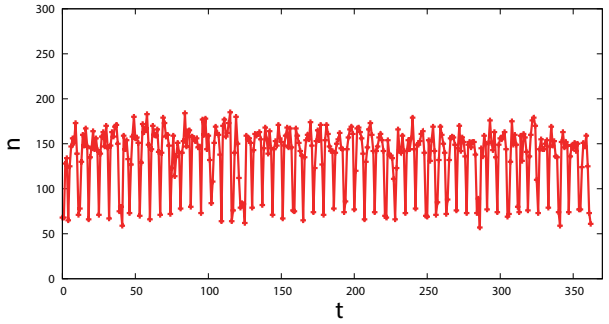


図 8: 各 t の記事数 n

$y \in Y$ を識別するために割り当てた色であり，図 7 の各円グラフの変遷から，分離・融合に携わるデイリートピックと関係度の高い年間主要トピックが変化している様子も視認できることがわかる．このように、提案可視化法では，年間主要トピックとデイリートピックを階層的に捉えながら活性度や関係度の動的な変化を視覚的に分析できることがわかる．

4 実施例

4.1 実験データ

本研究では実験データとして，毎日新聞データベースより，2013 年 1 月 1 日から 2013 年 12 月 31 日の新聞記事の中で 1 面，2 面，3 面，経済面，社会面の記事を文書ストリームとして使用した．1 日あたりの記事数 n のグラフを図 8 に示す．これによりおおよそすべての日において偏りなく記事が書かれていることが分かる．また，これらの記事において，1 日に 1 回以上出現している単語という条件のもと BOW 表現に変換した．その語彙の総数は，1,333 となった．

