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5回)
SIG-AM-05-01

消費カロリーを暗黙的に管理する 散歩ナビゲーションシステムの提案

Proposal of Walking Navigation System with Implicit Control of Consumed Calorie

奥村 尚史[1]　佐々木 渉[1]　高間 康史[1*]

Takafumi Okumura[1], Wataru Sasaki[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: 近年，運動不足に起因する生活習慣病の対策として，誰でも気軽に行うことができるウォーキングが注目されている．特に，活動量計との併用は運動効果を実感でき，長期の運動に対するモチベーションの持続が期待されている．しかし，目標のカロリーを消費するためには歩行中常にカロリーを意識する必要があるため，運動する義務感や煩わしさが継続性を阻害する要因となりうる．そこで本稿では，消費カロリーを暗黙的に管理する散歩ナビゲーションシステムを提案する．

1　はじめに

本稿は，消費カロリーを暗黙的に管理する散歩ナビゲーションシステムを提案する．近年，運動能力低下が問題視されており，交通機関やコンビニエンスストアの発達による利便性の向上や基礎体力低下による運動持続の困難などが理由として挙げられている[12]．そこで，誰でも気軽に行うことができるウォーキングが運動不足の改善手段として注目されているが，長期に渡る運動の継続と運動の単調さから運動効果を実感することは難しい．そのため，ウォーキングに対するモチベーションを維持することは容易ではなく，ウォーキングの習慣化を支援することが重要と考える．

また近年，歩数計や活動量計が普及しつつあり，手軽に運動効果を実感できることから運動に対するモチベーションの持続につながることが期待されている．しかし市販の活動量計では自身でカロリーを管理する必要があり，目標のカロリーを消費するためには歩行中常にカロリーを意識することになる．そのため，カロリー消費を意識した運動に対する義務感や管理することへの煩わしさが継続性を阻害する要因となりうる．そこで本稿では，消費カロリーを暗黙的に管理する散歩ナビゲーションシステムを提案する．

提案するナビゲーションシステムでは，現在の消費カロリーをリアルタイムに計算し，目標カロリーに対する残余カロリーを算出する．提案手法は，次の目的地を反復的に推薦する往路モードと，自宅までの帰宅ルートを推薦する復路モードに大別される．往路モードでは，各交差点をノードとした道路ネットワークを構築し，残余カロリー以内で到達可能な目的地を繰り返し推薦する．復路モードでは，残余カロリーをなるべく満たす帰宅ルートを推薦することで，目標カロリーを消費できる散歩ルートの推薦を可能とする．また，歩行者向けのナビゲーションは，ルート案内のための手法や進路情報の提示方法などがヒューマンインタフェースの分野で研究されている[2][8][9]．本稿では，わかりやすさや安全性を考慮して，地図ベースではない直感型ナビゲーションインタフェースを提案する．

上述の通り，提案システムは，往路推薦モード，帰宅推薦モード，ナビゲーションインタフェースなど複数の要素から構成される．このうち，帰宅推薦モードに関して，探索アルゴリズムの分野では，既定のコストを満たすルートの推薦手法は未だ確立していない．そこで本稿では，既定のカロリーを消費することを目的とした帰宅ルート推薦アルゴリズムを提案し，予備実験結果を示す．

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

2 関連研究

2.1 継続性を考慮した運動支援システム

運動不足に起因する生活習慣病を予防するための取り組みが，全国の病院や自治体を中心に活発化しており，その改善に向けた研究も行われている[3]. 生活習慣病の予防には適度な運動を継続することが重要であるが，長期に渡る運動の継続とそれに伴う身体への負担から継続的な運動に対するモチベーションを維持することは容易ではなく，一人で継続しようとしてもなかなか長続きしないのが現状である. このため，ユーザの継続的な運動への取り組みを支援するための研究が数多く行われている[11][14].

山村ら[14]は，利用者同士の交流を促す SNS (Social Network Service) の仕組みを活用したウォーキング支援システムを提案している．このシステムでは，毎日の歩行実績を管理し，視覚的に確認できる仕組みを提供する．家族や友人など現実世界の人間関係をシステムに持ち込み共有することで，歩行状況の確認やコミュニケーションを容易にし，対抗意識や仲間意識を促進する．また，顔見知りのいないユーザを考慮したライバル関係を採用することで, 誰でも仮想的な競争相手を確保することができる.

酒田ら[11]は，Situated Music をユーザへの刺激として提供するインタラクティブジョギングシステムを提案している．Situated Music とは，ユーザの状況や環境に即し選択･再生される楽曲やそのための枠組みである．ユーザに無理なく目標を達成させる手法として，ユーザの意識・無意識に働きかけ運動量を自然に調整する方法がある．酒田らはユーザのジョギングペースから時間内に目標距離を走破可能か推定し，ペースの乱れをテンポの速い曲や遅い曲，また，エフェクトをかけて調整することで，ユーザの身体への過負荷を未然に防止することを目指している.

2.2 経路推薦のための歩行ナビゲーションシステム

今日の高度情報化社会の進展により，様々な移動体情報端末が開発された結果，カーナビゲーションシステムだけでなく歩行者を誘導する歩行者ナビゲーションシステムも普及しつつある．カーナビゲーションの経路推薦では，移動手段が車であることから移動距離は長く，交通渋滞や移動時間などを考慮した経路情報を付加することで経路の選定が可能となる．しかし，対象が歩行者である場合には移動距離は短いことや，歩行者の嗜好パターンが複雑であるなどの違いがあるため，カーナビゲーションの推薦システムをそのまま適用するのは困難である．従って，歩行者を対象とした経路推薦手法は確立しきれていないのが現状である．このため，歩行者向けの経路推薦を目的としたナビゲーションシステムが数多く研究されている[1][4][7][13].

荒井ら[1]は，歩行者の嗜好を反映した最適経路を提供することを目的として，屋内に特化した経路探索手法を提案している．屋内空間では，道路という概念がなく，道路ネットワークの構築・利用が困難である．そこで荒井らは，屋内にある店舗などのオブジェクトを構成する輪郭線が交わる頂点をノード, それらを結ぶ線をエッジとすることで屋内型ネットワークデータを構築している．さらに，車椅子やベビーカーの所有の有無をユーザに確認し，その結果をルート推薦時に昇降機の利用として反映させることで, ユーザの負担を考慮した最適経路を提供する.

生田目ら[7]は，歩行者の歩行履歴情報から信頼性のある位置情報をノードとして抽出し経路ネットワークの構築を行うことで，施設内における歩行者向けのナビゲーションシステムを提案している．歩行者の複雑な行動パターンに対応することを目的として，歩行者が歩行するエリアこそが経路であると定義している．ユーザの歩行履歴情報から取得した位置情報のうち，同じ緯度・経度となる位置情報の頻度に基づきユーザの歩行頻度の高い経路を構成する位置をノードとして抽出する．各ノードの信頼性を基準に経路ネットワークを構築することで，ユーザへの推薦歩行経路を提供する.

山本ら[13]は，ユーザから投稿された情報の時間・位置データを利用し，ユーザの現在地を考慮したリアルタイムのナビゲーションシステムを提案している．ユーザが投稿する情報の種類は，バリア情報，店の内容とその評価情報，その他のスポット情報に分けられる．バリア情報は道路などの状況を示すもので経路の推薦に活用し，店やスポット情報は目的地の推薦に活用する．現状のバリアフリーマップは頻繁に更新が行えないために，情報の即時性が懸念されているが，山本らが提案しているシステムを用いることで，その時点における所在地に適した鮮度の高い情報を得ることが可能となる.

3 消費カロリーを暗黙的に管理する散歩ナビゲーションシステム

3.1 システム構成

提案システムは，ウォーキングを通して運動を行いたいユーザを対象とし，3 軸加速度センサおよび

GPS センサを搭載したスマートフォン用アプリケーションとしての利用を想定する．提案する散歩ナビゲーションのシステム構成を図 1 に示す．提案システムは，道路情報を RDF 形式で格納する FusekiServer およびカロリー情報の管理・推薦サービスを行う Server，Googlemap[1]により作成したシミュレータ，ユーザが所持するスマートフォンから構成される．FusekiServer で管理する道路情報は，各交差点の位置情報である．ユーザへのナビゲーションは Android を用いたナビゲーションインタフェースにより行うが，推薦ルートの閲覧・検証はシミュレータを用いて行うこともできる．

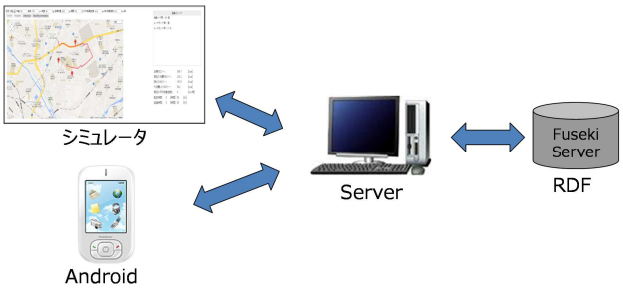


図 1　消費カロリーを管理する散歩ナビゲーションのシステム構成

3.2 カロリーを考慮した散歩ルート推薦

提案システムでは，ユーザからプロフィールとして取得する性別・年齢・身長・体重・目標体重，および目標体重を達成するまでに要する期間からカロリーを算出する．厚生労働省が定義している下記の算出式[6]を用いて，各ルートに対するユーザの消費カロリーを求める．

$$C = RMR * W * METS * T \quad (1)$$

$$RMR \fallingdotseq BMR * 1.2 \quad (2)$$

$$BMR(\text{Man}) = 66.47 + (13.75 * W) + (5.0 * H) + (6.76 * Age) \quad (3)$$

$$BMR(\text{Woman}) = 655.1 + (9.56 * W) + (1.85 * H) + (4.68 * Age) \quad (4)$$

ここで，C[kcal]は消費カロリーを表し，座位安静時代謝量 RMR[kcal]，体重 W[kg]，時間 T[h]から算出される．RMR は，基礎代謝量 BMR[kcal]から概算を求めることができ[10]，BMR は，欧米にて確立された評価法であるハリス・ベネディクト計算式を用いて体重 W[kg]，身長 H[cm]，年齢 Age から算出する．

散歩ルート推薦システムのフローチャートを図 2 に示す．提案システムによる推薦は，次の目的地を反復的に推薦する往路モードと，自宅までの帰宅ルートを推薦する復路モードに大別される．往路モードでは，帰宅時に消費するカロリーを考慮しながら，反復的に目的地を選択し，ユーザに推薦する．図 2 において，t 回目の反復における目的地候補集合をS_t，その中から選択・推薦される目的地をP_tとし，また，現在地からS_t内の各目的地を経由した帰宅地点までに消費すると思われるカロリーを推定消費カロリー，目標カロリーから現在までの消費カロリーを差し引いた値を残余カロリーとする．推定消費カロリーが残余カロリーより小さい目的地がS_t内に存在するとき，S_tの中からランダムに一つ目的地を選択することでP_tを決定する．この条件を満たす目的地がS_t内に存在しない場合は，新たな目的地を推薦せずに復路モードに移行する．なお，残余カロリーの初期値は目標カロリーとする．

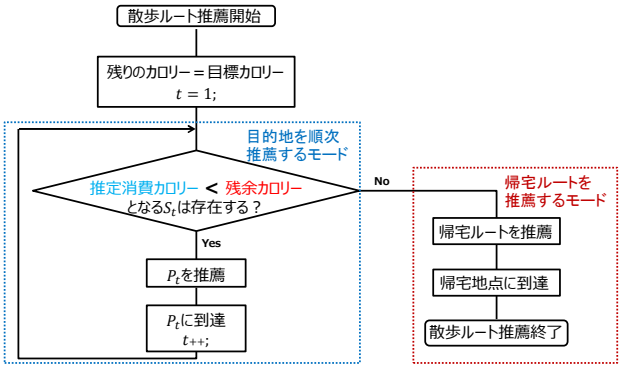


図 2　散歩ルート推薦システムのフローチャート

図 3 は推薦経路をシミュレータ上に表示したものである．シミュレータ上では，目的地までのルート上に位置する交差点にマーカを設置し，通過したルートをラインにより表示している．提案システムでは，目的地へ向かう道中にユーザが消費するカロリーをリアルタイムに計算し，残余カロリーを算出する．ユーザが推薦した進路方向以外へ移動すると，再び目的地までのリルートを行う．これによりユーザは必ずしもシステム側の示す進路方向に従う必要はなく，カロリー消費のためのナビゲーションを意識せずに目標カロリーを消費できる．また，ユーザの運動ペースを考慮して目標カロリーの更新を行う．経過時間が散歩予定時間の半分を過ぎたとき，ユーザの現在の平均移動速度が普段の平均移動速度を下回っていれば，現在の歩行ペースで予定時間通りに散歩を終えることが可能な値まで目標カロリーを低減する．目的地に到達すると，再び残余カロリーに応じて目的地を推薦する．以降推薦目的地の候補がなくなるまで繰り返す．

[1]https://maps.google.co.jp/maps?hl=ja&tab=wl

図 3　推薦されたルート

前述の通り，推薦できる目的地がなくなると復路モードへ移行し，残余カロリーをなるべく満たす帰宅ルートを推薦する．図 4 に示すシミュレータ上では，通過したルートを赤線，帰宅経路として推薦したルートを青線により表示している．最適なルートを推薦するためには，考えられる全てのルートを探索する必要があるが，探索範囲は一般に膨大となるため，リアルタイムに最適ルートを求めることは困難である．そこで本稿では，道のりの近似値を求めることで探索範囲を絞り込む手法を提案する．以下に帰宅ルート推薦アルゴリズムを示す．

図 4　推薦された帰宅ルート

Step1　現在地から各交差点x_iを経由した帰宅地点までの距離d_{i1}を計算する
Step2　$d_{i2} = d_{i1} * \{(\sqrt{2}+1)/2\}$
Step3　d_{i2}を平均移動速度を用いてカロリーc_iに変換する
Step4　c_iが残余カロリー値に近い上位 25 交差点$\{x_i^*\}$を中継交差点候補として抽出する
Step5　現在地からx_i^*を経由した帰宅地点までの最短経路d_{i3}を計算する
Step6　d_{i3}のカロリー換算値が残余カロリー値に最も近いルートを推薦する

Step 2 では，図 5 に示す考えに基づき，2 点間の最短経路の道のりを近似している．2 点間の最短経路の道のりは距離（図中赤線）以上となるが，最大値を直角二等辺三角形（図中の黒線）の 2 等辺の和と想定する．現在地とx_iの距離，およびx_iと帰宅地点間距離の和であるd_{i1}に対し，最小値d_{i1}と最大値$\sqrt{2}d_{i1}$の平均値として，この経路の道のりを近似する．Step 4 では，抽出する交差点の数を上位 25 個としているが，その値は 4 節に示す実験結果に基づき決定した．Step 5 では，x_i^*を経由した帰宅地への最短経路を求めているが，探索範囲を限定しているため 4 節で示すように短時間での計算が可能である．

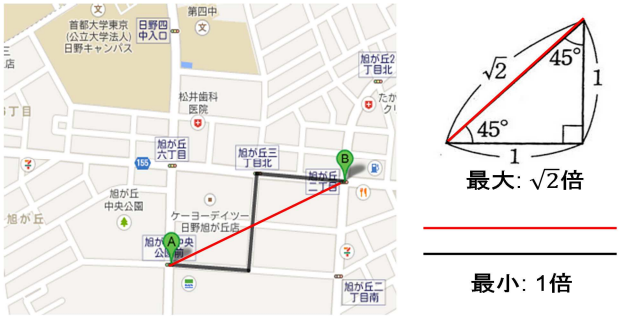


図 5　2 点間の道のり近似値の計算

3.3 直感型ナビゲーションインタフェース

3.2 節に示した散歩ルート推薦結果に基づき，進路情報をユーザへ提示するためのナビゲーションインタフェースを提案する．サーバから推薦ルートを構成する各交差点の位置情報を WebSocket 通信を用いて Android に転送する．このとき TooTallNate / Java – WebSocket・GitHub をライブラリとして使用し，通信のフォーマットに JSON を利用した．

Android により開発した直感型ナビゲーションインタフェースのスクリーンショットを図 6，図 7 に示す．サーバから推薦ルート上の交差点位置情報を取得し，交差点ごとに次の進路情報を矢印により提示する（図 6(a)）．端末の向きとサーバから取得する進路方向を考慮し，リアルタイムに進路情報を更新・表示する（図 6(b)，図 7(a)）．ユーザが交差点に到達すると，端末の振動により到達の告知を行い進路情報を更新する（図 7(b)）．

(a) 進路情報　(b) 端末向き(右45度)

図 6　直感型ナビゲーションインタフェースのスクリーンショット（1）

(a) 端末向き(右180度)　(b) 進路情報の更新

図 7　直感型ナビゲーションインタフェースの
スクリーンショット（2）

4　帰宅ルート推薦の評価と考察

3.2 節で提案した帰宅ルート推薦アルゴリズムの計算時間および推薦精度について，最適解（経路）を求めた結果と比較することで, その有効性を示す.

8 つの地点（復路出発地点）から，提案手法による推薦ルートで帰宅した時の残余カロリー値を表 1 に示す．このとき，各復路出発地点から帰宅地点到着までの間に消費したいカロリー（復路目標カロリー）およびその距離換算値は，表の 2,3 行目に示す値としている．この距離換算値は，20 代の平均的成人男性の場合を想定した算出値とする．また，3.2 節の Step4 で抽出する交差点数を上位 5 個，10 個，15 個，20 個，25 個，30 個にした場合それぞれについて，得られた経路による残余カロリーおよび経路探索に要した平均処理時間を示している．近似値を用いずに，残余カロリーが最も小さくなる最適解を求めたときの残余カロリー値を，上位 5 経路について表 2 に示す．このとき最上位の経路を探索するのに要した平均処理時間は 3 分 8 秒であった.

表 1　提案手法による残余カロリー値（絶対値）

復路出発地点	A	B	C	D	E	F	G	H	平均処理時間[s]
帰宅地点までの距離[m]	590	1017	762	935	236	982	491	849	
復路目標カロリー[kcal]	178	137.9	161.8	145.6	211.2	141.2	187.2	153.6	
距離換算値[m]	1780	1379	1618	1456	2112	1412	1872	1536	
上位 5 交差点[kcal]	1.69	0.85	1.6	3.5	3.2	0.18	0.38	4.4	2.26
上位 10 交差点[kcal]	1.69	0.09	1.6	0.18	2.0	0.18	0.38	2.8	3.21
上位 15 交差点[kcal]	1.69	0.09	1.6	0.18	2.0	0.18	0.38	1.79	3.95
上位 20 交差点[kcal]	1.69	0.09	0.10	0.18	2.0	0.18	0.38	1.41	5.15
上位 25 交差点[kcal]	1.69	0.09	0.10	0.18	0.57	0.18	0.38	1.41	5.83
上位 30 交差点[kcal]	1.69	0.09	0.10	0.18	0.57	0.18	0.38	1.41	7.09

表 2　最適解探索による上位 5 経路の
残余カロリー値（絶対値）[kcal]

復路出発地点	A	B	C	D	E	F	G	H
1 位	0.00	0.09	0.10	0.09	0.38	0.18	0.38	1.41
2 位	1.12	0.74	0.47	0.10	0.56	0.38	0.47	1.69
3 位	1.69	0.75	0.75	0.18	0.57	1.13	0.85	1.79
4 位	1.78	0.75	1.40	1.04	0.57	1.32	1.04	2.25
5 位	1.79	0.85	1.40	1.31	0.75	1.59	1.22	2.26

抽出交差点数が増えるほど，残余カロリーは減少するものの処理時間も増加することがわかる．表 2 に示す結果と比較すると, 抽出交差点数 25 個以上では，全地点について最適解を求めた場合の 3 位以内に入っていることがわかる.

計算時間に関しては，提案システムではユーザの体調や気分を考慮し，任意のタイミングで帰宅ルート推薦モードに切り替わることを想定している．従って, あらかじめルートを計算しておくのではなく，リアルタイムに求める必要がある．計算機からの応答時間により発生するユーザの待ち時間が 1.5〜6 秒の場合に，心理的負担が最も抑えられることが先行研究により示されている[5]．この知見を考慮すると，上位 25 交差点を探索範囲とすることが適切と考える．このとき，カロリーの誤差（残余カロリーの絶対値）は平均 0.575kcal であり，これを式(1)により距離に換算すると 5.8m となる. これは暗黙的なカロリー管理の数値として十分な性能と考える.

また，中継する交差点の数を 2 個に増やした場合も検証した．近似値による抽出交差点数を上位 25 個としたときの中継交差点数 1 個と 2 個の結果を表 3 に示す.

表 3　中継交差点数の違いによる残余カロリーの
比較（単位[kcal]）

復路出発地点	A	B	C	D	E	F	G	H	平均処理時間[s]
中継交差点数：1 個	1.69	0.09	0.10	0.18	0.57	0.18	0.38	1.41	5.83
中継交差点数：2 個	0.66	0.75	0.00	0.56	0.94	0.00	0.47	0.19	81.02

目的地 B，D，E，G では，中継する交差点の数は 1 個の方が優れた結果となったが，A，C，F，H では 2 個の方が優れた結果となった．平均値を取ると 1 個では 0.575kcal，2 個では 0.446kcal となり 2 個の方が優れているが，処理時間は大幅に増加していることがわかる.

中継する交差点数が 2 個の場合に探索範囲となる経路は，交差点数 1 個の場合を包含するため，目標カロリーに近い経路が探索範囲に存在する可能性は高くなる．しかし，本稿で提案する近似値と実際の

最短経路の誤差は交差点ごとにまちまちであるため，表 3 に示すように必ずしも 2 個の方が優れた結果が出るとは限らない．一方，探索範囲の拡大による処理時間の増加は大きく，実用的な側面から利用できるとは言い難い．よって，中継する交差点の数は 1 個で，近似値により抽出する交差点の数は上位 25 個のとき，暗黙的なカロリー管理に十分な精度を実用的な処理時間で実現可能と考える．

5 おわりに

本稿では，ユーザがカロリーを意識せずに目標のカロリーを消費できることを目的として，消費カロリーを暗黙的に管理する散歩ナビゲーションシステムを提案し，帰宅ルート推薦アルゴリズムに関する予備実験を行った．

帰宅ルート推薦アルゴリズムにおいて，近似値により抽出する交差点数と精度および処理時間の関係について検証を行った結果，上位 25 個の交差点を抽出した場合に，処理時間・精度の両面で良い結果が得られた．また，中継交差点数を 2 個に増やした場合の実験も行ったが，精度向上効果はそれほど見られない一方，処理時間が大幅に増加するため実用には適さないことを示した．

今後は，直感型ナビゲーションインタフェースの視認性や操作性をさらに向上させ，ユーザインタフェースとしての完成度を高めることで，利用者の負荷軽減が期待できる．また，ナビゲーションインタフェースに関しては，案内のわかりやすさや，歩行時の安全性などの観点から，実験協力者による評価実験を行う予定である．

参考文献

[1] 荒井 亨, 戸川 望, 柳澤 政生, 大附 辰夫 : 屋内歩行者ナビゲーションにおける歩行者の嗜好を反映させる経路探索手法，電子情報通信学会技術研究報告.ITS, Vol. 106, No. 266, pp. 47-52, 2006

[2] 福井 良太郎，白川 洋，歌川 由香，重野 寛，岡田 謙一，松下 温：携帯電話における歩行者ナビゲーション情報の表示方法に関する提案と評価，情報処理学会論文誌, Vol. 44, No. 12, pp. 2968-2978, 2003

[3] 石原 礼子, 馬場園 明, 亀 千保子, 八尋 玄徳, 西岡 和男：生活習慣病予防事業におけるメンタルヘルスの変化と生活習慣改善および身体的健康度改善との関連，日本衛生學雑誌，Vol. 60, No. 4, pp. 442-449, 2005

[4] 川端 将之，日裏 博之，上田 真由美，上島 紳一：利用者コンテキストを考慮した歩行者ナビゲーションシステムの利用可能性について，情報研究:関西大学総合情報学部紀要, Vol. 22, pp. 91-104, 2005

[5] 小松原 明哲，横溝 克己:計算機応答時間の人間工学的許容範囲に関する一考察，人間工学, Vol. 24, No. 3, pp. 195-202, 1988

[6] 厚生労働省：健康づくりのための運動指針 2006，エクササイズガイド, p. 6, 2006

[7] 生田目 宏昭，神戸 英利，三井 浩康，小泉 寿男：歩行履歴情報を基にした歩行者ナビゲーションシステムの構築，情報処理学会 データベース・システム研究会報告, Vol. 2008, No. 7, pp. 13-18, 2008

[8] 長岡 哲郎，矢内 裕之，長谷川 孝明：上腕部での振動により歩行者道案内を行う WYSIWYAS 案内バンドについて，電子情報通信学会技術研究報告.ITS, Vol. 106, No. 616, pp. 37-42, 2007

[9] 二宮 直也，戸川 望，柳澤 政生，大附 辰夫：歩行者ナビゲーションにおける微小画面での視認性とユーザの迷いにくさを考慮した略地図生成手法，電子情報通信学会技術研究報告.ITS, Vol. 106, No. 266, pp. 53-58, 2006

[10] 佐川 貢一，石原 正，猪岡 光，猪岡 英二：歩行形態の違いを考慮した消費カロリーの無拘束推定，計測自動制御学会東北支部 第 183 回研究集会, No. 183-7, pp. 1-8, 1999

[11] 酒田 信親，興梠 正克，大隈 隆史，蔵田 武志：Situated Music:インタラクティブジョギングへの応用，ヒューマンインタフェースシンポジウム 2005, pp. 459-462, 2005

[12] 若吉 浩二，高橋 豪仁，今枝 和与，岸田 悟，長谷川 芳彦，石川 元美，田辺 正友 : 小学生児童における運動能力･運動習慣の経年的変化: スポーツ教室開催の影響（自然科学）. 奈良教育大学紀要，自然科学, Vol. 54, No. 2, pp. 39-47, 2005

[13] 山本 浩司，安村 禎明，片上 大輔，新田 克己，相場 亮，宮城 政雄，桑田 仁 : ユーザの投稿情報に基づく経路ナビゲーション，人工知能学会全国大会論文集, Vol. 18, pp. 1B3-01, 2004

[14] 山村 豊，井上 悦子，吉廣 卓哉，中川 優：生活習慣病予防のための SNS の仕組みを用いたウォーキング継続支援システム，電子情報通信学会 第 19 回データ工学ワークショップ(DEWS2008), pp. A8-6, 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5回)
SIG-AM-05-02

RDF データベースを対象とした データ分析支援ツールの提案

Proposal of Data Analysis Support Tool for RDF Database

田代 航一* 高間 康史

Koichi Tashiro, Yasufumi Takama

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

Abstract: 本稿では，RDF の特徴を考慮した分析支援ツールを提案する．RDF は分散した情報の結合や，構造化されていない情報の扱いが容易という特徴があり，活用事例が増えてきている．しかし，従来主流の関係データベースとは異なる特徴を有するため，専用の分析支援ツールが必要と考える．本稿では，構造化された部分データ空間の抽出，接続性の高いリソースの発見， ログデータを対象とした時系列データの抽出を行うツールを提案し，テキストデータマイニング統合環境 TETDM を用いて実装したプロトタイプを示す．

1. はじめに

本稿では RDF の特徴を考慮したデータ分析支援ツールを提案し，テキストデータマイニング統合開発環境 TETDM[1][4][5]にて実装したプロトタイプを示す．

Resource Description Framework (RDF) とは W3C[2] で規格化されている情報規格であり，リソースに関する情報を主語・述語 (プロパティ) ・目的語の 3 要素 (トリプル) で表現する．トリプルはグラフ構造で表すことが可能であり，述語や目的語の追加・更新を容易に行える．また，分散した情報の結合や，構造化されていない情報の扱いが容易という特徴がある．RDF で記述されたデータは，SPARQL などの RDF クエリ言語により，検索・操作することが可能である．

近年，RDF の特徴を生かし，ガバメント分野や科学技術分野などでデータを RDF で記述し活用する事例が増えてきている[6][7]．それに伴い，RDF データベースを分析する必要性が増してきていると考える．しかし，RDF データベースは複数のデータベースを横断して操作・検索を行うことが可能であることや，スケーラビリティが高く複雑なグラフ構造をしていることなど，従来主流の関係データベースとは異なる性質を有するため，既存のデータ分析ツールをそのまま適用することは難しく，専用の分析ツールが必要であると考える．

一方，分散されているデータマイニングツールを統一的に扱うことを目的とした，テキストデータマイニングによる統合開発環境 TETDM が公開されている．データマイニングのシステムやツールは，研究者・開発者が特定の分析に用いるために独自に開発しているケースが多く，また開発したツールは公開されない事も多い．TETDM はそういったツールを TETDM のモジュールとして公開・配布し，再利用することを目的としており，これにより分析者はツールの開発から始めることなく，分析に専念することが可能となる．

本稿では，RDF の特徴を考慮したデータ分析支援ツールを 3 種類提案する．提案する分析支援ツールは，構造化された部分データ空間を抽出し， 共通の述語を持つ主語の抽出･テーブル作成を行うツール，接続性の高いリソースの発見のために，複数エンドポイント間の共通リソース抽出を行うツール，そしてログデータを対象とした時系列データ抽出である．これらを TETDM のモジュールとして実装し，実際の RDF データベースに適用した事例を示す．

* 連絡先 首都大学東京 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: tashiro-koichi@ed.tmu.ac.jp

[1] http://tetdm.jp/pukiwiki/index.php
[2] http://www.w3.org

2. 関連研究

2.1 RDF の活用事例

表 1 に，RDF データを公開しているサイトの一例を示す．欧米諸国では，政府や公共機関の統計データを，RDF を利用した Linked Open Data (LOD) として公開しており，代表的なものとして英国の DATA.GOV.UK や米国の DATA.GOV がある．また，Wikipedia[3] を RDF により記述した DBpedia も代表的な LOD である. 科学技術関連ではライフサイエンスの分野において，タンパク質の知識ベースとして，単一データベースではあるが UniProt が公開されている．さらに日本でも同様に，福井県鯖江市や，横浜市の芸術関連施設を公開している Yokohama Art Spot[12]などが RDF を採用している．また，日本語版 Wikipedia を RDF 化した DBpedia Japanese も公開されている．また，ライフサイエンスの分野では，データベースの統合が重要であるとの考えから，ヒト遺伝子データベース H-InvDB が公開されている[13]．さらには，食生活の管理・分析を行なう Web サービスとして FoodLog があり，ライフログデータの記述に RDF が用いられているなど, 様々な分野のデータに対して RDF が利用されている．

表 1: RDF の活用事例.

サイト名	URL
DATA.GOV.UK	http://www.data.gov.uk/
DATA.GOV	http://www.data.gov/
DBpedia	http://wiki.dbpedia.org/About
UniProt	http://www.uniprot.org/
福井県鯖江市	http://www.city.sabae.fukui.jp/pageview.html?id=11552
Yokohama Art Spot	http://lod.ac/apps/yas/
DBpedia Japanese	http://ja.dbpedia.org/
H-InvDB	http://www.h-invitational.jp/hinv/ahg-db/index_ja.jsp
FoodLog	http://www.foo-log.co.jp/index.html

2.2 RDF データ分析

現在，RDF データを対象とした分析に関する研究は，データ構造の理解を目的としたスキーマの可視化と，データ分布の理解を目的としたグラフ検索支援の 2 つに大別できる．以下ではそれぞれについてまとめる．

2.2.1 データ構造の理解

Goyal らは, RDF で記述されたデータを読み込み，情報関係をグラフ形式で表す RDF Gravity[4]を公開している． RDF Gravity は RDF クエリ言語の 1 つである RDQL によりトリプルを検索し，検索結果のグラフを表示することができる．Deligiannidis ら[14]は，データ構造とデータの理解を目的とし，RDF データ検索と可視化を行う Paged Graph Visualization (PGV) を提案している．通常，グラフ可視化ツールは，グラフ全体の可視化を行った後で，無関係なデータを除外し，目的とするデータの探索・可視化を行うものが多い．これに対し，PGV では， 小さなグラフから段階的に大規模な RDF オントロジ関連のデータを探索し，可視化を行う．

2.2.2 データ分布の理解

後藤ら[3]は，メタデータを対象とした探索的検索行為を支援する DashSearch LD を提案している．探索的検索とは，探索目的を少しずつ明確化して知識を獲得する情報検索であり, 検索空間を推移しつつ，途中の絞り込み検索をするという行為を繰り返し行う．これにより，情報要求の具現化だけでなく，検索空間を理解することができるため，検索に最適なクエリの入力が可能になるとしている．飯塚ら[1]は，複数の RDF のデータをマージして得られた単一のデータセットから頻出部分グラフパターンを抽出することで，RDF データを対象としたグラフ検索に必要なクエリを自動生成する手法を提案している．これにより，データ構造の把握をすることなく，適切なクエリの選択が可能となるため，類似するデータや比較対象となるデータの検索が容易になるとしている．

2.3 テキストデータマイニング統合開発環境 TETDM

TETDM (Total Environment for Text Data Mining) [4][5]とはテキストデータマイニングのための統合開発環境であり，Java で構築されている．複数のデータマイニング技術を柔軟に組み合わせて使える環境を目標としており，複数の研究者・技術者が別々に開発したデータ分析ツールをそれぞれモジュールとして扱うことが可能となっている．ユーザはモジュールがセットされたパネルを任意の枚数並べて分析を行う．各パネルにおいて，処理ツールと可視化

[3] http://www.wikipedia.org/

[4] http://semweb.salzburgresearch.at/apps/rdf-gravity/index.html

ツールをそれぞれ1対1で組み合わせてセットする．また，複数のパネルを連動して利用することも可能であり，様々な観点から分析を行うことが可能である．

徳永ら[10][11]はオープンソースの統計解析を行う開発実行環境である R 言語[5] や，ニュージーランドのワイカト大学にて開発・公開されているデータマイニングツールである Weka[6] を，TETDM のモジュールとしてシステム化し，組み合わせて使うことを提案しており，TETDM と既存のデータマイニングツールの融合も進められている．

TETDM を用いたツールの開発事例として，高間ら[8]は，専門用語の候補を抽出し，専門用語辞書を作成する作業を支援するツールを提案している．また，梶並[2]は情報系の専門教育における TETDM の有効性を指摘し，実際に大学の講義において，受講者がモジュール開発を行った事例を報告している．

この様に，幅広い用途で TETDM は利用されており，今後も適用範囲がさらに広がることが期待されている．

3. 提案する分析支援ツール

3. 1 共通の述語を持つ主語の抽出・テーブル作成

従来のデータ分析やデータマイニングツールでは，表形式に構造化されたデータを対象とする場合が多い．これに対し，RDF データベースには，多くの述語 (プロパティ) が格納されており，その使用頻度は述語ごとに大きく異なることが一般的である．一般的な関係データベースにおいても欠損値が存在することはあるが，RDF データベースにおいて全ての述語を属性として表形式に構造化した場合，スパース度が非常に高くなる可能性が高い．従って，RDF データを分析するためにデータの構造化を行う場合，密度がある程度高い部分空間の抽出を行う必要があると考える．

本稿では，図 1 に示すように，最大公約数的に共通の述語を持つ主語の抽出を行い，行を主語，列を述語としたテーブルの出力を行うツールを提案する．これにより，データ分析に必要となる部分空間の抽出が可能になり、同時にどういった述語が多く使用されているかが把握できるので，データ分布の理解が可能となる．以下，部分空間抽出の手順を図 2 に示す具体例により説明する．始めに述語を抽出し，各述語を持つ主語の件数を求める．次に最大件数の述語とそれ以外の各述語において AND 検索を行ない，両述語を持つ主語の件数を求める．図 2 の例では，最大件数を持つ述語 B と他の述語それぞれについて AND 検索を行う．以降，最大件数をとる述語を AND 検索に追加していく．図 2 では，A が追加されている．AND 検索によるヒット数が全て 0 になるか，追加する述語がない場合，繰り返しを終了する．

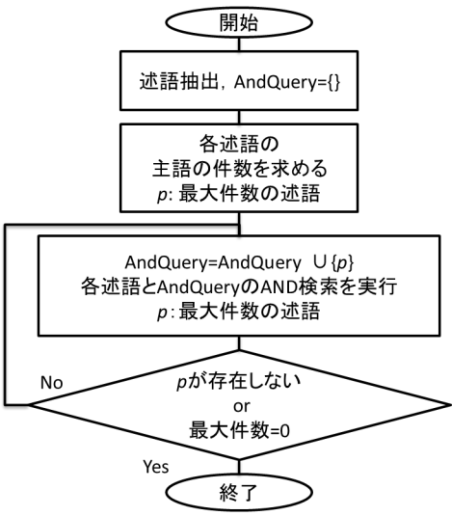


図 1: フローチャート．

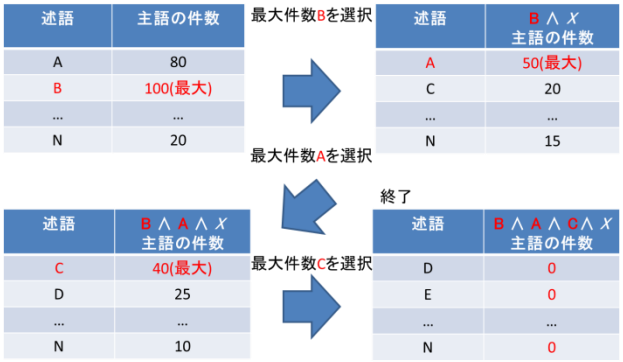


図 2: フローの具体例．

3. 2 複数エンドポイント間の共通リソース抽出

現在，LOD に代表されるように大量のデータが公開されており，それらのデータベースを横断して検索を行うことが，LOD 活用における一つの利点として挙げられる．Tim Berners-Lee は 2010 年の TED

[5] http://www.r-project.org/index.html
[6] http://www.cs.waikato.ac.nz/ml/weka/

University[9]にて，公開されている複数のデータを繋ぎ合わせることによって新たなデータの見方が可能になり，単体のデータからは得られない新たな結果が得られると述べている．しかし，複数のエンドポイント間のデータから意味のある結果，すなわち特徴的なリンク関係を発見するには，あらかじめデータ構造の把握が必要であり，単一のデータベースを対象とする場合と比較して，分析に手間がかかると考える．

そこで本稿では，接続性の高いリソース，すなわち 2 つの SPARQL エンドポイント間で共通するリソースを抽出し，そのリソースが持つ述語を把握するツールを提案する．これにより，複数データベースを横断した検索の手がかりを得ることが可能となり，データのリンク関係、すなわちデータ構造の理解につながることが期待できる．図 3 に示すように，2 つの SPARQL エンドポイントにおいて双方に出現しているリソースを共通リソースとして抽出する．また，各エンドポイントにおいて，共通リソースの主語あるいは目的語となる述語も同時に抽出する．

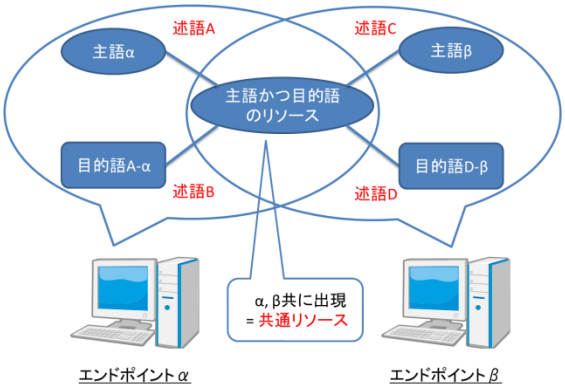


図 3: 複数エンドポイント間の共通リソース抽出．

3. 3 時間情報に基づくデータ分析支援

2.1 節で述べたように，RDF で記述されたデータには統計データやログデータも多くあり，それらには時系列データを扱っているものも多いため，時系列分析を行うことも想定される．あらかじめ分析したい時系列データが決まっていれば，そのデータのみ分析すればよいが，実際は様々な時系列データを探索的に分析したい場合も多いと考える．この場合には，時系列データの観点からデータ構造を理解する支援が有効であると考える．