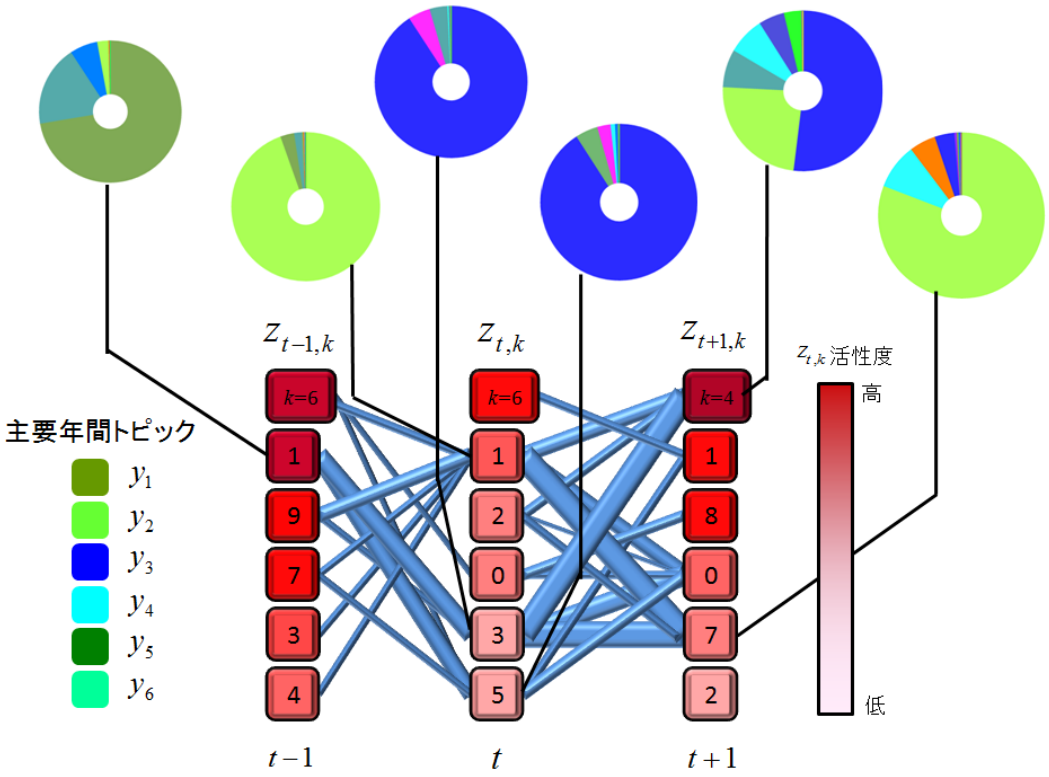


図 10: t=7/21 のデイリートピックタイムライン

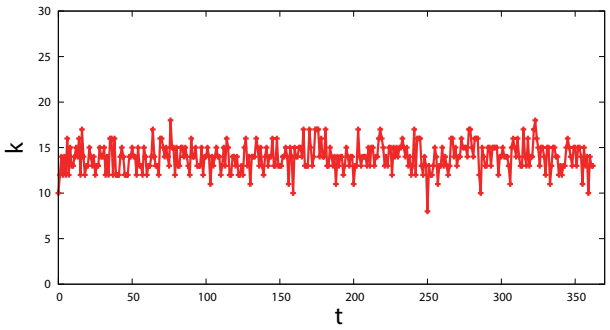


図 9: デイリートピック数 k の変動

4.2 実験

2013 年の実験データに対し，潜在トピック推定法を適用したところ，抽出された年間主要トピック $y \in Y$ の総数は 18 個となった．また，各 t 日において，デイリートピックを推定したところ，各 t 日のディリートピック数 K_t は，図 9 に示すような結果となった．図 9 の横軸は t となっており，縦軸は各 t 日のディリートピック数となっている．すべての日において，デイリートピックが一定数以上生成されていることや，日によって数が変動していることが確認できる．

2013 年のデータを用いて可視化システムの View A を表示させたのち，ディリートピックの詳細として，t=7 月 21 日を選択し，View C を表示した．このとき，階層的トピックダイナミクス分析が可能なディリートピックタイムラインを表示させたものが図 10 である．2013 年，7 月 21 日は，参院選挙の開票日であり，第 $t-1$ 日はその前日，第 $t+1$ 日は化参院選挙の開票日直後の日となる．図 10 より，第 t 日のディリートピック $z_{t,1}$ が第 $t+1$ 日のディリートピック $z_{t+1,4}$ と $z_{t+1,7}$ に分離している様子が見てとれる．ただし，ディリートピック間の関係度の強さが可視化されているだけでなく，活性度ランキングと活性度の値が可視化されているため，第 t 日のディリートピック $z_{t,1}$ は，活性度ランキング上トップで，活性度もかなり高い第 $t+1$ 日のディリートピック $z_{t+1,4}$ と結ばれている．しかし，第 t 日のディリートピック $z_{t,1}$ は，むしろ活性度ランキングが低めの第 $t+1$ 日のディリートピック $z_{t+1,7}$ と関連が強いことがわかる．ただし，第 t 日のディリートピック $z_{t,1}$ と第 $t+1$ 日のディリートピック $z_{t+1,7}$ は，活性度を示す色がほぼ同色であることから，活性度は変化していないが，$z_{t,1}$ や $z_{t+1,7}$ よりも，より活性度の高いディリートピック（例えば $z_{t+1,4}$）が第 $t+1$ 日に現れ，上位を占めてたと解釈できる可能性がある．

ところで，第 $t+1$ 日のディリートピック $z_{t+1,4}$ は，第 t 日のディリートピック $z_{t,1}$ と $z_{t,3}$ が融合したものであると解釈できる可能性がある．また，これらの $z_{t,1}$ と $z_{t,3}$ に関して，年間主要トピックとの関係率を示す

円グラフを表示させたところ，図 10 より，$z_{t,1}$ は，ほぼ黄緑の年間主要トピックと関係し，$z_{t,3}$ は，ほぼ青の年間主要トピックと関係していることがわかる．さらに，$z_{t,1}$ と $z_{t,3}$ が融合したと解釈した $z_{t+1,4}$ の円グラフを確認すると，その主要な年間主要トピックは，ほぼ黄緑の年間主要トピックと青の年間主要トピックであることが見てとれるため，$z_{t,1}$ と $z_{t,3}$ の融合したものが $z_{t+1,4}$ であるという解釈をより裏付けている可能性が期待される．ディリートピック $z_{t,1}$，$z_{t,3}$，$z_{t+1,4}$ のそれぞれと関連の深い第 $t+1$ 日の新聞記事をランキングしたことろ，以下のような記事と関連が高かった．$z_{t,1}$ は，エジプトでの武装勢力の攻撃に関する記事や，中国の影響力を懸念する記事，中国でのテロ行為．$z_{t,3}$ は，投開票日の話題や，与党と野党の攻防．$z_{t+1,4}$ は，自民圧勝，ねじれ解消，アベノミクスが指示されたという解釈が報じられた記事であった．この事例の解釈は，ユーザに委ねるものの，提案可視化法では，ディリートピックの分離・融合の観点だけでは捉えきれない，より詳細なトピックダイナミクスの分析が可能になる点で，提案法の有効性が示唆されていると思われる．