そこで本稿では，時系列データを抽出し，ヒストグラムとして可視化するツールを提案する．時間情報が付与されている目的語を抽出し，分析対象とする．図 4 に示すように，目的語に時間情報をもつ述語を選択し，その述語の主語に対して他の述語，目的語を抽出する．図 5 に示すように抽出した述語と目的語から分析したい組み合わせを選択することにより，その目的語を横軸，時間の階級を縦軸としてヒストグラムを描画する．

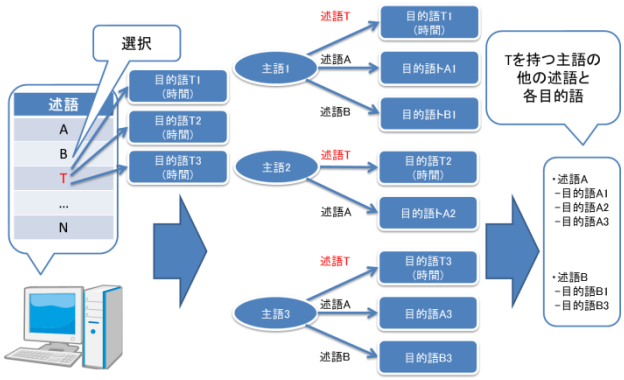


図 4: 時間情報に基づくデータ分析支援．

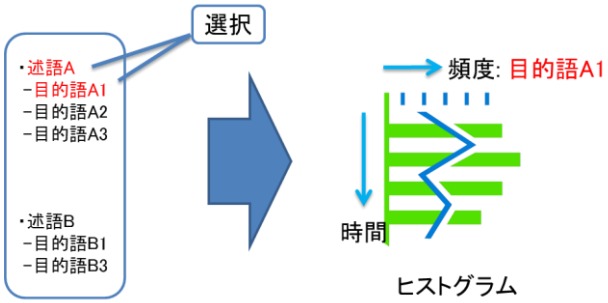


図 5: ヒストグラムにて可視化．

4. インタフェース

3 節で提案した 3 種類のツールについて，TETDM モジュールとして実装した．RDF の解析や SPARQL による問合せの処理に関しては，Apache Jena[7] を用いている．そのため，TETDM をコビルドする際に Jena の jar ファイルを含める様にしている．

図 6 に，3.1 節で提案した共通の述語を持つ主語の抽出・テーブル作成ツールのスクリーンショットを示す．このモジュールは，3.1 節のフローを行う処理ツールと，Type の表示・選択およびテーブルの表示を行う 2 種類の可視化ツールから構成される．モジュールの利用手順は以下の通りである．

1-1: SPARQL エンドポイントを入力
1-2: Type の表示及び選択
1-3: 検索する述語数の制限
2: テーブルを表示

[7] http://jena.apache.org/

3: テーブルをファイルに出力

ここで，ステップ 1-2，1-3 は任意であり，指定しない場合は全ての Type，述語を検索する．

図 7 は, ステップ 1 にて米国政府[8]の SPARQL エンドポイント を指定し，1-2 で述語 type の目的語として組織を表す FOAF の語彙である http://.../Organization を選択した場合の出力の一部である．この目的語を持つ主語は 176 件存在し，その中から主語 68 件, 述語 9 件のテーブルが抽出されている．

図 6: 共通の述語を持つ主語の抽出･テーブル作成ツールのスクリーンショット．

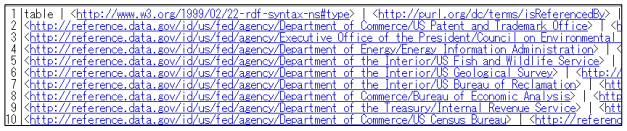


図 7: テーブルの出力例．

図 8 に, 3.2 節で提案した複数エンドポイント間の共通リソース抽出ツールのスクリーンショットを示す．このモジュールは，共通リソースとその共通リソースの述語を抽出する処理ツールと，その結果を表示する可視化ツールから構成される．モジュールの利用手順は以下の通りである．

1: 2 つの SPARQL エンドポイントを入力

2: 共通リソースを目的語に持つ述語 (A) ・共通リソース (B) ・共通リソースを主語に持つ述語 (C) の表示

図 9 は，ステップ 1 にて DBpedia Japanese [9]と DBpedia [10]の SPARQL エンドポイントを指定した場合の結果の一部である．http://dbpedia.org/ontology/Person が共通リソースの 1 つとして抽出されている．このリソースが関係する主語は両エンドポイントで共通であるが，目的語は異なっていることがわかる．

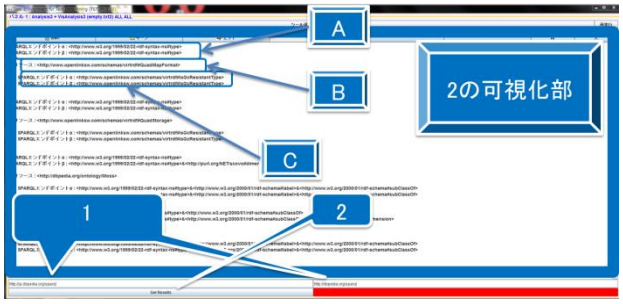


図 8: 複数エンドポイント間の共通リソース抽出ツールのスクリーンショット．

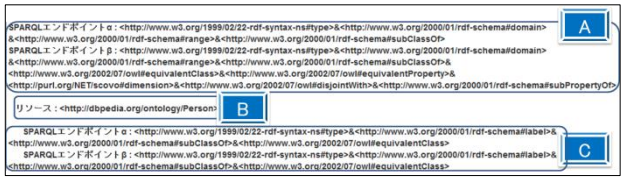


図 9: 共通リソースの抽出例．

図 10 に，3.3 節で提案した時間情報に基づくデータ分析支援ツールのスクリーンショットを示す．このモジュールは, 3.3 節で述べたヒストグラム生成を行う処理ツールと，述語・目的語の選択及びヒストグラムの描画をそれぞれ行う 2 種類の可視化ツールから構成される．現在では時間情報として XML Scheme の日付データ型である xsd:date，すなわち [yyyy-mm-dd]の形式の時間情報を持つデータに対して 5 つの階級幅において年の粒度での分析を可能としている．モジュールの利用手順は以下の通りである．

1-1: SPARQL エンドポイントを入力

1-2: 検索する述語数の制限

2: 述語の表示

3-1: 述語の選択

3-2: 選択した述語と共通の主語を持つ述語とその目的語の表示

4-1: 検索対象の述語の選択

4-2: 検索対象の目的語の選択

4-3: 年・階級間隔の入力

5: 選択した目的語の階級間隔内頻度の表示

ここで，ステップ 1-2 は任意であり，指定しない場合は全ての述語を検索する．

図 10 の左部は， ステップ 1-1 にて DBpedia Japanese，ステップ 3-1 で http://dbpedia.org/ontology/deathDate を選択し，ステップ 4-1 で http://dbpedia.org/ontology/deathPlace，ステップ 4-2 で http://ja.dbpedia.org/resource/東京都，ステップ 4-3 で年を 2000，階級間

[8] http://services.data.gov/sparql
[9] http://ja.dbpedia.org/sparql
[10] http://dbpedia.org/sparql

隔を 2 としそれぞれ指定した場合の結果である．このように時系列データの分析を行うことで，データ分析支援を行うことが可能である．

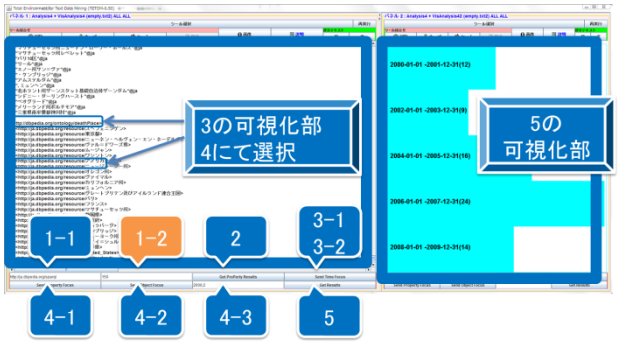


図 10: 時間情報に基づくデータ分析支援ツールのスクリーンショット．

5. まとめ

本稿では RDF の特徴を考慮した分析支援ツールを 3 種類提案し，TETDM を用いて実装したプロトタイプを示した．今後は各ツールの機能・インタフェースの向上のために，ユーザが指定可能な検索条件等の追加や，可視化表現の改善を行う予定である．3.1 節のツールに関しては，ユーザがテーブルの行・列の指定を可能にすることや，許容可能な欠損値の指定を可能とする予定である．3.2 節のツールに関しては，2 つ以上の SPARQL エンドポイントを指定可能にすることや，データ構造をよりわかりやすく提示することを考えている．3.2 節のツールに関しては，粒度としてより詳細な時間情報を扱えるように拡張する事を考えている．また，上記以外の新たなツールの開発や，開発したツールの有用性について評価・検討を行うことも予定している．特に様々なデータマイニングツールを統一的に扱うことを目的とした TETDM を用いて開発している意義として，提案・構築したツールを公開することで，フィードバックとしてより多くの意見を取り入れて反映させていくことも重要と考える．

参考文献

[1] 飯塚京士，佐藤宏之，イコプラムディオノ，村山隆彦: RDF データを対象としたグラフ検索におけるクエリ生成方式の検討，第 11 回セマンティックウェブとオントロジー研究会, SIG-SWO-A502-08, 2005

[2] 梶並知記: TETDM を利用した情報系専門教育の実践例，第 27 回人工知能学会全国大会，3B3-NFC-01a-5, 2013

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

[4] 砂山渉，高間康史，ダヌシカボレラ，西原陽子， 徳永秀和，串間宗男，松下光範: テキストデータマイニングのための統合環境，第 25 回人工知能学会全国大会, 1B2-NFC3-10, 2011

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

[6] 総務省: 総務省におけるオープンデータに係る実証実験，http://www.opendata.gr.jp/committee/docs/20130122_4_rikatu.pdf，2013/09/14 現在

[7] 総務省: 総務省におけるオープンデータに関する技術の検討状況について，http://www.kantei.go.jp/jp/singi/it2/densi/wg/dai1/siryou8.pdf，2013/09/14 現在

[8] 高間康史，阿部美里: テキストデータマイニング統合環境を利用した看護記録からの専門用語辞書作成支援ツールの提案，第 27 回人工知能学会全国大会，3B3-NFC-01b-1, 2013

[9] Tim Berners-Lee: The year open data went worldwide, http://www.ted.com/talks/lang/en/tim_berners_lee_the_year_open_data_went_worldwide.html, 2013/09/14 現在

[10] 徳永秀和，杉村拓哉: R と Weka を活用した TETDM ツールの開発，人工知能学会情報編纂研究会第 6 回, TETDM-01-SIG-IC-06-07, 2011

[11] 徳永秀和: R によるテキストマイニング用 TETDM モジュール開発，第 27 回人工知能学会全国大会，3B3-NFC-01b-2, 2013

[12] 松村冬子，小林巌生，嘉村哲郎，加藤文彦，高橋徹，上田洋，大向一輝，武田英明: Linked Open Data による博物館情報および地域情報の連携活用,じんもんこん 2011 論文集, Vol. 2011, pp. 403-408, 2011

[13] 村上勝彦， 山崎千里， 今西規: ヒト遺伝子データベース H-InvDB の RDF 化と Endpoint の公開，第 27 回人工知能学会全国大会, 1N5-OS-10c-3, 2013

[14] Leonidas Deligiannidis, Krys j. Kochut, Amit P Sheth: RDF Data Exploration and Visualization, CIMS `07 Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience, pp. 39-46, 2007

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5回)
SIG-AM-05-03

オンライン対戦型クイズシステムによる学習支援環境

Learning Support Environment by On-Line Competition Quiz System

渥美 峻[1*]　砂山 渡[1]　川本 佳代[1]　西村 和則[2]
Takashi Atsumi[1]　Wataru Sunayama[1]　Kayo Kawamoto[1]　Kazunori Nishimura[2]

[1] 広島市立大学大学院 情報科学研究科
[1] Graduate School of Information Sciences, Hiroshima City University
[2] 広島工業大学 工学部 電気システム工学科
[2] Department of Electrical Systems Engineering, Faculty of Engineering, Hiroshima Institute of Technology

Abstract: Recently, learning environments with e-learning or with gamification are developed. Since learners have been familiar with on-line communication through Social Network Services (SNS), it is expected that learning environment with network is ready to be available. In this study, a learning support environment by on-line competition quiz system is proposed. That is, learners can have high motivation to learn with ranking of experience points and human-human competition. Quiz is set as four multiple choices to continue learning easily and to be prepared various question sets through the Internet.

1　はじめに

近年の学習環境には，e-learning などネットワークを介して行われるものや，ゲームの要素を取り入れた学習が用いられるようになってきた．多くの学習者は，SNS などネットワークを通じたオンラインコミュニケーションに精通するようになってきており，またインターネットのブラウザゲーム，携帯電話やタブレット型端末で行うゲームがはやっている現状がある．そのため，ネットワークを利用して手軽に行えるゲーム型の学習環境が実現可能になり，その効果が期待されている．

そこで本研究では，オンライン対戦型クイズシステムによって学習を支援する環境を提案する．すなわち，対人での対戦ゲームとすることで，他人の存在と競争心による学習のモチベーションを与える．また学習意欲を維持するために，ゲームの要素となっている経験値やランキングを導入する．単純な四択クイズとすることで，手軽に学習を継続でき，よくある学校教育の問題のみならず，さまざまな分野の問題集合を準備できる環境とする．

以下，2 章で関連研究について述べる．3 章で構築したオンライン対戦クイズシステムの構成について述べる．4 章で学習意欲を向上させる改良版のクイズシステム，5 章でその評価実験について述べ，最後に 6 章で本稿を締めくくる．

2　関連研究

本章では，本研究で提案するオンライン対戦型クイズシステムに関わる先行研究について述べる．

これまでに，さまざまな目的に対して，オンラインクイズシステムを導入したアプリケーションが開発されている．教育目的としては，携帯電話を端末として電子メールによりクイズを出題する教育システム [1]，Web を使用して分散資源をパッケージ化した学習リソースを提供しながらクイズを出題するシステム [2]，Web を使用して出題とともに採点や集計ができるシステム [3] などが挙げられる．

これらのクイズシステムでは，システムが出す問題に対して個人が解答する形式となっている．本研究においては，ネットワークに接続した端末を利用して，複数人で競いながら，あるテーマについての学習を行える環境を構築する．また問題と解答の集合を元にした四択問題形式とすることで，さまざまなテーマについての学習を可能にする．

学習を目的としたオンライン対戦型システムの研究においては，各学習者が作成した問題と回答のカード群をランダムに裏向きに並べ，神経衰弱の要領で多くのペアを獲得することを目指して，能動的で競争的な学習を実現した研究 [4] や，論理的思考能力の育成に向けて，各学習者が作成したプログラム同士を対戦させ

ることで，よりよいプログラム作成に努めさせる研究 [5] などがあり，他の学習者の存在や競争心をモチベーションとして学習に結びつけている．これらは，スキル獲得を目指したシステムとなっているが，本研究においては知識の定着を目的として，その経験を積み重ねられるシステムを目指している．例えば英単語の記憶のような，ともすれば単調な作業を強いられる知識の定着において，競争的な環境を導入することで，意欲的に学習できる環境を目指す（3 章）．

e-learning はその有効性が声高に主張される一方，学習過程において学習者の学習意欲を高め，維持させることは困難と言われている．この問題を解決するため，回答を褒めたり叱ったりする機能や回答時間に制限を加えたシステム [6] や，学習者の意欲を高めるため，授業前や授業中に学習目標の設定と確認をさせたり，授業後に他の学生と成績や出席状況を比較する試み [7] などがある．他の学習者の存在や競争心によるモチベーションの増加は，一時的な効果はあると考えられるが，その意欲を維持して学習を進めてもらうためには，学習経過に伴った指標の提示が必要と考えられる．そこで本研究では，オンライン対戦クイズにおいて，その対戦内容と結果に対して学習者に経験値を与え，対戦を重ねることで経験値とそれに応じた段位が上昇する枠組みを構築する（4 章）．

教育システムへのゲームの導入について，古くは，ビジネス分野を対象に生産管理，在庫管理，リスク分析などのゲーミングモデルが開発されてきた [8]．以降，経済，経営分野を中心に，多様な分野の教育においてゲームを使用した教育システムが開発されている [9, 10, 11]．これらの教育システムの多くは，シリアスゲームとして，システマチックに知識やスキルを身につける学習を主目的に設計されており，必ずしもエンターテインメント性を重視していない．

一方，学習者の継続的な意欲の向上を目的として，近年ゲーミフィケーション [12] という言葉が用いられ，コンピュータゲームにおけるユーザのゲームへのはまりを実現する，特徴的なシステムや仕様を表す内容が注目されてきている．これには，ゲームの進捗に応じた報酬や経験値の付与，レベルシステム，課題リストとその達成を扱うミッションなどが含まれ，一つの行動の結果が積み重ねられて次につながることで，一定の満足感を与えるとともに，区切りを感じさせず，継続的に意欲を持続できる要因となっている．そこで本研究においても，これらの要素を取り入れたシステムにより，意欲を持って持続的に学習できる環境を目指す．

3 学習のためのオンライン対戦型クイズシステム

本章では，学習のためのオンライン対戦型クイズシステムの構成と詳細について述べる．

3.1 オンライン対戦クイズシステムのハードウェア構成

オンライン対戦クイズシステムは，実際にユーザが対戦を行うクライアントと，各クライアント間で対戦の制御をするサーバによって構成される．クライアントはサーバを介すことで他のクライアントとの通信対戦を可能とする．サーバは全てのクライアントから情報を受信し，通信状況の管理や対戦者同士の情報の共有，対戦時の時間の同期，対戦成績や解答履歴の保存を行う．

3.1.1 システムサーバ

サーバは新しいクライアントが通信を接続をしてきたときに，クライアントの情報を保持する．すなわちログインしたユーザに対して，クイズを行った日時，クイズとして出題された問題，解答結果，最終順位などを記録する．また，クイズの対戦を行う際に，問題の配信と各クライアントの解答情報の取得を行い，対戦の進行を制御する．

3.1.2 システムクライアント

クライアントはクイズ対戦を行うためにサーバと通信する．サーバに接続する際には，サーバに登録されているユーザごとの ID とパスワードを入力して通信を開始する．クライアント側で学習を希望する分野を選択すると，同じ分野を選択した学習者同士の対戦グループが生成され，クイズを開始することができる．

3.2 クイズの出題形式と出題分野

本システムにおいて出題されるクイズは四択問題としている．これは，解答入力時の手間を抑えるとともに，さまざまな分野の問題を容易に用意できるようにするための仕様となっている．一つの出題分野の問題数は 50 から 100 程度を想定しており，同一分野の問題おいては,「年号」や「人名」など同形式の解答が選択肢に並ぶようにする．これにより，インターネット上などで情報がまとめられているページから，50 から 100 の問題と解答のリストを収集することで，手軽に１つの分野の問題集合を登録することが可能となる．

また，用意する問題の分野は，中学や高校などでの学習に用いられる，英単語や年号などのいわゆる暗記ものを初めとして，さまざまな用語とその説明など，対になるものであれば何でも出題が可能となる．たとえば，お笑い芸人の名前とその人が属するコンビ名，ある製品とそれを販売している会社名，といったペアによる出題も可能で，問題と解答のペアを csv 形式の 1 つのファイルとして与えることで，1 つの分野を簡単に登録することが可能となっている．

出題の際には，出題されていない問題の解答が誤答の選択肢となる．そのため，同じ解答となる複数の問題が含まれないようにする必要がある[1]．

3.3 オンライン対戦クイズシステムの利用手順

学習者は，各自の PC 上でクライアントを起動し，ログインを行う．ログイン後に，学習を希望する分野を選択することができるので，1 つの分野を選択する．同時刻に同じ分野を選択している他の学習者がいる場合，その学習者同士でクイズ対戦を開始することができる．対戦終了後は，再び学習分野を選択する画面に戻り，学習を終了する際にはログアウトしてシステムを終了する．

4 学習者の意欲を補う対戦型クイズシステム

本章では，筆者らが過去に行った研究 [13] の実験結果を踏まえ，より学習者の意欲を補い，高い学習効果が得られるよう改善したシステムについて述べる．

実験結果から，以下のことがわかった．

1. 実力差がある対戦では，対戦に参加できない利用者が出る
2. 対戦クイズにより学習する前には一定の事前知識が必要

そこで，1. を改善する方法として，早押し形式により解答権を取得する方法をやめ，全員が解答でき，正解者で得点を分配する山分け形式に変更する．関連して，解答権を取得してから表示していた選択肢を，最初からすべて見えるようにする．これにより，全参加者が全問題に参加してクイズを楽しむことができるようにする．

また，2. を改善する方法として，事前知識がない参加者でも，他の参加者の解答を参考に解答を行うことができるように変更する．最低限の事前知識得るための学習を行った上で対戦に参加することを推奨するが，事前知識がない問題に対しても，他の解答者に追随して正解を自ら選択できる機会を増やし，クイズに参加できる機会を増やす．また対戦として，点差が開きにくくなることにより，緊張感を維持する効果を狙う．

これらに加え，継続して学習の動機づけを与える方法として，対戦結果に応じて参加者に経験値を与え，その積み重ねで段位が上がるようにする．これまでは一回ごとの対戦により区切られていたが，毎回の対戦の積み重ねにより，その達成度が目に見えるようにすることで，継続して参加する意欲の向上と維持を狙う．

以下でこれらの詳細について述べる．

図 1: 改良版オンライン対戦クイズシステムの画面

4.1 対戦クイズの形式

本節では，改良版のクイズシステムにおける対戦形式について述べる．

クイズ対戦は，4 択の山分けクイズとして，2 人から 4 人の学習者同士で行われ，1 回の対戦は 10 問で行われる．1 問の解答時間は最大 8 秒として，すべての学習者は 4 つのうちの 1 つの選択肢を選ぶ (図 1)．なお，他の学習者の解答内容は，学習者を表すアバターを用いて，各選択肢の横にリアルタイムで他の学習者にわかるように表示される．

正解時には，式 (1) による点数 Sc を与える．ただし，N は参加人数，$correct$ は正解者数を表し，$order$ は，最初に解答した人のみ 5 点が与えられる．すなわち，4 人対戦時には，正解者数が 4 人，3 人，2 人，1 人のときに，正解者にそれぞれ 10 点，16 点，20 点，40 点がベースとして与えられ，最初の正解者には 5 点がさらに加算される．

[1] 解答時に同じ選択肢が現われないようにすることは，プログラムで容易に対応が可能と考えられる．

表 1: 対戦結果と 1 問正解に対して与えられる経験値

現在の段位	1 位	2 位	3 位	4 位	1 問正解
1-10	90	70	50	40	3
11-20	80	60	40	30	3
21-30	70	50	35	25	4
31-40	60	40	30	20	5
41-50	50	25	20	10	6
51-	40	15	10	5	7

表 2: 「語句」の事前テストと事後テストの結果（被験者平均）

システム	事前テスト	事後テスト	得点差
経験値あり	42.3	61.4	19.1
経験値なし	41.8	51.1	9.3

表 3: 「四字熟語」の事前テストと事後テストの結果（被験者平均）

システム	事前テスト	事後テスト	得点差
経験値あり	56.8	74.8	18.0
経験値なし	56.9	71.4	14.5

$$Sc = 10 \times \frac{N}{correct} + order \qquad (1)$$

また，不正解時の減点はない．10 問終了時の合計点により，対戦結果の順位付けが行われる．

これにより，学習者全員がすべての問題に参加することができ，全く知識がない問題に対しても，他の学習者の解答を参考に解答することができる．また，単に他の学習者の解答に追随するよりも，自ら早く解答した学習者に 5 点をアドバンテージとして与え，不正解時の減点をなくすことで積極的な解答を促す．

4.2 学習者の学習意欲を補う機能

本節では，継続して学習の動機づけを与える方法として設けた機能，経験値，段位，ランキングについて述べる．

学習者には，対戦結果に応じて参加者に経験値を与え，その積み重ねで段位が上がるようにする．一回の対戦時に学習者が得られる経験値を表 1 に示す．

また，経験値が 100 に達するごとに，1 から始まる段位が 1 つずつ上がるようにし，参加者の間で段位の高い順にランキングを表示する．ランキングは，各対戦の終了時に確認できる．これにより，対戦時の積極的な解答ならびに正解を覚えようとする学習意欲の向上と維持を狙う．

5 改良版オンライン対戦クイズシステムの学習効果の検証実験

本章では，4 章で述べたシステムの学習効果を検証した実験について述べる．本実験により，改良したクイズシステムが学習者の意欲を補う効果，ならびに高い学習効果が得られるかどうかを検証する．

5.1 実験手順

被験者は電気工学を専攻する大学生，大学院生の 16 名とした．学習分野は，就職試験のための SPI 問題集から集めた,「語句（ことわざ）」と「四字熟語」とし，問題は各テーマ 100 問ずつの計 200 問とした．

実験は，被験者の学習意欲を補う要素として導入した，経験値，段位とランキングを明示する被験者と明示しない被験者に分けることで行い，このグループ分けは事前テストの点数によりグループ間で差が出ないように行った（以降これらのシステムを，経験値あり，経験値なしのシステムと呼ぶ）．事前テストは 100 問を 10 問ずつのセットに分け，その中で解答を選択してもらう形式で，1 問 1 点の 100 点満点として行った．

どちらのシステムを利用する場合においても，被験者には，約 3 週間の期間に,「語句（ことわざ）」と「四字熟語」のそれぞれについて，1 日に両テーマの対戦を 5 回ずつ，のべ 8 日間で合計 40 回の対戦を行ってもらった[2]．なお対戦は，8 名の被験者が他の被験者と均等にマッチングするようにした．評価は事前テストと同形式の事後テストの結果と，対戦クイズのログをもとに行うこととした．

5.2 実験結果と考察

表 2 と表 3 に，事前テストと事後テストの結果を示す．両テーマにおいて，経験値ありのシステムでは事前テストの点数に比べて大きく点数が上昇した．経験値なしのシステムでも，事後テストの点数は上昇したが，経験値ありに比べてその上昇量は少ない結果となった．

表 4 と表 5 に，被験者ごとの事前テストと事後テストの結果と最終段位，各問題で一番早く正解した割合，総正解率と平均解答時間を示す．事後テストの点数上位 5 名の平均点は，経験値ありとなしで,「語句」で 22.2 点,「四字熟語」で 5.6 点の差が生じており（点数上位

[2]各テーマの問題は全部で 100 問となっているため，各問題が平均 4 回出題される．

表 4: 「語句」の事前，事後テストの結果と最終段位，各問題で一番早く正解した割合 (1st) と総正解率 (%)，平均解答時間（事後テストの点数が高い被験者順，解答時間の下線は最初に正解した割合が高い 3 人を表す）

経験値あり					経験値なし				
事前	事後	段位	(1st) 正解率	解答時間	事前	事後	段位	(1st) 正解率	解答時間
65	94	31	(44.0)94.0	3.3	76	76	29	(22.0)95.2	3.5
85	94	26	(14.0)95.2	3.6	69	66	28	(37.2)93.6	3.4
72	89	29	(36.0)95.6	3.0	56	61	29	(24.4)94.0	3.4
49	86	38	(52.8)96.0	2.9	32	56	19	(4.4)80.0	4.4
3	58	15	(22.4)76.4	3.8	35	51	24	(38.4)82.8	3.4
42	45	20	(19.2)98.0	3.4	39	49	16	(11.6)82.0	4.3
12	14	22	(14.0)83.6	4.8	15	34	22	(42.8)83.6	3.4
10	11	20	(0.8)83.6	4.7	12	16	17	(3.2)88.4	4.5
平均	–	25.1	90.3	3.7	–	–	23.0	87.5	3.8

表 5: 「四字熟語」の事前，事後テストの結果と最終段位，各問題で一番早く正解した割合 (1st) と総正解率 (%)，平均解答時間（事後テストの点数が高い被験者順，解答時間の下線は最初に正解した割合が高い 3 人を表す）

経験値あり					経験値なし				
事前	事後	段位	(1st) 正解率	解答時間	事前	事後	段位	(1st) 正解率	解答時間
51	100	29	(44.0)96.0	2.7	77	100	32	(25.6)95.6	3.0
66	94	34	(38.4)95.6	2.9	57	93	29	(25.2)94.0	3.1
80	94	29	(9.2)96.4	3.5	74	86	27	(22.0)97.2	3.1
79	91	38	(62.0)95.2	2.5	70	79	32	(37.6)96.4	2.9
50	82	20	(23.2)83.2	3.8	50	75	29	(57.2)93.6	2.7
52	57	18	(22.4)94.8	3.1	67	67	18	(6.8)93.2	3.6
55	45	27	(9.6)88.0	4.2	29	36	16	(5.6)92.8	3.8
21	35	23	(2.4)89.6	4.1	31	35	18	(12.4)90.0	3.6
平均	–	27.3	92.4	3.3	–	–	25.1	94.1	3.2

5 名の t 検定：「語句」$t(4) = 3.73, p < .05$,「四字熟語」$t(4) = 2.49, p < .10$），事後テストで 80 点以上の被験者の人数を比べると，経験値ありのシステムを用いた被験者においては,「語句」で 4 人,「四字熟語」で 5 人となったのに対し，経験値なしのシステムを用いた被験者においては,「語句」で 0 人,「四字熟語」で 3 人となった．これらのことから，経験値ありのシステムを用いた被験者の半数は，システムの利用により高い学習効果が得られたことがわかる．

最終的に到達した段位の被験者平均を比較[3]すると，両テーマにおいて，経験値の有無でおよそ 2 段の差が生じた（t 検定：「語句」$t(7) = 1.97, p < .10$,「四字熟語」$t(7) = 2.62, p < .05$）．加えて，経験値は表 1 のように，段位が上がるにつれて増えにくい仕様となっているため，特に高段位の被験者については，低段位の被験者の 2 段差に比べて，より大きな差が生じていると考えられる．このことから，経験値ありのシステムにおいては，被験者が意欲的に経験値を増やそうと取り組んだことがわかる．

正解率はシステムによらず全被験者において高くなり，他の被験者の解答を参考にして，正しい解答を選択する経験を積めていたことがわかる．しかし，全被験者が事後テストで高い点数とならなかったのは，正解時にその解答を記憶に留めようとしたかどうかが強く関わっていると考えられる．すなわち，ゲームに勝つことだけを目的に他の被験者の解答をまねたり，惰性でゲームを続けても効果は薄いことが伺える．

表 4 と表 5 の，各問題で一番早く正解した割合が高い 3 人の被験者について，その平均解答時間（表中の下線部）を比較すると，経験値ありのシステムの方が早く解答していたことがわかる．これらの学習者においては，経験値の効果により，被験者の意欲と集中力を高め，結果として高い学習効果を得ることができたと考えられる．しかし，平均解答時間においては，シス

[3]経験値なしのシステムでは被験者に表示はしていない．

テムの有無によって大きな差がないことから，経験値ありのシステムにおいては，早く解答した人とそうでなかった人との解答時間に開きが生じており，被験者内に学習意欲の差があったことが伺える．これらのことから，経験値ありのシステムにおいては，一部の学習者の意欲を高められる効果があった反面，対戦に勝てない学習者には，経験値や段位のランキング表示によって，学習できている人との差を確認，意識させられるために，学習意欲が減退する可能性があったと考えられる．そのため，同程度の知識を持った人同士が対戦することや，テーマに対する知識が極端に少ない学習者には，各自で一定の解答ができるようになった後に対戦を行ってもらうこと，経験値に応じた適切なハンディキャップを設定することなどにより，多くの学習者の意欲を維持できるシステムへの改良が望まれる．

6 結論

本研究では，オンライン対戦クイズシステムを用いて，学習者の学習意欲を高め，効果的に学習を図るためのシステムを提案した．実験により，対戦型とすることで学習の動機づけが与えられること，ならびに学習継続の指標となる経験値を導入することで，学習意欲の維持と向上に役立てられることを確認した．今後の課題として，クイズの正解時にその解答を記憶に留めやすい工夫を導入することや，対戦で勝ちにくい学習者の学習意欲を維持する仕掛けの導入が挙げられる．

参考文献

[1] ヒンクルマンダン，奥田統己，ジョンソンアンドリュー，石川園代，グロースティモーシ：携帯電話を端末とするオンライン双方向教育システムの開発と効果測定，札幌学院大学人文学会紀要, Vol.83, pp.173 – 202, 2008

[2] 梅田恭子，原崇，安田孝美，横井茂樹：高速通信回線における XML を活用したオンラインクイズシステムの提案と開発，情報処理学会研究報告，コンピュータと教育研究会報告, Vol.2001, No.40, pp.25 – 32, 2001

[3] 岡田源也，舩曳信生，中西透，天野憲樹:WEB ベースの教育支援システム”NOBASU”の拡張と評価，電子情報通信学会技術研究報告，ET，教育工学, Vol. 107, No. 205, pp.75 – 80, 2007

[4] 望月智也，金子敬一：ネットワーク対戦型ゲームに基づく教育システムの開発，電子情報通信学会技術研究報告，ET，教育工学, Vol.102, No.697, pp.115 – 120, 2003

[5] 佐々木整，森川哲史，竹谷誠：対戦型ゲームを利用した論理的思考能力育成教材の開発，電子情報通信学会論文誌, Vol.J83-D-I, No.6, pp.635 – 643, 2000

[6] 島田麗聖，高橋健一，上田祐彰：e-learning システムにおける学習意欲向上についての研究，電子情報通信学会技術研究報告，ET，教育工学, Vol.109, No.163, pp.13 – 18, 2009

[7] アブドサラムダウティ，中山洋，山口正二：目標設定と評価教示による意欲向上を目的とした授業支援システム，教育情報研究：日本教育情報学会学会誌, Vol.25, No.1, pp.3 – 13, 2009

[8] G.W. Dickson: Research in Management Information Systems: The Minesota experiments, Management Science, Vol.23, No.9, pp.913 – 923, 1977

[9] 吉川肇子：防災教育にゲーミングを生かす，自然災害科学，Vol.24, No.4, pp.363 – 369, 2006

[10] 檜山敦，山下淳，西岡貞一，葛岡英明，広田光一，廣瀬通孝: ユビキタスゲーミング：位置駆動型モバイルシステムを利用したミュージアムガイドコンテンツ，日本バーチャルリアリティ学会論文誌, Vol.10, No.4, pp.523 – 532, 2005

[11] 臺有桂，西村多寿子，国井由生子，河原智江，田口理恵，田中奈津子，田高悦子：地域看護学教育におけるゲーミング・シミュレーションを活用した健康危機管理演習の試み，横浜看護学雑誌，Vol.2, No.1, pp.25 – 32, 2009

[12] 井上明人：ゲーミフィケーション，NHK 出版, 2012

[13] 渥美峻, 砂山渡：オンライン対戦クイズゲームによる汎用型学習システム, 第 26 回人工知能学会全国大会, 3L1-R-12-1, 2012

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5回)
SIG-AM-05-04

Web画像による語義・概念の視覚的な提示に関する検討

Visualization of Senses and Complex Concepts by Web Images

林 良彦[1] *
Yoshihiko Hayashi[1]

[1] 大阪大学大学院 言語文化研究科
[1] Graduate School of Language and Culture, Osaka University

Abstract: Images, deemed language-independent to some extent, can be utilized as an effective medium for representing/disambiguating user's search intent in interactive information access systems, such as a cross-language information retrieval system. If a target query concept (e.g. beaver) could be decomposed into a set of salient concept-feature pairs (e.g. beaver builds dams; beaver chews on woods, etc.), the depiction of each concept-feature pair may greatly help manifest the user's search intent. Given this motivation, we conducted an investigation on the Web-imageability of complex concepts (concept-feature pairs), which the psycholinguistic semantic feature database developed by McRae and his colleagues provides. More specifically, this paper reports on the human assessment results on how the images acquired from the Web by means of an image search engine, given a concept-feature pair, can adequately deliver the meaning. The results clearly show that physical motions and feeding behaviors performed by animate beings were more adequately depicted in the acquired Web images. This paper also discusses possible utility of visual images in interactive information access systems, and further argues mutual grounding of visual and linguistic information.

1 はじめに

単語が指示するアトミックな概念，さらには，語句により想起される複合的な概念が画像や記号により視覚的に提示できれば，インタラクティブな情報アクセスにおいて，ユーザの意図の明確化を支援する手段として利用できると考えられる．特に，画像や記号が (ある程度までは) 言語独立であることを考えれば，言語横断情報検索のような「言語の壁」を越える必要があるタスクにおいて，有用な手段となると考えられる．

このような考えに基づき，本稿の著者らは，言語横断情報検索において，クエリ単語の翻訳曖昧性解消の手がかりとして Web 画像[1]を用いることを提案し [7]，その有効性について評価を行った [9]．その結果，(1) 適切なクエリ単語の翻訳を選択するために画像手がかりは有用であるが，(2) 検索過程の効率性の向上には必ずしも寄与しない，ことが分かった．また，クエリ単語の表す概念が Web 画像によりどの程度表現されるかについての調査を行い，(1) 形状を有する具体的な事物が Web 画像により適切に表されうることが多いだけでなく，(2) 抽象的な概念も自然現象やある種の動作を表すものは，その状況を想起させる視覚的表現が Web 画像として取得できる場合が多いことを確認した [8]．

以上の一連の研究[2]では，品詞が名詞である単独の単語を調査の対象としていた．しかしながら，たとえ具体的な事物 (のクラス) を指示する単一の名詞であっても，我々はその事物が有する顕在性の高い特徴をいくつか心的に思い浮かべている，あるいは，心的に保持していると考えられる．例えば,「蛙」といえば，色が「緑色」であり,「ピョンと跳ねる」という行動が特徴的といったことが「蛙」の心的な意味表現の構成に含まれていると考えることができる．

本研究では，単語が指示する概念の有する一つの特徴を**意味属性**と呼ぶ．また，ある概念が有するある意味属性，すなわち，概念と意味属性の組み合わせ (例: 蛙はピョンと跳ねる) を**概念意味特徴**と呼ぶ．

ここで，概念意味特徴は対象概念を主体として含む一種の複合概念と考えることもできることを注意しておく．また，概念は，それに関する概念意味特徴の集合により規定されると見なすことができるので，各概念意味特徴を表す，あるいは，想起させる非言語メディ

*連絡先：大阪大学大学院・言語文化研究科・言語情報科学講座
〒 560-0043 豊中市待兼山町 1-8
E-mail: hayashi@lang.osaka-u.ac.jp

[1]Web から取得できる画像，特に画像検索エンジンにより取得できる画像を本研究では Web 画像と呼ぶ．

[2]NTT コミュニケーション科学基礎研究所との共同研究による．

アによる感覚的表現 (画像や音などのメディア情報) が適切に得られるならば，それらのメディア情報もやはりインタラクティブな情報アクセスなどにおける支援の手段として利用できることが期待できる．

ところで，従来より認知心理学，あるいは，心理言語学の分野では，単語を刺激として被験者に提示し，それに対する反応を収集することにより意味属性の体系を構築する試みが行われてきた [3]．そこで本稿では，そのような心理言語学的なアプローチにより収集された概念意味特徴から検索クエリを構成し，これを用いた Web 画像検索により収集された Web 画像が当該の概念意味特徴をどの程度，視覚的に適切に表しうるか (以下，**概念意味特徴の Web 画像適合性**) について調査した結果を報告する．

本稿の具体的な構成は以下のとおりである．まず 2 節では，心理言語学的な意味属性の体系，および，概念意味特徴のデータを提供するものとして利用するデータベース [12](以下，**McRae のデータベース**) について説明する．次に 3 節では，概念意味特徴の Web 画像適合性の調査の概要について述べる．主要なアイディアは，情報検索における性能評価指標を用いて概念意味特徴の Web 画像適合性を評価する点にある．続く 4 節では，Web 画像適合性の評価結果について検討する．より具体的には，検索クエリ生成の方略，概念意味特徴の特徴タイプ，概念意味特徴の構成要素の意味的な分類の観点からの検討を加える．また 5 節では，Web 画像をインタラクティブ情報アクセスにおける手がかり情報とする可能性や，言語と画像を相互にグランディングする可能性について議論する．

2 McRae のデータベース

2.1 概要

McRae のデータベース [12] は，基本レベル (basic level)[3]に属すると考えられる，生物・無生物に関する 541 種類の英語名詞により表される概念に対して，多数 (700 名以上) の被験者が心理実験的設定のもとで与えた英語による記述に基づいており，結果として，2,526 種類の意味属性と，これに基づいて上記 541 の概念に対して付与された総計 7,526 の概念意味特徴が提供されている．

本データベースにおいては，ターゲットとなる各概念の規定･記述に関する豊富な情報を有しているが，その中で最も基本的な情報は，各概念に対してある概念意味特徴が何人の被験者によって付与されたかという一種の頻度データである．

[3]日常生活においてオブジェクトを分類する際に参照されていると考えられる概念階層におけるレベル (抽象すぎず詳細すぎない) を指す．

表 1: 概念 beaver に対する記述の一部.