5 まとめ

本研究では，文書ストリームに対し，タイムライン上で，日々のデイリートピックの関係度を可視化することで，分離・融合するデイリートピック間のダイナミクスを分析するだけでなく，年間主要トピックとデイリートピックの階層的な関係を示しながら，トピックの活性度を可視化することで，より多くの観点からダイナミクスを視覚的に捉える文書ストリームのトピックダイナミクスにおける階層的ビジュアリゼーション法を提案した．提案法では，まず，文書ストリームに含まれる文書情報を 1 年単位で BOW 表現し，未知数の潜在的な多重トピックをパラメータ学習によって決定できる HDP-LDA を用いて推定する．ただし，1 年間すべての文書データに HDP-LDA を適用して年間主要トピックを推定し，日々の文書データに HDP-LDA を適用してデイリートピックを推定する．次に，年間主要トピック y に関して，第 t 日における活性度 $f_y(t)$ およびデイリートピック z に関して，活性度 $f_z(t)$ を求める．そして，年間主要トピックとデイリートピックの関係度を求める．このようにして得られた年間主要トピックとデイリートピックに関して階層的に文書ストリームにおけるトピックダイナミクスを活性度，関係度の観点から可視化し，視覚的分析が可能な環境を構築した．

実データとして，毎日新聞データベースを文書ストリームと見なし，階層化ビジュアリゼーション法を含む提案法の有効性を評価した．2013 年のデータを用いることにより，提案法では，まず，活性度をデイリートピックの活性度ランキング用い，さらに色を用いて活性度の値を可視化することで，デイリートピックの分離・融合を捉えるだけでなく，分離したデイリートピックが活性化したのか，沈静化したのか，また，融合したデイリートピックが活性化したのか，沈静化したのかという情報に加え，活性度の値はどの程度変化したのかがわかり，より詳しくデイリートピックの活性度の変化を分析できる可能性を示した．また，注目したデイリートピックの年間主要トピックとの関係率を示す円グラフを選択的，かつ複数同時に可視化することにより，分離・融合しているデイリートピックと，年間主要トピックの関係が，変わらない場合や変化することを確認することができ，文書ストリームのトピックダイナミクスを階層的に分析できる可能性も示し，提案法の有効性を示した．

参考文献

[1] D.M. Blei and J.D. Lafferty, “Dynamic topic models,” Proceedings of the 23rd International Conference on Machine Learning, pp.113–120, ICML ’06, ACM, 2006.

[2] A. Ahmed and E.P. Xing, “Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering,” Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA, pp.219–230, 2008.

[3] A. Ahmed and E.P. Xing, “Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream,” UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8-11, 2010, pp.20–29, 2010.

[4] 佐々木謙太朗 C 吉川大弘 C 古橋　武 C “複数のトピックの時間的依存関係を考慮した時系列混合モデル C” 人工知能学会論文誌 C vol.30Cno.2Cpp.466–472C2015D

[5] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, “Hierarchical dirichlet processes,” Journal of the American Statistical Association, vol.101, pp.1566–1581, Dec. 2006.

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

参照関係の可視化による論文サーベイの効率化

The Visualization of Citation Information and Its Application in Literature Survey

井上絢翔[1]　韓 東力[2]

Ayato Inoue[1]　Dongli Han[2]

[1]日本大学大学院 総合基礎科学研究科

[1] Graduate School of Integrated Basic Sciences, Nihon University

[2]日本大学文理学部 情報科学科

[2] Department of Information Science, College of Humanities and Sciences, Nihon University

Abstract: 論文サーベイは学術論文の執筆において重要なタスクの一つである。必要な論文を収集する方法はいくつか存在するが、その中でも我々は論文間の参照関係に注目し、「どのような理由で参照を行っているのか」といった情報を明らかにすることで効率的に論文サーベイを行うことができるのではないかと考えた。本研究では論文間参照関係の可視化を行うことにより論文サーベイの効率化を図った。

1. はじめに

論文を執筆する上で論文サーベイは必要不可欠である。その論文サーベイによく利用されているツールとして、CiNii[1]や Google Scholar[2]、CiteSeerX[3]などの電子図書館が挙げられ、これらを使った検索手法としてはキーワード検索や論文間参照関係に注目した探し方が挙げられる。

キーワード検索はキーワードとの照合により論文を検索する方法である。キーワード検索に関する既存研究は既にいくつもの既存研究が存在する[4][5]。

一方、論文間の関連性に着目して検索を行う方法としては、大きく分けて「共参照を利用した方法」と「過去の論文をたどっていく方法」が挙げられる。共参照とは、気になる論文(以下「起点論文」と呼ぶ)と同じ論文をなるべく多く参照している論文、もしくは同じ論文になるべく多く参照されている論文は関連度が高いのではないかという考え方で収集していく方法で、こちらも既にいくつもの既存研究が存在する[6][7]。

過去の論文をたどっていく方法は起点論文を 1 つ選定し、その論文が参照している論文や、さらに参照論文が参照している論文という順にサーベイの対象を広げていくという方法である。本研究ではこの「過去の論文をたどっていく方法」に焦点を当てていく。

関連度の高い過去の論文は起点論文が直接参照しているものだけとは限らない。例えば、起点論文がある論文Ａの手法を参考にしていたとする。その論文Ａがさらに別の論文Ｂの手法を参考にしていた場合、起点論文は間接的に論文Ｂの手法を参考にしていたといえる可能性がある。既存の電子図書館である CiNii[1]や Google Scholar[2]などでも、このように起点論文が直接参照していない論文も探索対象に含みたい場合は、気になる被参照論文を一つ選んで、それを起点論文として改めておいてさらにその被参照論文をたどっていくことはできる。しかし、この方法ではたどり着いた被参照論文を全て表示すると情報が多くなりすぎてしまう。実際に、論文の関係性に着目して可視化を行った研究として清水ら[8]や渡部ら[9]の研究が挙げられるが、どちらも論文の数が多すぎて見た目が乱雑になりすぎているという問題点を挙げている。そこで我々は論文間の関係性を明らかにし、これを利用して関連論文を絞り込むことで効率的な文献検索が可能になるのではないかと考えた。

論文間の関係性を明らかにすることができれば、「被参照論文の手法を改良して利用している文献が欲しい」や「被参照論文の実験結果と比較している論文を読みたい」などといったような検索が行えるようになり、多くの論文候補から検索目的にそぐわない論文をシャットアウトすることができるように

なる。このような検索が行えるような文献検索システムの構築が本研究の最終目標である。

2. 論文間の参照関係付与

このようなシステムを作るにはもちろん論文間の参照関係の付与が必要になる。その方法を大きく分けて機械的に参照関係の付与を行うものと、手動による付与が挙げられる。前者に関するものとして難波ら[10]、小出ら[11]と Teufel ら[12]の研究がある。難波らは論文間の参照タイプを 3 種類に分類しているが、効率的な論文サーベイを行うのに分類数が不十分と思われる。小出らと Teufel らの研究では、それぞれ 9 種類と 12 種類の参照理由を定義し機械学習を用いて参照理由を付与しているが、精度は最大で 60%～70%台に留まっている。論文間関係の解明を最終目標とするような研究では上記の精度でも一定の有効性があるかもしれないが、自動付与された参照理由を異なる目的で再利用する場合には、連鎖的誤りを回避するためにはより正確な分類結果が必要であろう。

それに対して手動付与ではより高精度のアノテーションを行うことができるが、時間がかかることや論文サーベイに精通している専門家を雇うのに多大なコストがかかることなどが問題点としてあげられる。

そこで我々はクラウドソーシングを利用することによりコストの問題に対処できるのではないかと考えた。クラウドソーシングの既存のサービスとしては yahoo クラウド[13]やランサーズ[14]などが存在している。これらのサービスはインターネット上の不特定多数の作業者に仕事を依頼する雇用形式で、低コストで迅速な作業が可能である。また、クラウドソーシングでは主にアンケート調査やデータ入力などの単純な業務が多い。それに対して論文間参照情報のアノテーションは比較的難易度の高いタスクである。そのため不特定多数の作業者がどの程度遂行できるのか、また専門家と比べるとどのような差があるのかなど大きな不安が挙げられる。

そこで我々は上記の懸念を念頭に、まずは論文間参照情報をアノテーションするためのプロトタイプを構築した。次に、論文サーベイに精通している大学教員を専門家に、大学生をクラウドソーシングで働く一般作業者に見立て、構築されたプロトタイプを利用してアノテーションしてもらった結果を比較した。この過程を通じて論文間参照情報のアノテーションにクラウドソーシングを利用する可能性の検討を行った[15]。

アノテーションのタグに関しては図 1 のようなタグの階層および種類を利用した。既存研究では 9~12 種類の参照理由が定義されていたが、いずれも単一階層で構成されているので、アノテーションやそれを利用した論文検索が容易ではないという問題がある[11][12]。本研究では図 1 のように 3 階層構造にし、少ない選択肢を複数回与えることで検索やアノテーションの負担を減らすことを目指した。