意味属性	BR ラベル	頻度
a_mammal	*taxonomic*	6
beh_-_builds_dams	*encyclopaedic*	20
beh_-_chews_on_wood	*visual-motion*	5
beh_-_cuts_down_trees	*encyclopaedic*	7
found_on_the_nickel	*encyclopaedic*	6
has_a_tail	*visual-form_and_surface*	24
has_sharp_teeth	*visual-form_and_surface*	24
hunted_by_people	*function*	7
is_brown	*visual-colour*	19
is_furry	*tactile*	18

例として，概念 beaver に対するデータの一部を表 1 に示す．表の各行は一つの概念意味特徴に対応する．これによれば，ビーバーは例えば，「哺乳動物の一種 (a_mammal) であり」,「ダムを作る (beh_-_builds_dams)」といった特徴があるが，これらの概念意味特徴はそれぞれ 6 人，20 人により与えられている．

2.2 意味属性キーワード

表 1 に例を示したように，多くの意味属性は，a_, has_, is_ などの一定の統制がなされた接頭辞部分と ”mammal”，”sharp_teeth”，”brown” といった比較的自由な単語 (列) による部分からなる．本稿では，接頭辞部分を**意味属性キーワード**[4]と呼ぶ．McRae らによれば，これらのキーワードは，意味属性をタイプ化するために導入されている．

各意味属性キーワードの意味はほぼ自明と思われるが，a_ (上位・下位関係)，has_ (全体・部分関係)，といった通常のオントロジー的な関係だけでなく，様態や機能を表す関係や，さらには，連想的な関係も含まれており，通常の語彙的オントロジーの言語資源には必ずしも含まれない豊かな情報が含まれている．これらの中で，beh_-_･beh_-_ というキーワードは，それぞれ生物・無生物オブジェクトの典型的な振る舞いを表すメタキーワードである．すなわち,「ダムを作る」(beh_-_builds_dams) という意味属性は，ビーバーを規定する上で顕在性の高い[5]振る舞いに関する意味属性である．

2.3 Rrain Region (BR) ラベル

公開されているデータにおける付加的な情報として興味深いものに，概念意味特徴への 2 つの体系による分類ラベルの付与がある．そのうちの一つは，Cree と

[4]公開されているデータからは，96 種の意味属性キーワードを抽出したが，これらの中には誤りと思われるものもあるほか，整理統合が必要と思われるものも散見される．

[5]実際，データベース中でこの意味属性を持つ概念は beaver のみである．

表 2: BR ラベルとその頻度分布.

BR ラベル	頻度
visual-form-and-surface	2,336
visual-color	424
visual-motion	339
tactile	245
sound	142
taste	84
smell	24
function	1,517
encyclopaedic	1,417
taxonomic	730

McRae により提案された脳領域階層分類 (brain region taxonomy) [2] と呼ばれる体系によるものである．表 1 の中央のカラムはこの体系におけるカテゴリのラベル (以下，**BR ラベル**) を示す．この体系には 10 種類 (概念階層を表す taxonomic も含んで) のカテゴリが含まれるが，そのうちの 7 つは知覚に関するものであり，特に視覚に関するものは 3 種 (form-and-surface, color, motion) に細分類されている．

表 2 に BR ラベルのデータベースにおける頻度分布を示す．延べ数においても概ね半数が知覚に関するカテゴリであり，特に視覚的なカテゴリがその大部分を占めていることが分かる．表 1 のビーバーの例における「木を齧る (”chews on wood”)」振る舞いは視覚的な動作であるのに対し,「ダムを作る (”builds dams”)」振る舞いは百科事典的知識として分類されている．

3 概念意味特徴の Web 画像適合性の調査

図 1 に意味属性の画像適合性の調査の概要を示す．以下，各ステップについて説明する．

3.1 概念意味特徴の選出

今回の検討においては，`beh_`，および，`inbeh_`の意味属性キーワードを持つ意味属性のみを対象とした．すなわち，生物・無生物の何らかの動作・振る舞いに関する概念意味特徴を調査した．この背景には，動作・振る舞いがある程度は画像化されやすいであろうという予想があり，また，どのような動作・振る舞いが Web 画像として取得可能であるかという興味がある．

この条件により，218 種の概念に対する 535 件の概念意味特徴を今回の検討の対象とした．表 3 に，535

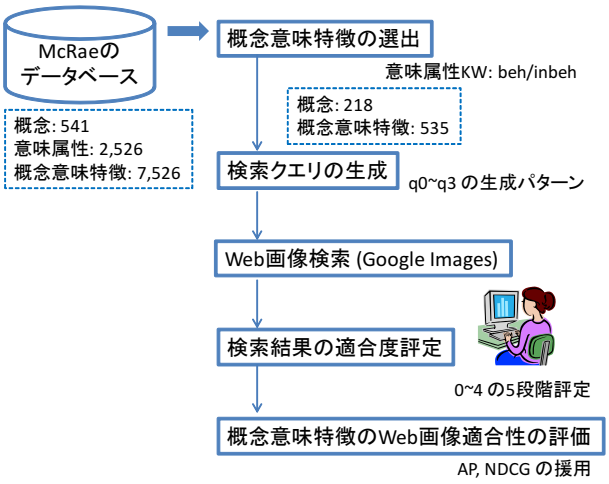


図 1: 概念意味特徴の Web 画像適合性の調査.

表 3: 概念意味特徴 (535 件) の分布.

	`beh`	`inbeh`
視覚的な BR ラベル	268 (f_a)	33 (f_c)
その他のラベル	151 (f_b)	83 (f_d)

件の概念意味特徴の `beh_`/`inbeh_`の区別，BR ラベルが視覚に関するものか・それ以外かの区別の組み合わせによる分布を示す．約半数 のものが生物による動作・振る舞いに関するものであり，かつ，視覚的であると分類されていることが分かる．本稿では，表における f_a から f_d を**概念意味特徴の特徴タイプ**と呼ぶ．

3.2 Web 画像の収集:画像検索クエリの生成方略

与えられた概念意味特徴に対して，以下の 4 つのパターンにより検索クエリを生成し，Web 画像検索エンジンを用いて Web 画像を収集する．

- q0: 概念を表す名詞のみを用いる (”beaver”)
- q1: McRae のデータベースに与えられている記述に基づく (”beaver builds dams”)
- q2: q1 において動詞を現在分詞形 (-ing 形) とする (”beaver building dams”)
- q3: 動詞を現在分詞形 (-ing 形) とし，概念名詞を後置する (”building dams beaver”)

検索クエリにより得られる検索結果はもちろん利用する検索エンジンに依存するが，上記の 4 つのパターンは概ね以下のような考えに基づいている．

- q0 および q1 はある意味でのベースラインである

図 2: Web 画像と適合度評定の例 (”beaver building dams”).

- 特に q0 は，対象とする概念意味特徴が当該の概念において非常に顕在性の高いものであれば，概念名詞のみによって，当該の概念意味特徴に対してある程度の適合性を有する画像が検索できるのではないか，という予測に基づく
- q2 および q3 においては，動作・振る舞いを直接的に表す動詞を現在分詞形 (-ing 形) としている．これは，画像 (特に写真) の持つ一般的な性質として，ある状況を描写するということがあり，そのために，Web コンテンツ中での画像に対する注釈・記述に動詞-ing 形が用いられているという可能性を考慮したものである

3.3 画像検索結果の適合度評定

上記によって生成した 2,126 件のクエリ[6]を Web 画像検索エンジン (Google Images) に送り，一つのクエリあたり最大 15 点，総計 27,970 点の画像を収集した．収集した Web 画像に対し，一名の評定者[7]により，0〜4 の 5 段階評価 (0:不適合，4:適合) により，検索された画像の概念意味特徴に対する適合度を評定した．

図 2 に検索結果と評定例を示す．また，評定結果の分布を表 4 に示す．適合度 2 以上のものを適合画像とするならば，半数近く (44.8%) の画像が適合画像という結果となった．以下の分析において，適合・不適合の 2 値判断が必要な場合は，この基準によっている．

3.4 概念意味特徴の Web 画像適合性の評価指標

ある概念意味特徴 (概念と意味属性の組み合わせ; 例: ”beaver builds dams”) に対して，前述のパターンにより生成したクエリを用いて収集した Web 画像集合がどの程度適切にその概念意味特徴を表しているかを，情報

[6]q1 パターンのクエリが”chicken cannot fly” のようになるものがあり，これらについては q3/q4 タイプのクエリを生成していない．

[7]実際には数名の評定者が分担して作業しているが，監督者が一定の基準となるよう調整を行っている．なお，評定者，監督者とも本論文の著者ではない．

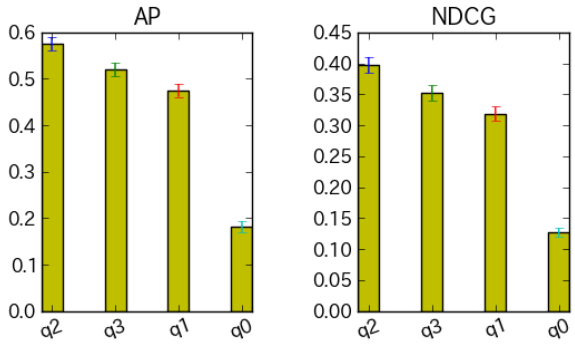


図 3: クエリタイプによる検索性能 (画像適合性) の比較.

検索においてよく用いられている平均精度 AP (Average Precision)，および，NDCG (Normalized Discounted Cumulative Gain) の二つの指標 [11] により評価することとした．すなわち，これらの指標の値が高ければ，当該の概念意味特徴の Web 画像適合性は高いものとして扱うということである．

4 概念意味特徴の Web 画像適合性の評価

4.1 検索クエリ生成の方略

クエリ生成の方略ごとに見た画像適合性の比較を図 3 に示す．この図には，動詞の-ing 形を用いる q2 タイプの画像適合性が他のタイプのクエリよりも有意に高い (AP，NDCG の双方について，$p < 0.01$; Mann-Whitney U-test) ことが示されている．このことから，本稿の以下の分析では，q2 タイプのクエリに対する結果について議論する．

ところで，図 3 には示されていないことであるが，ここまでの適合性判定をより甘い基準にし，適合度 0 の画像のみを不適合と考えると，q0 タイプのクエリの AP 値の結果が他のクエリタイプのものを上回る．このことは，特定の概念意味特徴が当該の概念において高い顕在性を有していることを示唆すると考えられる．例えば，`airplane` という概念に対する q0 タイプのクエリは，”airplane”であるが，これにより検索される画像集合には飛行機が飛んでいる画像が多く含まれる．もちろん，”airplane flies”という概念意味特徴はデータベースに存在するのであるが，わざわざこれを検索クエリにする必要はないことになる．

4.2 概念意味特徴の特徴タイプ

表 3 に示した f_a から f_d の特徴タイプから見た画像適合性の比較を図 4 に示す．

表 4: 評定された適合度の分布.

適合度	0	1	2	3	4
画像件数	6,705	8,734	4,025	1,910	6,596
総計: 27,970	15,439 (55.2%)		12,531 (44.8%)		

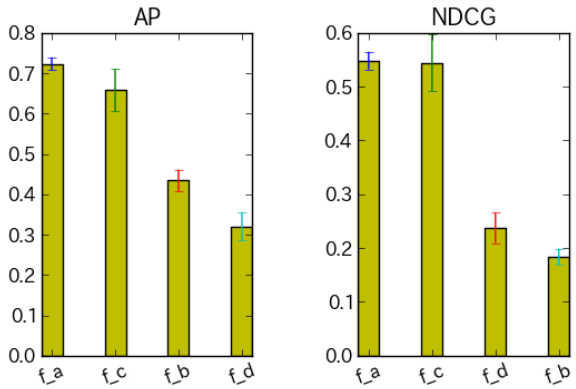


図 4: 概念意味特徴の特徴タイプによる検索性能 (画像適合性) の比較.

この図からは概ね以下のことが言える.

- 視覚に関する BR ラベルを有する概念意味特徴タイプ (f_a と f_c) の画像適合性がそれ以外の知覚ラベルを有する概念意味特徴タイプ (f_b と f_d) よりも有意に高い. このことは, McRae のデータベースで付与されている視覚に関する BR ラベルの妥当性を裏付けていると考えられる
- しかしながら, f_a と f_c の間の差は有意ではなく (AP については, $p = 0.192$; NDCG については, $p = 0.440$), 視覚に関する概念意味特徴を対象とする限り, 動作・振る舞いの主体が生物であるか無生物であるかという観点からは, 画像適合性に有意な差がない
- ただし, 無生物の主体の振る舞いを表す f_c タイプにおいては, 画像適合性のバラ付きが非常に大きい
- 視覚以外の知覚ラベルを有する概念意味特徴タイプ (f_b と f_d) に関しては, AP 値において両者に有意な差 ($p < 0.01$) が見られた

4.3 概念意味特徴の構成要素の意味的な分類

本節では, Princeton WordNet [8] における lexicographer file[9]を粗い意味分類の体系として用い, 概念意味特徴の構成要素 (概念を表す名詞, 動作・振る舞いを表す動詞, および, これらの組み合わせ) の意味的な特徴と Web 画像適合性との関わりを調べる.

lexicographer file は, そもそもは, WordNet を編纂する作業上の要請から設けられたもの [13] であるが, いくつかの研究 (例えば [10]) において, 粗い意味分類の体系として用いられている. 本研究でも lexicographer file を意味グループと考え, そのファイル名を意味グループのラベルとして用いる. このために, McRae のデータベース中の対象となる概念名詞, これらに対する概念意味特徴において使われる動詞に対して, 人手により意味グループラベルを付与した.

4.3.1 概念を表す名詞

表 5 に本稿で対象とする 218 種の概念名詞に対する意味グループの分布を示す. McRae ら [12] は, 生物 (living thing)・無生物 (nonliving thing) を研究の対象としたと述べているが, いくつかの概念名詞は, WordNet の分類においては, 抽象的なカテゴリ (*communication*, *possession*) に属する. 約 6 割強 (*animal*, *plant* に属するもの) が生物にあたり, それ以外の具体物に関するカテゴリが無生物にあたる.

表 5: 概念名詞の意味から見た分布.

意味グループ	概念名詞の数
animal	120
artifact	78
food	11
plant	3
communication, *substance*	各 2
body, *possession*	各 1

図 5 に概念名詞の意味グループ別により検索性能 (画像適合性) を比較する.

AP と NDCG の二つの指標によって傾向の違いが見られるが, 画像適合性が比較的高いグループ *animal*, *food*, *substance* と比較的低いグループ *artifact*, *plant* に分けることができ, 特に動物 (*animal*) は, 比較的安定して画像適合性が高いとみることができる一方で, 人工物 (*artifact*) は一貫して画像適合性が低い傾向にあるとことが分かる. この傾向は, 図 4 に示した傾向, すなわち, f_a, f_c タイプの概念意味特徴が f_b, f_d タイ

[8] http://wordnet.princeton.edu/
[9] http://wordnet.princeton.edu/man/lexnames.5WN.html

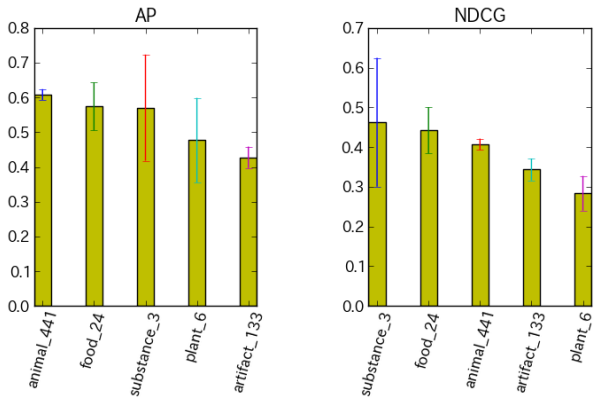


図 5: 概念名詞の意味グループによる検索性能 (画像適合性) の比較.

プの概念意味特徴よりも高い画像適合性を示すという結果と符合している．以上より，Web からは動物による動作がより画像として得やすいという傾向があることが分かる．

4.3.2 動作・振る舞いを表す動詞

表 6 に本稿で対象とする 535 件 の概念意味特徴において用いられる動詞に対する意味グループの分布を示す．この意味グループを用いて，Web 画像適合性の高いと思われる動詞の意味的な傾向を調べる．

表 6: 動作・振る舞いを表す動詞の意味から見た分布.

意味グループ	動詞の数
motion	169
creation	117
consumption	108
communication	34
change	19
contact	16
competition, perception	各 15

図 6 に動詞の意味グループ別により検索性能 (画像適合性) を比較する．

まず，*consumption*, *motion* の二つの意味グループは，これらに属する動詞の数も多く，また画像適合性も比較的安定して高い．*consumption* に属する代表的な動詞は，”eat”, ”chew”, ”drink” などであり，生物の摂食行動に関連するものが多い．*motion* に属する代表的な動詞は，”fly”, ”swim”, ”crawl”, ”travel” などであり，やはり生物の基本的な行動である移動に関するものが多い．表 5 に示したように，McRae のデータ

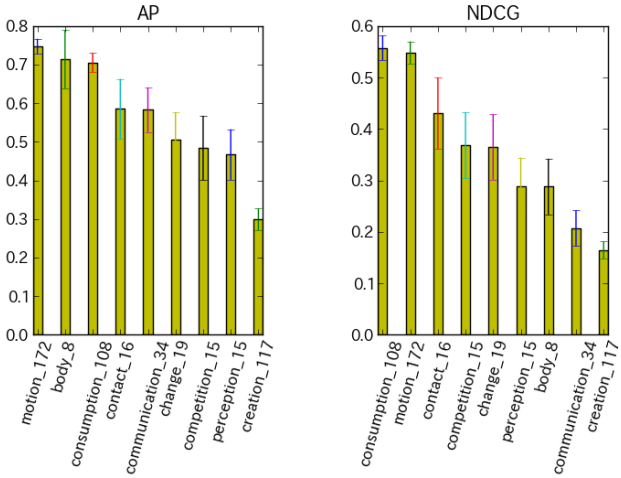


図 6: 概念名詞の意味グループによる検索性能 (画像適合性) の比較.

ベースには *animal* の意味グループに属する概念が多く含まれていが，動物の摂食行動や基本的な移動動作は，画像として Web 上から取得しやすい傾向にあることが分かる．

一方，*creation* の意味グループは，頻度は高いものの画像適合性は低い傾向にある．この意味グループに属する代表的な動詞は，”produce”, ”build”, ”make”, ”give”, ”do” などである．これらの動詞による意味特徴においては，焦点は create されるものに当てられると考えられるが，create されるものが具体的な形状を持つものとは限らない (例: ”accordion produces music”) ことから，画像適合性は低い傾向にあるものと考えられる．

4.3.3 概念名詞と動作・振る舞い動詞の組み合わせ

最後に，意味グループのレベルでみた場合の概念名詞と動詞の組み合わせで頻度の高いものを表 7 に示す．

表 7: 頻度の高い概念名詞+動詞の組み合わせ.

意味グループ組み合わせ	頻度
animal+motion	126
animal+consumption	114
animal+creation	63
artifact+creation	49
animal+communication	35

図 7 に概念名詞と動詞の組み合わせにより検索性能 (画像適合性) を比較する．

ここまでの議論から予想されるとおり，概念名詞:*animal*+動詞:*motion* の組み合わせが AP，NDCG 両方の

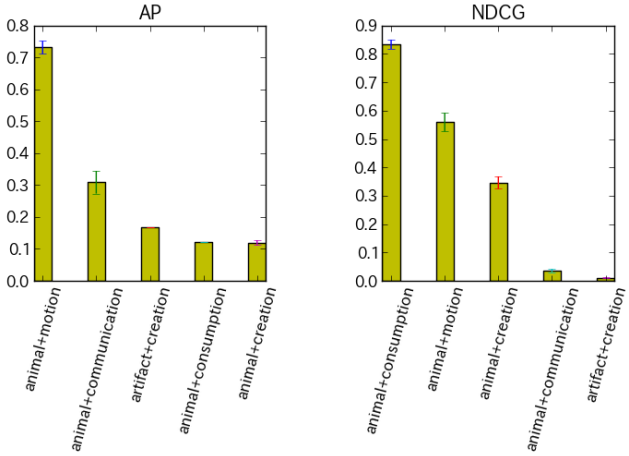


図 7: 概念名詞と動詞の組み合わせによる検索性能 (画像適合性) の比較.

尺度で安定して高い検索性能 (画像適合性) を示している．その一方で，概念名詞:*animal*+動詞:*consumption* の組み合わせは，NDCG では高い画像適合性を示しているが AP では低いという傾向があり，検索された画像の中にはにおいては，必ずしも適合度が高くないものも混在しているという状況が想像される．

5 関連研究と議論

5.1 関連研究

近年，概念の具体性 (concreteness) [14] に代表されるような，心理言語学的な属性を自然言語処理や情報検索に導入しようという傾向が見られる [10, 17]．このような心理 (言語) 学的な属性の中で，最も興味深いものは，やはり McRae のデータベースで提供されているような意味属性 [10] であろう．意味属性を用いて記述された概念意味特徴は，当該の概念に対して人々が持つ顕在性の高い情報を担っており，既存の言語資源からは得がたいタイプの情報を提供する．Silberer ら [15] は，McRae のデータベースによる意味属性の情報を従来からのテキスト情報と統合することにより単語の意味的類似度・連想度の精度が向上することを示している．

このような意味属性の持つ問題点の一つとして，意味属性が被験者実験により導出されるためにそのカバー率が低いことから，実用的な言語処理への適用が限定されることが挙げられる．この問題に対して，Baroni ら [1] は，大規模コーパスから有用な意味属性を抽出する手法を提案している．一方，Silberer ら [16] は，McRae のデータベースにおける意味属性に相当する高次の視覚的属性 (visual attribute) を与えられた画像から抽出することを試みている．より具体的には，ImageNet [4] (WordNet 中の synset と対応する画像を収集したデータベース) における画像と McRae のデータベースにおける概念記述特徴を訓練データとして用い，視覚的属性抽出のモデルを学習している．

[10] semantic feature norm (意味属性記述) という用語が関連研究を含めて用いられているが，本稿では単に意味属性とした．

5.2 議論: 言語と画像の相互グランディング

言語表現の意味を画像に求めること，逆に画像が伝える意味を言語として表現することは，相互に関連した問題であり，また，情報アクセスという応用においても重要である．後者は，画像コンテンツに対する言語注釈を与えるという問題であり，情報アクセスの高度化に直結する課題である．前者に関連しては，例えば，辞書エントリに対して意味的に適合する画像を求めようとする藤田らの試み [6] や，Web 画像を取り込んでマルチメディア百科事典を構成しようとする藤井らの試み [5] があるが，これらの大きな意味でのグランディングを行うとする場合と，筆者らの以前の研究のように意味の区分のための手がかりとして画像を用いる場合とでは，意味を絶対的に表すものか，複数の意味を相対的に区分するためのものかという違いを考慮する必要があり，情報アクセスにおける手がかりとしては，より後者の条件に適合した画像を収集・判別する手法の開発が必要となる．

本研究で目指すところの一つは，概念意味特徴で表されるような概念を主体 (主語) とした複合的な概念を画像で表現する可能性の探求であり，主体となる概念名詞は生物・無生物を問わず具体物であった．しかしながら，特に他動詞を用いた概念意味特徴では，目的語として現れるものは必ずしも具体物とは限らない (例: accordion produces music)．このことから，抽象概念 (この例では `music`) に適合する，あるいは，それを想起させる画像を具体物を主体とした概念意味特徴に適合する画像からある程度は得ることができる可能性があると考えられる．

6 おわりに

言語表現が担う意味を言語独立性の高い非言語メディア情報を用いて適切に表現することができれば，インタラクティブな情報アクセスにおける有用な手がかりとして利用できるほか，コミュニケーション支援や言語学習などにも適用が可能であると考えられる．

本研究では，非言語メディア情報として画像を取り上げ，ある概念 (例:beaver) に対して心理言語学的な手法により導かれた概念意味特徴 (例:beaver builds dams;

beaver chews on woords) [12] がどの程度 Web 画像検索により収集できる Web 画像により表現されうるかを調査した．この結果，動物の摂食行動や，主に物理的な動き・移動に関した基本行動が Web 画像により表現されやすいことを確認した．また，英語による概念意味特徴において，より良い画像検索結果を得るためのクエリ生成手法について評価を行い，動詞の-ing 形を用いることが有用であることを確認した．

しかしながら，情報アクセスシステムに組み込める形の支援機能を実現するためには，ある概念意味特徴に基づいた検索クエリによって検索された画像が，その概念意味特徴をどの程度適切に表しているかを予測する手法の開発が必要であり，Web コンテンツの言語的な解析と画像解析を統合した手法の開発が必要である．

謝辞

本研究は，JSPS 科研費#24650123 (挑戦的萌芽研究:「画像による言語的意味のグランディングの可能性の探求」) の助成を受けた．

参考文献

[1] Baroni, M., Murphy, B., Barbu, E., and Poesio, M. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, (34):222–254.

[2] Cree, G.S., and McRae, K. 2003. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology*, 132:163–201.

[3] Cree, G.S., and Armstrong, B.C. 2012. Computational Models of Semantic Memory. In: Spievey, M.J, et al. (Eds), *The Cambridge Handbook of Psycholinguistics*, pp.259–282, Cambridge University Press.

[4] Deng, J., Dong, W., Socher, T., Li, L.J, Li, L., and Fei-Fei-, L. 2009. ImageNet: A large-scale hierarchical image database. *Proc. of CVPR 2009*, pp.248–255.

[5] Fujii, A., and Ishikawa, T. 2005. Toward the automatic compilation of multimedia encyclopedias: Associating images with term descriptions on the Web. *Proc. of Web Intelligence 2005*, pp.536–542.

[6] Fujita, S. and Nagata, M. 2010. Enriching dictionaries with images from the Internet. *Proc. of COLING 2010*, pp.331–339.

[7] Hayashi, Y. and Savas, B. and Nagata, M. 2009. Utilizing images for assisting cross-Language information retrieval on the Web. *Proc. of WIRSS 2009*, pp.100–103.

[8] Hayashi, Y. and Nagata, M. and Savas, B. 2010. Exploring the visual annotatability of query concepts for interactive cross-language information retrieval. *Proc. of AIRS 2010*, pp.379–388.

[9] 林 良彦，永田昌明，Bora Savas. 2012. 言語横断情報検索における画像手がかりを用いたインタラクティブな翻訳曖昧性解消の評価. 情報処理学会論文誌:データベース, Vol.5, No.2, pp.26–35.

[10] Kwong, O.Y. 2012. *New Perspectives on Computational and Cognitive Strategies for Word Sense Disambiguation*, Springer.

[11] Manning, C.D, Raghavan, P., and Schütze, H. 2008. *Introduction to Information Retrieval*, Cambridge University Press.

[12] McRae, K., Cree, G.S., and Seidenberg, M.S. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments, and Computers*, 37(4):547–559.

[13] Miller, G. 1998. Nouns in WordNet. In: Fellbaum, C. (Ed), *WordNet, An Electronic Lexical Database*, pp.23–46, The MIT Press.

[14] Paivio, A., Yuille, J.C., and Madigan S.A. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76 (1, Part 2):1–25.

[15] Silberer, C., and Lapata, M. 2012. Grounded models of semantic representation. *Proc. of the 2012 Joint Conference on EMNLP*, pp.1423–1433.

[16] Silberer, C., Ferrari, V., and Lapata, M. 2013. Models of semantic representation with visual attributes *Proc. of ACL 2013*, pp.572–582.

[17] Tanaka, S. Jatowt, A., Kato, M.P., and Tanaka, K. 2013. Estimating content concreteness for finding comprehensible documents. *Proc. of The Sixth ACM WSDM Conference*, pp.475–484.

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5回)
SIG-AM-05-05

映像コンテンツとコメントを利用した 指示的動画要約生成・提示手法の提案

Proposal of Indicative Video Summary Generation and Presentation Method using Video Contents and Comments

于 多[1]　高間　康史[1*]

Duo Yu[1] and Yasufumi Takama[1]

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

Abstract: 本稿では，ニコニコ動画などのコメントが付与された動画から指示的要約を生成する手法を提案する．動画は Web 上の主要コンテンツの一つであるが，テキストコンテンツと異なり，見るべき動画の効率的発見は困難である．提案手法では，映像の特徴とコメントの関連性を考慮し，多数のユーザが共感している区間を複数抽出する．抽出した各区間から GIF アニメーションを生成し，それらを並列提示することで効率的な閲覧を支援する．本稿では並列提示に関する予備実験の結果について報告する．

1　はじめに

本稿では，ニコニコ動画[1]などのコメントが付与された動画から指示的要約を生成する手法を提案する．2005年頃から，動画共有サービスが急速に発展してきている．動画共有サイトも次々と登場し，特に，YouTube[2]やニコニコ動画に代表される動画共有サイトの利用者が急増している．2011年のマイボイスコムのインターネット調査結果[1]によれば，直近１年間に利用したことがある動画共有サイトは12,477名の回答者のうち「YouTube」が93.9%，「ニコニコ動画」が46.5%となっている．しかし，公開コストの低下により膨大な量の動画が投稿されるようになった結果，短時間で興味ある動画を探すことは困難となっている．

ユーザの動画検索や選択を支援する有効な手段の一つに動画要約が挙げられる．動画要約とは，動画内の各シーンがどのような内容を表現しているかを静止画あるいは短い動画により提示する手法である．利用目的に応じ，指示的動画要約と報知的動画要約の二つのタイプに分けられる[2]．指示的動画要約は，元動画を観るべきかどうか，自分の関心に合うかどうかを判断する際の手がかりとして用いられる要約であり，動画の重要な部分や視聴者の関心を引く部分を抽出して提示する．これに対し，報知的動画要約は元動画の代替として用いる要約であり，動画の主要な内容を，全体構成もわかるように提示する必要がある．本稿では，指示的動画要約を対象とする．

一方，ニコニコ動画をはじめとする，動画にリアルタイムのコメントを付加できるサービスが注目されている．このサービスにより，ある特定のシーンに対して投稿されたコメントは右から左へ流れるように動画の画面上に表示される．このようなコメントは動画コンテンツと関連するため，指示的動画要約生成の際に，重要な手がかりとして利用可能と考える．

本稿では，ニコニコ動画を対象として指示的動画要約を生成する事を目的として，映像コンテンツとコメントの両方を考慮した重要区間抽出手法を提案する．また，抽出した各区間を並列に提示する手法も提案する．提案手法では，映像の特徴とコメントの関連性を考慮し，元動画から多数のユーザが共感している区間を複数抽出する．抽出した各区間から GIF アニメーションを生成し，それらを並列提示することで，効率的な閲覧を支援する．本稿では，複数区間の並列提示手法に関する予備実験の結果について報告する．

2　関連研究

動画要約に関する研究には，これまで様々な種類

[1] http://www.nicovideo.jp/
[2] http://www.youtube.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 東京都日野市旭が丘６－６
E-mail: ytakama@sd.tmu.ac.jp

や手法がある．本節では，それらを三つの部分に分けてまとめる．

2.1 映像コンテンツを用いた映像要約

映像要約に関する従来研究では，映像コンテンツ内の特徴を用いたものが多い．料理映像の特徴を利用した要約手法[3]では，調理の全体的な流れを視覚的・直感的に理解することを目的としている．調理動作および料理や材料を示す部分を特に重要部分とし，動き検出により重要部分を決定し，要約映像を作成する．

映像中の情報を用いたニュース映像要約手法[4]では，ニュース映像の全体を把握することを目的として「冒頭部」を重要部分と定義し，カット点検出手法，人物領域認識手法とカメラワーク・テロップ検出手法を用いてこれを抽出することで，映像要約を実現する．これらの研究の多くは，特定ジャンルに対象を絞り，その対象に固有の特徴を有効活用しているが，本稿では対象に依存しない映像要約手法を対象とする．

また，映像要約の要素技術として用いられる代表的な技術にカット検出があり，さまざまな手法が提案されている．多次元特徴空間解析に基づく映像のカット検出手法[5]では，映像フレームから抽出した特徴量が特徴空間内で連続的に移動する特性を利用し，多次元特徴空間での軌跡を解析する．カット点検出手法に関する研究では，映像のジャンルにより用いられる特徴が異なるため，ジャンルに応じてカット点の基準を決めることが検出精度向上のために重要である．

2.2 コメントを用いた動画要約

YouTube やニコニコ動画など利用者が動画を投稿できるサービスの出現とともに，動画タイトルやタグ･コメントといった, 投稿動画に付与されるテキストを利用した動画要約に関する研究が増加してきている．特に，ニコニコ動画は他の動画共有サイトと異なり，動画にリアルタイムのコメントを付加できるサービスを提供しているため，ニコニコ動画を対象とする研究が多く存在する．

青木らは，映像に伴う音楽のサビ部分の検出や映像要約の生成を，コメント頻度に基づき行う手法を提案し，その可能性について検証している[6]．この研究では，様々なジャンルの映像に対して評価した結果，うまくいくケースとそうでないケースがあることを示している. 映像要約は一部失敗しているが，要約の正しさではなく面白さを評価するユーザにとら，受け入れられる可能性があるとしている[6]．多胡らは，動画に投稿されたコメントを意味的，量的に分析し，動画の各シーンに対するソーシャルアノテーションとみなしてシーン検索・動画要約を行うシステムを提案している[7]．動画を等間隔のユニットに分割し，コメントに基づくキーワードオントロジーを利用してそれらにスコア付けする手法を用いる．キーワードオントロジーは，人手でキーワード群を構築したものであるため，精度が高いものの構築コストが問題になると考える．

2.3 動画要約の可視化手法

これまで，動画要約に関する研究の多くは元動画より短い動画を作成して提示する手法を採用している．代表的なものに，動画内容の把握や特定シーンの再視聴を支援するため，動画要約をいくつかのサブネイルで並べて表示する手法[8]や，動画要約画像を対応する時間軸に配置して可視化する手法[9]などがある．また，小関らは，動画から抽出した画像を漫画的技法を用いてアニメーション化する「ぱらぱらアニメ」を提案している[10].

3 映像コンテンツとコメントを利用した指示的動画要約生成手法

本稿では，効率的な閲覧を支援する指示的動画要約の生成を目的として，映像コンテンツとコメントを用いた重要区間抽出手法と並列提示手法を提案する．指示的動画要約を生成するため，動画の重要な部分や視聴者の関心を引く部分を抽出する必要がある．動画の重要な部分を抽出する際には，画素値など映像内の情報だけではなくコメントなど映像以外の情報も同時に分析することで，動画作成側と視聴者側の両方の視点を考慮することができると考える.

指示的動画要約生成手法では, 図 1 に示すように, まず, ニコニコ動画から動画, 基本情報（動画の ID, タイトル，再生時間とコメント数など）とコメントを取得して記録する．次に，画素値を利用して元動画を分割する．分割した動画の各区間に対し，コメントを利用して重要度を付与し，重要度の高い順に四つの区間を選択する．さらに，選択した区間それぞれについて，コメントやキーワードが集中する部分空間を抽出することで区間の短縮を行う．

並列提示手法では，単一の動画を生成する従来手法とは異なり, 短縮した 10 秒間の区間に対してそれぞれ GIF アニメーションを作成し，それらを時間順に並べて表示する．

3 節では，区間分割，区間選択･短縮，並列提示のそれぞれについて具体的に説明する．

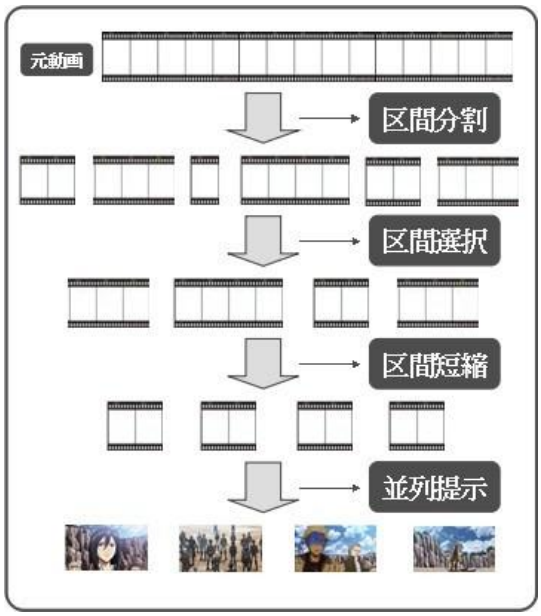


図 1 指示的動画要約の生成手順

3.1 区間分割

区間分割では，映像の瞬間的な変わり目（カット点）を検出することで元動画を分割する．まず，元動画から連続しているフレームを全部取得する．時間的に隣接する二枚のグレースケール化したフレーム $f(t\text{-}1)$と $f(t)$の間で, 対応する画素(x,y)の濃淡値の差 I_tを式(1)により求める．

$$I_t = \left| I(x, y, t) - I(x, y, t-1) \right| \qquad (1)$$

I_tがある閾値dより大きい場合，(x,y)を変化点としてカウントし，変化点の総数S_tを計算する．図2は，「アニメ」,「ゲーム」,「音楽」,「エンタメ」,「生活」と「スポーツ」の6ジャンルからそれぞれ一般的な動画を取得し，S_tを計算した結果である．横軸は，動画のフレーム数を示し，縦軸はS_tである．S_tが大きい場合，フレーム間の変化が大きいとみなせる．図2を見ると,「ゲーム」ジャンルの動画は他のジャンルよりもフレーム間の大きな変化が多く発生していることがわかる．また,「アニメ」と「スポーツ」の動画は変化の大きいところが全区間にわたって比較的均一に分布しているのに対し,「音楽」の動画は特定区間に集中していることもわかる．さらに,「エンタメ」と「生活」の動画の多くは動画投稿者が撮影・編集したものであるため，他のジャンルよりも変化が少ないと考える．

連続する 2 枚のフレーム間で S_t が大きければカット点と判断するが，エフェクトなどの影響により，実際にはカット点でない箇所も検出してしまうことがある．そこで，本稿では以降のプロセスで必要とするよりも多めにカット点を検出し，誤検出区間の統合などの後処理を行う．動画の長さに基づき，抽出する区間数 N を定め，カット点数を $2(N\text{-}1)$と設定する．

後処理として，誤検出を除去するため短い区間を合併する．隣接する 2 つのカット点に挟まれた区間に対し，その長さが 10 秒あるいは元動画の 4%の長さのいずれか長い方よりも短い場合に，前後に隣接する区間の短い方と合併する．