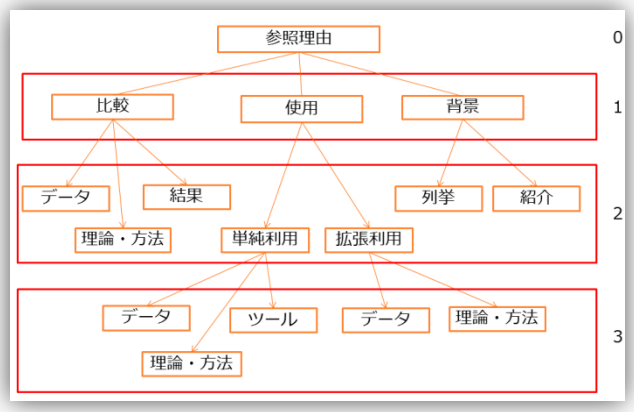


図 1 論文間参照関係の分類

実験に使用したデータは「言語処理学会年次大会発表論文集」に掲載された論文で、本文が日本語で書かれているものに限定した。

実験結果により、いくつかの課題が残ったものの、論文サーベイにそれほど精通していない一般作業者でも専門家に近い、良質なアノテーションを行う可能性が十分あることが示唆された。

3. 可視化システム

今までの研究で論文間参照情報のアノテーションにクラウドソーシングを利用することに関して、良質なアノテーションを行う可能性が十分あることが分かった。そのため本研究では十分なアノテーションが行われたものと仮定して、論文間参照情報を利用した可視化システムを構築する。3.1 では大まかなシステムの流れ、3.2 では論文間参照情報データベース、3.3 ではインターフェース・機能に関してそれぞれ説明していく。

3.1. システムの流れ

システムの流れは図 2 のようになっている。まずユーザが起点論文を選択し、その起点論文が参照している論文の情報(論文タイトル・著者等)や、どのような理由で参照を行っているのか(以下「参照理由」)などの内容を論文間参照情報データベースで検索をする。そして検索でヒットしたデータをもとに可視化を行い、ユーザに提示していく。単純にヒットした物を提示するだけでは論文候補が雑多になっ

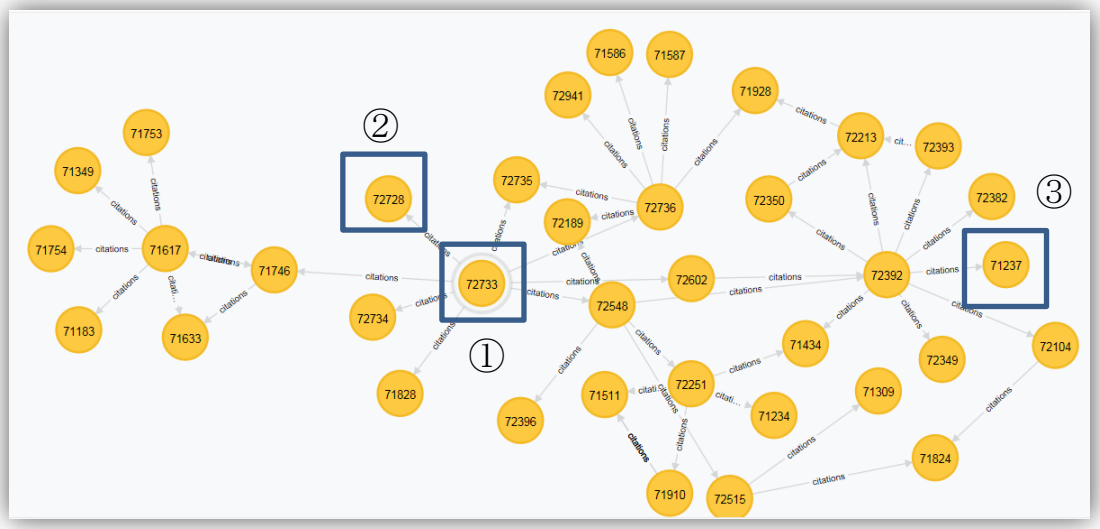


図 3 Neo4j ウェブインターフェース実行画面

てしまうので、その場合はユーザに表示したい参照理由を選択してもらい、フィルタリングをかけることで必要な論文だけを表示する。

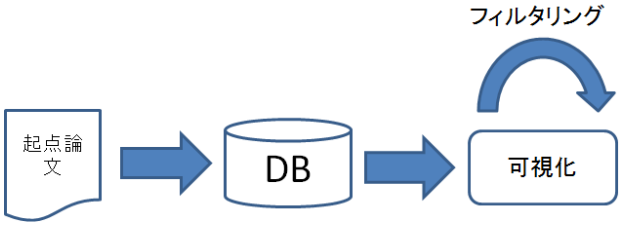


図 2 システム構成図

3.2. 論文間参照情報データベース

本研究では主に論文のタイトルや著者等の「論文情報」と、どの論文がどの論文を参照しているのか、どのような理由で参照を行っているのかといったような「論文間の参照情報」の 2 つの情報を扱う。これらのデータベース作成はグラフデータベースである neo4j[16]を利用して構築した。

neo4j ではグラフ構造のデータを扱うことができ、「ノードとノードがどのような関係性で結ばれているのか」といったような表現でデータを格納でき、本研究のように「論文と論文がどのような関係で結ばれているのか」といった情報を取り扱う場合には最適であると思われる。

図 3 は neo4j に実装されているウェブインターフェースの実行画面で、実際に構築した論文間参照関係のデータベースを使用している。図 3 の①は起点論文を表している。②は距離 1 の論文で起点論文が直接参照しているものである。③は距離 3 の論文で、今回は距離 3 までに限定して論文を表示しているため、終端ノードとなっている。

3.3. インターフェース・機能

図 4 に構築している可視化システムのインターフェースを示す。

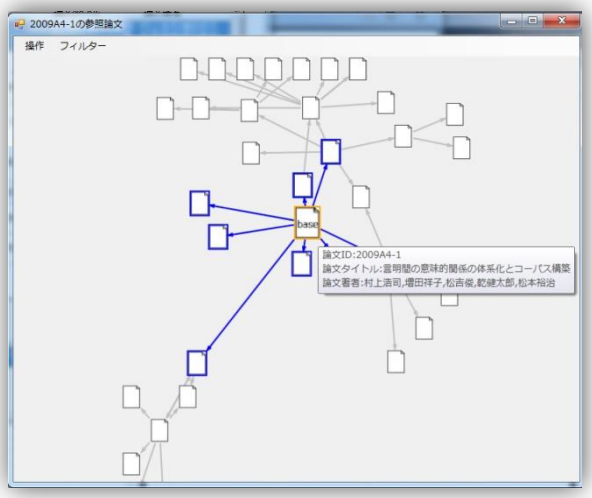


図 4 参照関係の可視化

図 7 参照情報

図 4 では起点論文が参照している論文、そしてさらにその論文が参照している論文、といったように参照をたどり、3 つ先まで参照関係を可視化している。論文やエッジが重なって見づらくなってしまった場合でも、マウス操作で論文の位置を移動することができるようになっている。今回のように表示される論文の数が多い場合には参照理由によるフィルタリングが効果的である。

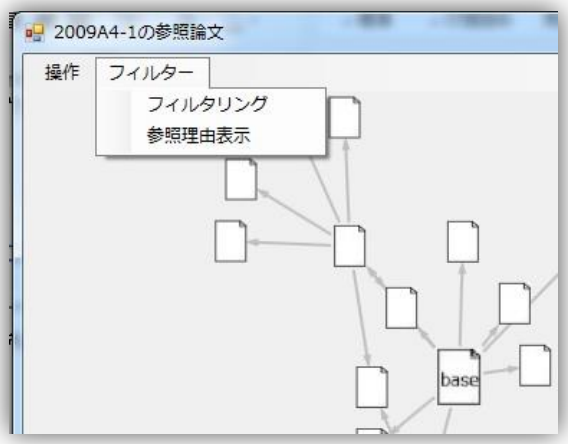


図 5 メニューバー

フィルタリングの実行と参照関係に関する情報の表示は図 5 のようにメニューバーから行えるようにした。そしてフィルタリングの設定画面は図 6 である。フィルタリングの設定方法は図 1 に示した階層型参照理由をもとに考案した。