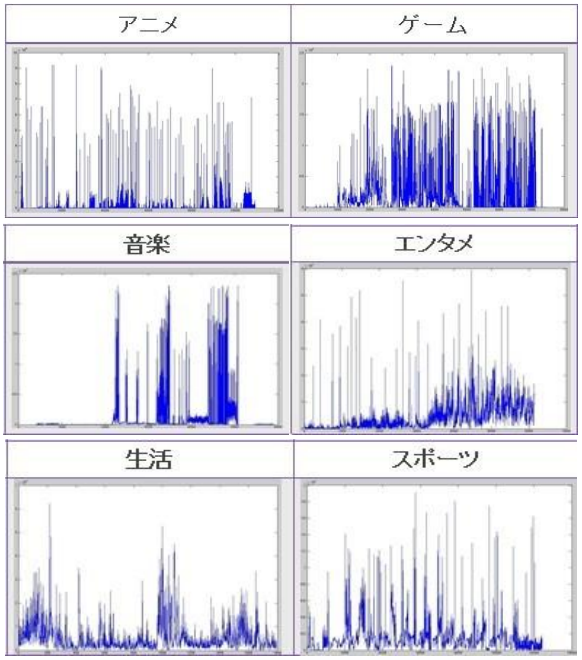


図 2 フレーム間画素濃淡値の差

3.2 区間選択・短縮

区間の選択では，ユーザの関心ある部分を選択することを目的として, 3.1 節で得られた各区間にコメント頻度などを用いて重要度を付与し，重要度の高い区間を選択する．

具体的には，区間 k の 1 秒あたりのコメント数を C_k/T_k と設定する. ここで, C_k を区間 k のコメント数, T_k を区間 k の長さとする．また，区間 k に含まれるコメントの割合を C_k/C と設定する．C は対象動画のコメント数である．以上に基づき，重要度 $IF(k)$を式(2)より求め，この値が上位の 4 区間を選択する．

$$IF(k) = \frac{C_k}{T_k} \times \frac{C_k}{C} \qquad (2)$$

区間短縮では，コメント数，およびキーワードを含むコメント数を利用することで区間を10秒に短縮

する．ここで，キーワードとは，コメント中に頻出する文字列とする．本稿では対象ジャンルを限定しない要約生成手法を目指すため，キーワードを辞書として用意しておく方法は適さないと考える．そこで，文字Nグラムを採用してキーワードを抽出する．Nは3とする．Nグラムを採用することにより，笑いを表現する「www」や「あああ」などの感嘆表現をキーワードとして扱うことが可能となる．

長さが10秒以上の区間kから部分空間を抽出する処理について，図3を用いて説明する．部分空間の長さは10秒に固定し，区間kの先頭から5秒ずつずらしながら，部分空間を抽出する．従って，隣接する部分空間iとi+1は5秒のオーバラップを持つ．

部分区間 i のコメント数を C_i，感嘆表現とキーワードを含むコメント数を K_i とする．$C_i \times K_i$ が最大となる部分空間を抽出し，3.3 節での GIF アニメーション生成対象とする．

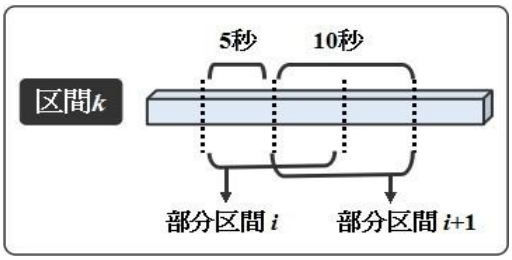


図 3 部分区間の定義

3.3 区間の並列提示

並列提示手法では，動画の要約結果を一目で把握可能とするため, 3.2 節で抽出した 4 つの区間を時間順に並べて表示する．各区間は 10 秒の GIF アニメーションとして提示され，同時に再生される．各動画は 10 秒と短いため，繰り返し再生する．

4 並列提示に関する予備実験

4.1 実験概要

3.3 節で述べた複数区間の並列提示手法は, 短時間で効率的に要約を確認できることが期待される一方, ユーザにとって認知的負荷が高い可能性もある．並列提示する区間数が増えるほど効率は高くなるが，負荷は高くなる事が想定される．そこで，単一の動画として生成した動画要約を提示した場合と，並列提示手法の比較を行う．

本実験では，20 代の工学系学部生・大学院生に，動画要約を見ながら対象動画に関する質問に解答してもらった．ニコニコ動画の「アニメ」から 6 動画を取得して，それぞれ単一の GIF アニメーション（1 分割），4 つの GIF アニメーションの並列提示（4 分割），6 つの GIF アニメーションの並列提示（6 分割）の 3 種類の要約を生成した．6 種類の動画に対し，作成した問題を図 4 に示す．出題数は 4 問であり，質問 1 と 4 は，内容の類似した動画や同じアニメーションの異なる話を組として，当てはまる動画を選択してもらう形式となっている．質問は，各動画の特徴または内容に関するものであり，実験協力者には各質問に対してそれぞれの動画要約を見てから解答してもらった．

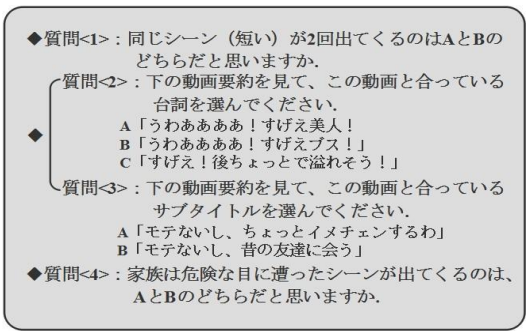


図 4 実験に用いた質問

本実験では，同じ動画を複数回見ないように配慮し, 実験協力者を 3 人ずつの 6 グループに分類した．そして，各実験協力者が 6 動画と 3 種類の要約を全部見るように，各要約と質問を組み合わせた．表 1 は 3 種類の要約提示方法と質問の組み合わせを示しており，これに基づき各実験協力者グループが解答した質問パターンを表 2 に示す．

表 1 要約提示方法と質問の組み合わせ

	動画①②	動画③④	動画⑤⑥
	質問1	質問2、3	質問4
1分割	N1	N2	N3
4分割	N4	N5	N6
6分割	N7	N8	N9

表 2 実験協力者グループと質問パターンの組み合わせ

グループ	実験協力者			1分割	4分割	6分割
A	A1	A2	A3	N1	N5	N9
B	B1	B2	B3	N4	N8	N3
C	C1	C2	C3	N7	N2	N6
D	D1	D2	D3	N1	N8	N6
E	E1	E2	E3	N4	N2	N9
F	F1	F2	F3	N7	N5	N3

図 5 に，グループ B の実験協力者に対する質問と要約の例を示す．実験協力者が各質問に対して「Play」ボタンをクリックすると，解答時間の記録が始まり，解答の選択肢の一つをクリックすると，解答時間の記録が終わる．各 GIF アニメーションの再生時間は 10 秒だが, 最後まで行くと先頭に戻って繰り返し再生される．

図 5 グループ B の実験ページの例

4.2 実験結果

表 3，4，5 に，各実験協力者の解答の正誤および解答時間について示す．ここで，実験協力者 F3 は他の協力者と比較して極端に長い解答時間となっている．また，実験協力者 B2，C1，C3 と F1 とは，対象動画について実験前に知っていた．

表 3 質問 1 の解答時間と正誤

質問 1			
	実験協力者	解答時間(秒)	正誤
1分割	A1	54	O
	D1	223	O
	A2	138	O
	D2	82	O
	A3	282	O
	D3	106	X
4分割	B1	51	O
	E1	82	O
	B2	73	O
	E2	31	O
	B3	26	O
	E3	35	O
6分割	C1	77	O
	F1	100	O
	C2	70	O
	F2	54	O
	C3	28	X
	F3	471	O

実験結果より，質問1と2についてはどの要約提示方法においてもほとんどの実験協力者が正解しているのに対して，質問3と4では不正解の人が多くなっていることがわかる．これらの結果は，要約提示方法よりも質問の難易度に依存すると考える．質問3は動画タイトルに関する質問であるため，動画タイトルの指す内容と視聴者の受け取り方が異なる可能性があると考える. 質問4は特定のシーンの解釈に関する問題であり，実験協力者ごとに解釈が異なったため正解者が少なくなったと考える．

表 4 質問 2 と 3 の解答時間と正誤

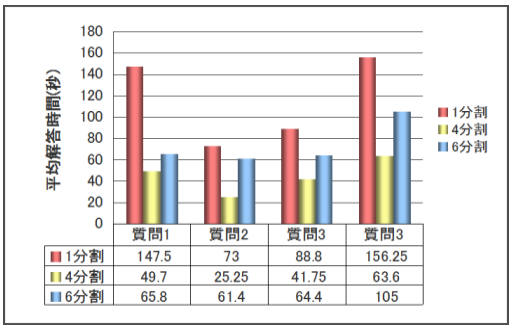
		質問 2		質問 3	
	実験協力者	解答時間(秒)	正誤	解答時間(秒)	正誤
1分割	C1	66	O	67	O
	E1	101	O	128	O
	C2	70	O	71	O
	E2	66	O	80	O
	C3	73	O	83	O
	E3	55	O	76	X
4分割	A1	26	O	31	X
	F1	56	O	81	O
	A2	37	O	75	X
	F2	244	O	31	O
	A3	14	O	30	X
	F3	168	O	107	X
6分割	B1	64	O	78	O
	D1	51	O	60	X
	B2	32	O	26	O
	D2	55	O	96	X
	B3	27	O	35	X
	D3	110	O	53	X

表 5 質問 4 の解答時間と正誤

質問 4			
	実験協力者	解答時間(秒)	正誤
1分割	B1	348	X
	F1	121	O
	B2	31	O
	F2	34	X
	B3	122	X
	F3	186	O
4分割	C1	20	O
	D1	52	X
	C2	68	O
	D2	63	O
	C3	41	X
	D3	94	O
6分割	A1	23	O
	E1	136	O
	A2	212	O
	E2	64	X
	A3	163	O
	E3	32	X

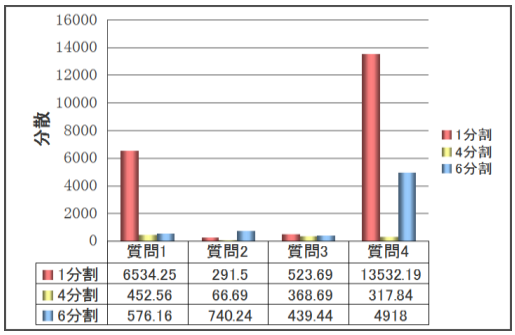
図6，7は，要約提示方法間で解答時間の平均と分散をそれぞれ比較したものである．これらの結果を見ると，質問2と3は解答時間の平均，分散共に小さいのに対し，質問1と4では解答時間の平均・分散共に大きくなっている．これは，質問1と4が比較問題であることが一因と考える．すなわち，基本的には二つの動画を視聴しなくてはならないため，単一の動画を対象とした場合よりも倍程度の時間がかかってしまうと言える．しかし，1分割よりも4，6分割の方が解答時間が減少していることから，複数動画を並列提示する提案手法の有効性が確認できる. 特に4分割が質問によらず解答時間が短く，分散も小さい傾向にあることから，認知的負荷の点でバランスがとれていると考える．更に，質問1のように内容ではなく映像の特徴のみで解答できる場合は, 4分割によ

り，短時間で正確な解答が可能と考える．



	質問1	質問2	質問3	質問3
1分割	147.5	73	88.8	156.25
4分割	49.7	25.25	41.75	63.6
6分割	65.8	61.4	64.4	105

図 6 平均解答時間



	質問1	質問2	質問3	質問4
1分割	6534.25	291.5	523.69	13532.19
4分割	452.56	66.69	368.69	317.84
6分割	576.16	740.24	439.44	4918

図 7 分散

実験後，3 種類の要約提示方法の見やすさについて実験協力者に感想を聞いたところ，質問 1 と 4 のように内容の近い動画の比較問題では，分割されていると集中して見ることが困難だと思う実験協力者が多数いた．また，6 分割が見やすいと思う実験協力者は一人もいなかった．これらの原因と考えられることとして，複数の動画が提示された場合，それらの時間的順序がわかりにくいという意見があった．そのため，時間的な順序がわかりやすいように提示する工夫が必要と考える．また，実験協力者の多くは 1 分割が最も見やすいと回答しているが，これは慣れの問題も大きいと考える．実験結果では，複数分割でも正解率は下がらず，解答時間は短縮されることが示されていることから，提示方法を更に改善すれば，多数の動画から見たい動画を選択する場合などに役立つと考える．

5 おわりに

本稿では，見るべき動画の効率的発見を支援する指示的動画要約を生成するため，ニコニコ動画を対象として映像コンテンツとコメントの両方を考慮した重要区間抽出手法を提案した．また，抽出した区間を並列提示する手法も提案した．

提案する重要区間抽出手法では，画素値など映像内の情報だけではなく，コメントなど映像以外の情報も同時に分析することで，動画作成側と視聴者側の両方の視点を考慮している．また，並列提示手法の利用可能性について予備実験を行った結果，4 分割程度であれば正解率を維持しつつ解答時間を短縮可能であることを示した．

今後は，提案する要約生成手法の評価実験を行う予定である．また，時間軸の追加などにより並列提示手法を改善することで，より見やすい提示手法を検討する予定である．

参考文献

[1] 自主企画アンケート結果:動画共有サイト（第２回）, http://www.myvoice.co.jp/biz/surveys/15713/index.html, (2013 年 9 月 8 日現在)

[2] 奥村学, 難波英嗣: テキスト自動要約 (知の科学), オーム社, p.6, 2005.

[3] 三浦宏一, 浜田玲子, 井手一郎: 料理映像の特徴を利用した要約手法の検討, 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解, Vol. 102, No. 155, pp. 15-20, 2002.

[4] 永橋功丞, 宮岡伸一郎: 映像中の情報を用いたニュース映像要約手法の研究, 第 70 回全国大会講演論文集, pp. 2-357-2-358, 2008.

[5] 岩元浩太, 山田昭雄: 多次元特徴空間解析に基づく映像のカット検出手法, 第 4 回情報科学技術フォーラム, Vol. 4, No. 3, pp. 65-66, 2005.

[6] 青木秀憲, 宮下芳明: ニコニコ動画における映像要約とサビ検出の試み, 情報処理学会研究報告, 2008-HCI-128/2008-MUS-75,Vol.2008,No.50, pp.37-42, 2008.

[7] 多胡厚津史, 中川博之, 田原康之, 大須賀昭彦:ニコニコ探検くらぶ: ソーシャルアノテーションとキーワード群に基づく動画要約, 情報処理学会シンポジウム論文集, Vol.2010,No.4,pp.47-50,2010.

[8] 中村貴洋, 青木秀憲, 宮下芳明: マンガ的手法を用いたニコニコ動画ナビゲーション, 情報処理学会研究報告. HCI, ヒューマンコンピュータインタラクション研究会報告, pp. 103-110, 2008.

[9] 笠松沙紀, 伊藤貴之: 動画像データの要約可視化インタフェースの一手法, 第 1 回データ工学と情報マネジメントに関するフォーラム(DEIM), E9-5, 2009.

[10] 小関悠, 角康之: 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, 情報処理学会シンポジウム論文集, Vol. 2005, No. 4, pp. 177-178, 2005.

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5回)
SIG-AM-05-06

LDA と Isomap を用いたカリキュラムの可視化

Visualization of Curriculum on LDA and Isomap

関谷 貴之[1*] 松田 源立[2] 山口 和紀[2]
Takayuki Sekiya[1] Yoshitatsu Matsuda[2] Kazunori Yamaguchi[2]

[1] 東京大学情報基盤センター
[1] Information Technology Center, the University of Tokyo
[2] 東京大学総合文化研究科
[2] Graduate School of Arts and Sciences, the University of Tokyo

Abstract: A curriculum is crucial to maintaining the integrity of various courses of universities, and an effort to improve a curriculum may include an analysis of the curriculum by comparing the curriculum with previous curricula or curricula of other universities. However, reading syllabi of each curriculum and comparing those curricula is too labor-intensive. We have been develping a method to project syllabi on a map by latent Dirichlet allocation (LDA) and Isomap for the comparison of curricula embodied by the syllabi. In this researh, we applied our method to analyzing curricula offered by information science-related departments. The result of analysis of the maps showed that our curriculum maps reflected characteristics of each department.

1 はじめに

情報科学の理論から計算機等のハードウェア，更には各種アプリケーションの応用分野など，大学では「情報系」の学部や学科は多い．しかし，学部や学科の名称は様々で，実際に学科で扱われるカリキュラムの内容も多岐にわたるため，ある学科がどのような教育や研究を行っているのかを知るのは容易でない．

一方我々は，カリキュラムの違いを概観し，これを比較する方法が必要だと考えており，シラバスとして公開されている文書を用いてカリキュラムのマップを作成して可視化する手法を研究している [14].

そこで我々は，情報系の教育や研究を行う学科のカリキュラムを可視化して，その違いを分析している [15]. まず，国内大学について，情報系の教育や研究を行っていると推測される学科を選び，各学科のシラバスを Web サイト等から取得した．次に取得したシラバスを用いて，カリキュラムマップを作成した．マップの作成に当っては，情報処理学会によるカリキュラム標準として，コンピュータ科学領域の J07-CS[18] を用いた．その結果，工学の分野で「コンピュータを利用する力を重視する」学科や，理学の分野で「計算機・情報そのものを研究対象とする」学科などの特徴がマップに反映されることが分かった．

*連絡先：東京大学情報基盤センター
〒 153-8902 東京都目黒区駒場 3-8-1 情報教育棟
E-mail: sekiya@ecc.u-tokyo.ac.jp

本発表では，カリキュラムの可視化方法について述べると共に，情報系学科のカリキュラムの分析結果の一部について報告する．以下，第 2 節では，本研究に関連のあるシラバスを用いた研究やカリキュラム分析の研究を挙げる．第 3 節ではカリキュラムの可視化手法について述べる．第 4 節では，取得した 5 つの学科のカリキュラムの比較結果について述べ，第 5 節で本研究をまとめる．

2 関連研究

高等教育機関において，個々の大学単位でのシラバスのデータベース化やライブラリ化は既に広く実施されているが，単一の大学だけでなく組織を跨がる形での試みも行われている．例えば，Joorabchi らはアイルランド国内の高等教育機関におけるシラバスライブラリの構築について報告している [10]．教員からシステムにアップロードされたシラバス文書を，その内容に応じて分類している．Chatvichienchai は長崎県下の大学における単位互換のために，関連する複数の大学で開講されている講義のシラバスを XML 化した上でこれを管理するリポジトリ構築の試みについて報告している [4]．このような複数組織のシラバスのデータベース化を進めた後には，組織毎の特徴を様々な観点で分析する手法が必要になると考えられる．

カリキュラムを構成する講義間の関係を分析する方法としては，履修者の成績データに対してマイニング

技術を適用する [17] など様々な手法が考えられるが，講義の内容を要約した情報を含むシラバスを用いることで，講義間の関係をより適切に反映できる．例えば美馬らは，自動的に専門用語を認識してクラスタリングする技術を用いた MIMA search[11] やシラバス検索のための分類システム [12] を開発している．井田らは対話的にカリキュラムを分析するツールを開発している [7, 6]．ここで，複数の組織のカリキュラムを相互に比較する上では，何らかの共通の基準を設定した上で比較することが望ましいが，いずれのツールにも，我々の知る限りはそのような試みは行われていない．一方，本研究ではカリキュラム標準を基準として，複数のカリキュラムを比較可能にする点に特徴がある．

石畑らは J07-CS の知識体系の各ユニットを用いて，理工系の情報学科の授業内容を分析している [8]．分析に当っては，多数の学科のシラバスを取り寄せた上で，各授業がカバーするユニットを複数の専門家が共同で調査している．テキスト処理によって自動的にマップを作る我々の研究は，このような専門家による調査の一助にもなると考えられる．

カリキュラムに限定せず，大量の文書の集合全体を可視化する方法としては，生成モデルに基づいて文書から直接二次元平面にマップするもの [9] がある．しかし，異なる文書集合を同一マップ内で比較するのは困難である．本研究では同一のマップに複数の文書集合を射影して比較できる点に特徴がある．