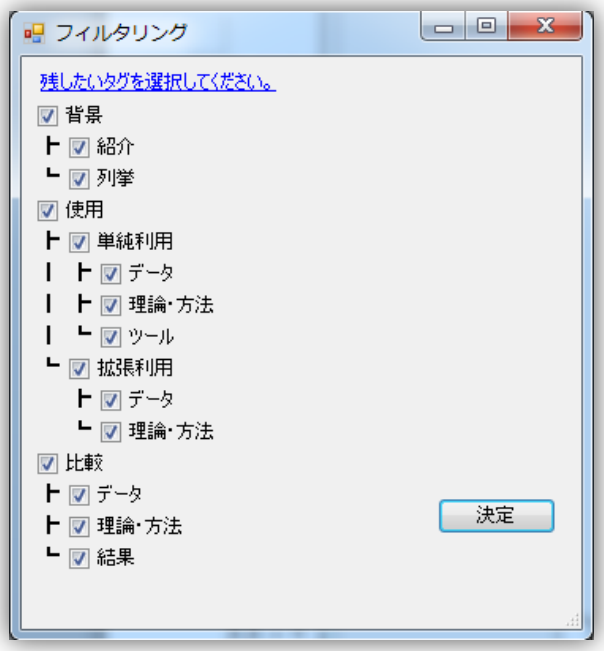


図 6 フィルタリング

このように階層式を導入することで、「背景を全て表示」や「データの比較だけを表示」といったような形でフィルタリングを変化させることにより、改めて再描写することができるようになっている。フィルタリングの対象は起点論文が直接参照している

被参照論文だけでなく、間接的に参照している被参照論文も全て含んでいる。これにより、1 章で述べているような「間接的に手法を参照している論文」等の探索も容易にできるようになっている。

図 7 は参照関係に関する情報で、参照理由や実際に論文の参照を行っている部分の情報などが表示される。これにより参照理由を確認し、最終的な必要な論文の取捨選択がしやすくなるようにした。図 7 の参照理由情報では情報量が多すぎる場合でも、被参照論文を一つ選択し、起点論文との関係性を最短距離で求めることができる。また、図 8 のように起点論文と興味を持った被参照論文との間にある参照理由のみが表示されるため、必要のない情報をシャットアウトすることができる。

図 8 起点論文との関係性表示機能

4. 評価実験

研究手法の有効性を検証するため客観評価、主観評価を行った。それぞれを順に述べる。

4.1. 客観評価

客観評価としては実際にフィルタリングシステムを利用することで検索タスクに対してどれだけ作業効率が上がるのかを評価した。検索タスクは 3 つ用意し、それに該当する解答を手作業で作成した。そして検索タスクに最適であると思われる参照理由でフィルタリングをかけることで検索効率を調査した。サーベイ検索効率は以下の式で評価した。

$$\frac{\textit{検索目的に合致した論文数}}{\textit{被参照論文数}}$$

また、タスクに利用した起点論文は以下の 3 編である。

1. 阿辺川武, 影浦峡:下訳から修正訳への訳文修正要因の分析, 言語処理学会第 14 回年次大会 pp. 253-256. (2008)
2. 大平真一, 山本和英:保険関連文書を対象とした校正支援システム, 言語処理学会第 18 回年次大会 pp. 243-246. (2012)
3. 大野潤一, 柴木優美, 山本和英 : Wikipedia のエントリ-リダイレクト間を対象にした同義関係抽出, 言語処理学会第 17 回年次大会 pp. 296-299. (2011)

探索は起点論文から 3 つ先までの参照関係を利用し、検索された被参照論文の数は論文 1、論文 2 が 14 編で、論文 3 が 24 編である。は論文 1 については「類似研究として翻訳を行っている研究」、論文 2 は「どのような理論や手法が利用されているのか」、論文 3 は「同義語抽出を行っている類似研究」をそれぞれタスクとして定めた。フィルタリング時に選択したタグは表 1 の通りで、1 が「表示」で 0 は「非表示」である。

表 1 フィルタリング時の選択したタグ

フィルタリングタグ			論文 1	論文 2	論文 3
背景		紹介	1	0	1
		列挙	0	0	0
使用	単純利用	データ	0	0	0
		理論･方法	1	1	1
		ツール	1	1	1
	拡張利用	データ	0	0	0
		理論･方法	1	1	1
比較		データ	0	0	0
		理論･方法	1	1	1
		結果	1	1	1

表 2 は客観評価のサーベイ検索効率を表している。フィルタリング前に比べて全体的に検索効率は向上しているのがわかる。特に起点論文 1 では 40%も向上していて効果が大きく表れている。フィルタリング後の正解個数の表示に関しては、起点論文 1 で 6 個中 5 個、起点論文 2 では 3 個中 3 個、起点論文 3 では 14 個中 10 個表示できた。表示個数を減らすのが目的なので論文を多く消しすぎてしまうことも懸念していたが起点論文 1 では 1 つを除いて全て、起点論文 2 では全て表示できていたという結果が得られた。

表 2 客観評価のサーベイ検索効率

	サーベイ検索効率	
	フィルタリング前	フィルタリング後
起点論文 1	43%	83%
起点論文 2	21%	30%
起点論文 3	58%	59%

今後の予定として、今回使用した解答やフィルタリングの内容に対する信憑性に疑問が残るため、今後それらの妥当性を検証していく。

4.2. 主観評価

主観評価では同じ大学で情報科学を専門とする学部生 10 人に本システムを使用してもらい、その使用感をアンケート調査することで評価した。評価はフィルタリングの実行・参照理由の表示が行えない物(以下システム 1)と、フィルタリングの実行・参照理由の表示が行える物(システム 2)との比較で行った。表 3 と表 4 はそれぞれシステム 1、システム 2 のどちらの方が使いやすかったか、またどちらの方が検索の効率が上がったと思うかのアンケート結果である。

表 3 どちらの方が使いやすかったかアンケート

	回答者数
システム 1 の方が使いやすい	1
システム 2 の方が使いやすい	8
変わらない	1

表 4 どちらの方が効率が上がったかアンケート

	回答者数
システム 1 の方が効率が良い	1
システム 2 の方が効率が良い	9
変わらない	0

表からわかるとおり、ほとんどの被験者からフィルタリングシステム・参照理由を利用してのシステムの方が使いやすい、効率が上がったという評価が得られた。

「システム 1 の方が使いやすかった」、または「変わらない」と答えた人の意見としては参照理由のタグの種類がよくわからないというものがあった。またシステム 1 の方が効率が上がったと答えた人に関してはフィルタリングシステムがうまく扱えず、関連性が高い論文までシャットアウトしてしまうという事が起きたのが原因ではないかと考えられる。これらの問題点の対策としてはチュートリアルを充実させることや、システム中の説明を増やす等のことが挙げられる。

その他の意見としては参照理由を可視化画面に表示してほしい、論文のアイコンが起点論文以外すべて同じなのでどの論文を見ていたのかわからなくなった等があった。これらのような UI の問題点は今後システムの改良の際に考慮する予定である。

5. まとめ

本研究では論文サーベイの効率化のための参照情報の可視化システムを構築し、その評価を行った。

実験結果から論文サーベイの検索効率向上を図れた他、システムに関するアンケートでは本システムを利用することで論文サーベイが効率的に行えたという意見を多く得ることができた。しかし、客観評価で用いた解答やフィルタリングの内容が本当に正しいのかどうかという信憑性に対しての疑問が残っているため、今後は解答の妥当性を検証していく予定である。その他、主観評価からの実験で得たアンケートからシステムの改善点などが意見として寄せられているため、これらに対応していくのも今後の課題である。

参考文献

[1] http://ci.nii.ac.jp/
[2] http://scholar.google.co.jp
[3] http://citeseerx.ist.psu.edu/
[4] R. Baeza-Yates, B. Ribeiro-Neto. “Modern Information Retrieval” (1999)
[5] K. Dobashi, H. Yamauchi., R. Tachibana. “Keyword Mining andVisualization from Text Corpus for Knowledge Chain Discovery Support”,Technical Report of IEICE, NLC2003-24, pp.55-60. (2003) (in Japanese)
[6] H. Small. “Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents”, Journal of the American Society for Information Science, Vol.24, No.4, pp.265-269. (1973)
[7] M. Kessler. “Bibliographic Coupling between Scientific Literatures”,Journal of the American Documentation, Vol.14, No.1, pp.10-25, (1963)
[8] 清水 成昭,竹中 豊文: 文献の参照関係を視覚化するアプリケーションの提案・実装,電子情報通信学会技術研究報告. IN,情報ネットワーク 109(449), 389-394, (2010)
[9] 渡辺 秀文,北川 晴香,齋藤 隆文: 文献の参照関係の可視化,情報処理学会 研究報告グラフィクスと CAD (CG) 2010-CG-139(6), 1-6,(2010)
[10] 難波英嗣,神門典子,奥村学: 論文間の参照情報を考慮した関連論文の組織化,情報通信学会論文誌,42(11), pp.2640-2649. (2001)
[11] 小出寛史, 韓東力: 論文間参照情報のデータベース化に基づく参照タイプの同定,自然言語処

理研究会報告 2012-NL-209(2), 1-7(2012)
[１２] Teufel, S.: The Structure of Scientific Articles -Applications to Citation Indexing and Summarization. CSLI Publications. (2010)
[１３] http://www.lancers.jp/
[１４] http://crowdsourcing.yahoo.co.jp/
[１５] 井上 絢翔,韓 東力: 論文間参照情報のアノテーションにおけるクラウドソーシングの利用検討, 言語処理学会 第 21 回年次大会 pp.736-739.(2015)
[１６] http://neo4j.com/