3 可視化手法

3.1 可視化の要件

本研究では，カリキュラムをシラバスの集合として捉え，そのシラバスを平面上に記号として配置したものをカリキュラムのマップと呼ぶ．そして，カリキュラムの特徴は，このマップ上の記号の分布として視覚的に把握できるものと仮定する．更に，異なる大学のカリキュラムや過去のカリキュラムとの比較を行うことを考えると，複数のカリキュラムを一つのマップの上に載せられることが望ましい．そこで，本研究では，以下の 3 つの条件を満たす可視化を提案している．

(A) 与えられたカリキュラムを基準として可視化が行えること

(B) シラバスの関連度がマップ上の距離として反映されること

(C) 異なるカリキュラムを同一の基準でマップできること

3.2 LDA

3.1 節で述べた条件 (A) を満たすために，与えられたカリキュラム標準の個々の専門的な領域に合致したトピックを LDA を用いて抽出し，そのトピックをカリキュラム分析の基準として用いることにした．LDA とは，確率的文書モデルに基づいて文書の集合から個々の文書を特徴づける潜在的なトピックを抽出する手法である．カリキュラム標準においては,「領域」が予め定義されているが，本研究では領域を LDA の確率的モデルとして再抽出する．これにより，カリキュラム標準に含まれない任意のシラバスが，カリキュラム標準の各領域にどの程度関連するかを計算できるので，各シラバスを共通のトピック空間上に配置することが可能となる．LDA の確率モデルの概要を以下に説明する．詳細は Blei らの論文 [3] を参照して欲しい．

1. 前提として，LDA では文書を確率的に選択された単語の集合 (bag of words) と捉える．そこで，文書 $\boldsymbol{w} = (w_m)(m = 1, \cdots, M)$ を，N 種類の語彙 $v_1, \cdots, v_N$ が M 個の単語 $w_1, \cdots, w_M$ として現れる多重集合 (同一文書中に単語として同一の語彙が複数回登場することを許す集合) として扱う．すなわち，各 $w_m \in \{v_1, \cdots, v_N\}$ である．

2. 最初に，Dirichlet 分布 Dirichlet($\boldsymbol{\alpha}$) に従って，文書を特徴づけるトピックの確率ベクトル $\boldsymbol{\theta} = (\theta_t)(t = 1, \cdots, T)$ が決定されると仮定する．ここで，トピックは T 種類，$\boldsymbol{\alpha}$ は T 次元 Dirichlet パラメータである．$\sum \theta_t = 1$ である．

3. 次に，文書 $\boldsymbol{w}$ に含まれる m 番目の単語 w_m のトピック $z_m (m = 1, \cdots, M)$ が，$\boldsymbol{\theta}$ の多項分布に基づいて決定される．$z_m \in \{1, \cdots, T\}$ である．

4. 最後に単語 w_m が，条件付き多項分布 $p(w_m | z_m, \boldsymbol{\beta})$ に従って語彙の中から選択される．ここで，$\boldsymbol{\beta} = (\beta_{nt})$ は $\sum_n \beta_{nt} = 1$ を満たす多項分布パラメータであり，トピック t と語彙 v_n の関連の強さを示す．

文書集合において，個々の文書に含まれる単語の頻度情報を LDA で処理すると，各文書はそのトピックの確率ベクトル $\boldsymbol{\theta}$ に基づいてトピック空間内で特徴づけられる．つまり，カリキュラムをシラバスの集合とみなして，LDA で処理する事で，個々のシラバスを $\boldsymbol{\theta}$ を用いて T 次元上で特徴づけることができる．

LDA においては，文書 $\boldsymbol{w}$ の集合のみが既知であり，他のパラメータは何らかの基準で決めるか，推定して決める必要がある．推定及び設定手法としては Gibbs サンプリング法 [5] などがあるが，本研究では広く使われている変分ベイズ法 [3] を用いた．

LDA におけるパラメータは，大きく分けて，生成モデルを規定する基準となるパラメータ ($\boldsymbol{\alpha}$，$\boldsymbol{\beta}$) と，ある与えられた文書がどのトピックに関係しているかを示すパラメータ ($\boldsymbol{\theta}$，z_m) の二種類がある．基準となる文

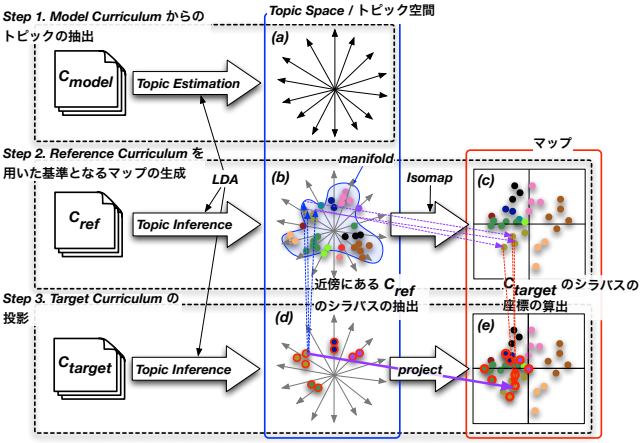


図 1: カリキュラムの可視化の手順の概念図

書集合を元に $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ を推定し，それ以外の文書に関しては $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ を固定して $(\boldsymbol{\theta}, z_m)$ を推定することで，与えられた全ての文書を基準パラメータによって特徴づけすることができる．ただし，$(\boldsymbol{\theta}, z_m)$ の推定は膨大な積分計算を必要とするため，一般に困難である．変分ベイズ法では，この推定を近似的に解決している．本稿では，$\boldsymbol{\theta}$ に関してのみ説明する．

$\boldsymbol{\theta}$ については，$\boldsymbol{\theta}$ の分布 $P(\boldsymbol{\theta})$ の推定問題を，$P(\boldsymbol{\theta}) \simeq \mathrm{Dirichlet}(\boldsymbol{\theta}|\boldsymbol{\gamma})$ で近似し，近似の最も良くなる T 次元 Dirichlet パラメータ $\boldsymbol{\gamma} = (\gamma_t)$ を推定する問題に帰着させる．$\boldsymbol{\gamma}$ は Dirichlet パラメータであり，$\boldsymbol{\theta}$ そのものの値ではないが，Dirichlet 分布の性質により，$\boldsymbol{\theta}$ の期待値 $E(\boldsymbol{\theta})$ は，合計を正規化した $\bar{\boldsymbol{\gamma}} = \boldsymbol{\gamma} / \sum \gamma_t$ で与えられる．そこで，本研究では，この $\bar{\boldsymbol{\gamma}}$ を，文書のトピック空間での位置と解釈する．ベクトルである $\bar{\boldsymbol{\gamma}}$ のトピック t に対応する座標値が，その文書とトピック t との関連の強さを表わす．

本研究では，$\boldsymbol{\alpha} = (\alpha_t)$ の全ての要素の値は一定値 α であると仮定する．これは，LDA で抽出されるトピックの生成率，つまりトピックに対応するカリキュラム標準の個々の専門的な領域の存在に偏りが無いと仮定することに対応する．J07-CS における与えられた各トピックは代表的なものが偏り無く提案されているので，この仮定は妥当である．

3.3 可視化の手順

本節では，LDA と Isomap を用いたカリキュラムの可視化の手順を説明する．分析には，次の 3 種類のカリキュラムを用いる．ここでは，カリキュラムはシラバス $\boldsymbol{w}_i$ の集合で表されている．

Target Curriculum, $C_{\mathbf{target}} = \{\boldsymbol{w}_i^{\mathbf{target}}\}$: 本手法を用いて分析するカリキュラム．

Model Curriculum, $C_{\mathbf{model}} = \{\boldsymbol{w}_i^{\mathbf{model}}\}$: C_{target} が対象とする学問分野を広くカバーする標準的，模範的なカリキュラム．その学問分野のトピックの抽出に用いる．

Reference Curriculum, $C_{\mathbf{ref}} = \{\boldsymbol{w}_i^{\mathbf{ref}}\}$: C_{target} の分析にあたって比較基準となるカリキュラム．

図 1 の可視化の手順の概念図に沿って説明する．

Step 1: Model Curriculum からのトピックの抽出

標準カリキュラム C_{model} を LDA で分析する (*Topic Estimation*) ことで，分析対象の学問分野におけるトピックの生成モデルのパラメータ $\boldsymbol{\beta}_{\mathrm{model}}$ を推定する (本研究では $\boldsymbol{\alpha}$ は既知としている)．以降の Step では，この $\boldsymbol{\beta}_{\mathrm{model}}$ を固定して，分析対象となるカリキュラムの標準的トピック空間として用いる (図 1(a))．任意のシラバス $\boldsymbol{w}_i$ に同一のトピック生成モデル $\boldsymbol{\beta}_{\mathrm{model}}$ を用いることが出来るので，3.1 節の条件 (A) を満たす．

Step 2: Reference Curriculum を用いた基準となるマップの生成

まず $\boldsymbol{\beta}_{\mathrm{model}}$ を利用して，C_{ref} の各シラバスの $E(\boldsymbol{\theta}) = \bar{\boldsymbol{\gamma}}(\boldsymbol{w}_i^{\mathrm{ref}})$ を LDA で推定する (*Topic Inference*)．これにより，C_{ref} の全シラバスを，基準となるトピック空間上に配置する (図 1(b))．しかし，トピック空間は T 次元空間であり，このままでは，カリキュラムの全体構造を可視化できない．そこで，本研究では，高次元空間内の多様体構造を保持したまま，これを広げるようにして次元を縮退する手法 Isomap[16] を用いる．

$$\boldsymbol{x}(\boldsymbol{w}_i^{\mathrm{ref}}) = \Pi_{\mathrm{Isomap}}(\bar{\boldsymbol{\gamma}}(\boldsymbol{w}_i^{\mathrm{ref}})) \tag{1}$$

但し $\Pi_{\mathrm{Isomap}} : \boldsymbol{R}^T \rightarrow \boldsymbol{R}^2$ は Isomap による写像変換．

ここで，$\boldsymbol{x}(\boldsymbol{w}_i^{\mathrm{ref}})$ は各シラバスの二次元平面上での位置を示す．この結果，3.1 節の条件 (B) を満たすような C_{ref} を二次元平面上に展開したマップが得られる (図 1(c))．Isomap の近傍数 k_{iso} については後述する．

Step 3: Target Curriculum のマップへの射影

まず Step 2 と同様に $\boldsymbol{\beta}_{\mathrm{model}}$ に基づいて，C_{target} の各シラバスの $\bar{\boldsymbol{\gamma}}(\boldsymbol{w}_i^{\mathrm{target}})$ を LDA で推定する (*Topic Inference*)．これにより基準トピック空間に C_{target} が配置される (図 1(d))．

次に，各シラバス $\boldsymbol{w}_i^{\mathrm{target}}$ を，Step 2 で求めたマップに，以下の式を用いて射影する．

$$\boldsymbol{x}(\boldsymbol{w}_i^{\mathrm{target}}) = \sum_{k=1}^{k_{\mathrm{neigh}}} \eta_k \boldsymbol{x}(\boldsymbol{w}_k^{\text{ref-neigh}}). \tag{2}$$

ここで，$\boldsymbol{w}_k^{\text{ref-neigh}} \in C_{\mathrm{ref}}$ は，トピック空間上で $\bar{\boldsymbol{\gamma}}(\boldsymbol{w}_i^{\mathrm{target}})$ の近傍となる k_{neigh} 個の比較基準シラバスである．η_k は $\sum \eta_k = 1$ を満たす加算重みであり，この制約下で $\sum \eta_k \bar{\boldsymbol{\gamma}}(\boldsymbol{w}_k^{\text{ref-neigh}})$ と $\bar{\boldsymbol{\gamma}}(\boldsymbol{w}_i^{\mathrm{target}})$ の距離を最小化することで計算する．この計算は，多様体次元縮退法の一種である LLE[13] で行われるものと同一である．この結

果，C_{target} の全てのシラバスが，Step 2 で求めた C_{ref} のマップと同一の二次元平面上に射影される (図 1(e))．異なるカリキュラムを同一のマップ上で比較できるので，3.1 節の条件 (C) を満たす．

4 実験:カリキュラムの可視化

4.1 パラメータの設定

LDA を用いた文書集合のモデル化においては，α の値を 1 より小さくすれば，各文書は少数のトピックとの関係が強くなり，1 より大きくすれば多数のトピックとの関係が同程度に強くなる傾向がある．その結果，本研究のカリキュラムのマップにおいては，α が小さいと同じトピックと関係が強いシラバスがトピック毎に固まって配置され，α が大きいと関係の強いトピックに拘わらずシラバスが一箇所に固まって配置される傾向がある．いずれにしても，シラバス同士の違いや関係を視覚的には把握し辛くなる．このような観点から $\alpha = 1$ とすることが適切であると考えられる．

また，3.2 節で述べたように，本研究ではカリキュラム標準における領域を LDA のトピックとして抽出して，これを基準としてトピック空間を形成し，カリキュラムマップを生成することから，トピックの数 T は領域の数と同じ値にすることが適切であると考えられる．

このようなパラメータの設定が妥当か否かを検証するために，Dirichlet パラメータ α とトピック数 T を決める予備実験を行った．本研究のカリキュラムの分析方法においては，LDA で抽出したトピックと文書との関連の強さが，Isomap を用いて得られるマップ上の文書の座標として適切に表現され，互いに関係のある文書がマップ上で近くに配置されるかが重要である．そこでこの実験では，予め分類された文書集合として UCI Machine Learning Repository[1] [1] が提供する Twenty Newsgroups Data Set(文書数:2000 個, 分類数:20 個) を用いた．文書群によって確率モデルが異なるため，本来はカリキュラム標準である J07-CS 自体をデータセットとして用いるべきである．しかしながら，J07-CS の文書集合のサイズ約 140 個程度では，パラメータの最適値を求めるには十分ではないと考えられる．そこで，広く用いられている UCI Machine Learning Repository を用いて，あくまで目安ではあるが，安定的に良い結果を返すと期待される α と T の値を求めている．

3.3 節で述べた LDA と Isomap を使った可視化手法を，パラメータの値を変更しながら適用して，同じ分類の文書 (ネットニュースの記事) がより近くに配置されれば，その時の値が適切だと判断する．判断指標として，Leave-one-out cross validation[2] に基づく k 近傍法による分類精度を用いた．全てのパラメータ値について，同じ分析処理を 10 回繰り返し，得られた分類精度の平均値で評価した．$\alpha = \{0.1, 1.0, 10\}, T = \{10, 20, 30\}$ の時の精度のグラフを図 2 に示す．この予備実験の結果から，トピック数 T の値を元データの分類数と同じ 20 かつ Dirichlet パラメータ α を 1 とした時の精度が高いことが分かった．そこで，当初想定した T = カリキュラム標準の領域数及び $\alpha = 1$ を採用することとした．

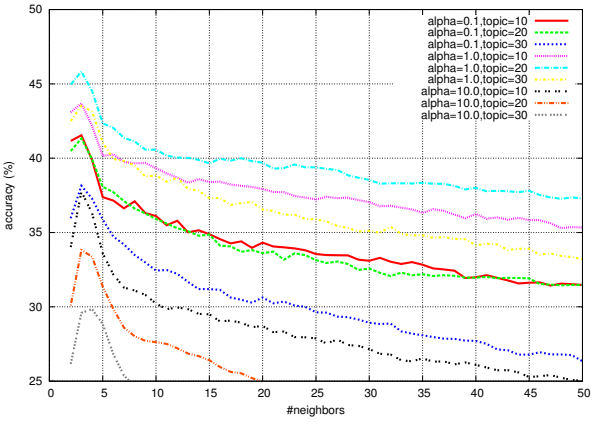


図 2: 異なる k の値に基づく k 近傍の一致率 (パラメータ $\alpha = 0.1, 1, 10$，トピック数 $T = 10, 20, 30$)

なお，3.3 節で説明した分析方法 Step 2 で用いる Isomap では，トピック空間内での近傍数 k_{iso} によって平面上のマップが変化する可能性がある．しかし，本研究で対象としたカリキュラムを分析した限りでは，ある程度小さい値であれば，その値の違いによってマップ上のシラバスの分布が大きく変わることは無かったため，Reference Curriculum の全てのシラバスを接続可能な (言い換えれば，全てのシラバス間で有限な距離が定義できる) 最も小さい値を用いることとした．4.2 節で述べる Reference Curriculum では $k_{iso} = 14$ となった．また，比較基準シラバス数 k_{neigh} については，接続可能という条件は必要ないため，比較的小さい値をとることとし，$k_{neigh} = 3$ とした．

4.2 Model 及び Reference Curriculum

J07-CS を Model Curriculum として，情報系学科のカリキュラムを分析するために，Body of Knowledge (BOK) =カリキュラム，領域=トピック，ユニット=シラバスと解釈した．J07-CS の領域は 15 個，ユニットの総数は 138 個で，個々の領域のユニットは 4〜14 である．ユニットの説明は比較的狭い範囲の事柄を列挙したもので，同じ領域に属する互いに関係のあるユニットであっても，共通の単語が無い或いは非常に少なく，LDA で抽出したトピックとの関連の強さが全く異なる場合も有り得る．そこで 3.2 節で述べたように，カリキュラム標準の各領域に合致したトピックを抽出するために，ユニットの説明にそのユニットが属する領域

[1] http://archive.ics.uci.edu/ml/
[2] 詳しくは，機械学習の教科書 [2] などを参考にして欲しい．

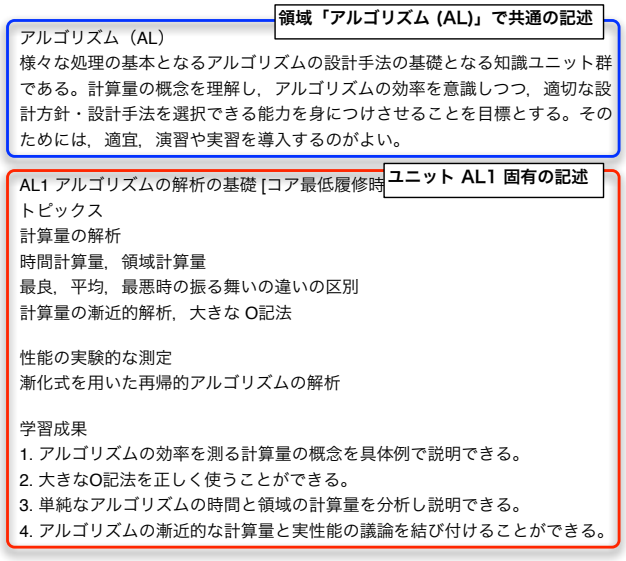


図 3: カリキュラム標準 J07-CS のユニットの記述に対する前処理

の記述を加える処理をして，領域を反映したトピックが抽出されやすいようにした．例えば，領域「アルゴリズム (AL)」のユニット「AL1 アルゴリズムの解析の基礎」を処理する際には，領域共通の記述 (図 3 の上部の青い四角に囲まれた部分) とユニット固有の記述 (図 3 の下部の赤い四角に囲まれた部分) を合わせたものを，シラバス相当の文書とした．その結果，同じ領域に属するユニットは，LDA の同じトピックとの関連が強くなり，LDA のトピックと J07-CS の領域とをほぼ一致させることができた．

図 1 に示すように，本可視化手法において Reference Curriculum はマップ全体の形状や配置の元となることから，分析対象の範囲を広くカバーするカリキュラムが必要である．ここでは，Model Curriculum にも用いた J07-CS の BOK と，情報処理学会が一般情報教育用の教科書として作成した「情報とコンピューティング」[3] と「情報と社会」[4] を用いた．前者の J07-CS BOK については，Model Curriculum と同様にユニットをシラバス相当の文書としたが，領域の記述をユニットに加える前処理は行わない．後者については，各教科書からテキストを抽出して，図 4 に示すように，「1.1 符号化の意義」「1.2 符号化の原理と自然数の符号化」のような節の単位で分け，全部で 69 個の節をそれぞれ独立したシラバス文書として扱った．その結果 Reference Curriculum のシラバスの数は合計で 207 個となる．ここで得られた Reference Curriculum のマップを図 5-a に示す．

一つのシラバスは様々な内容を含むが，3.2 節で述べ

[3]川合慧監修，河村一樹編著，オーム社，2004 年
[4]川合慧監修，駒谷昇一編著，オーム社，2004 年

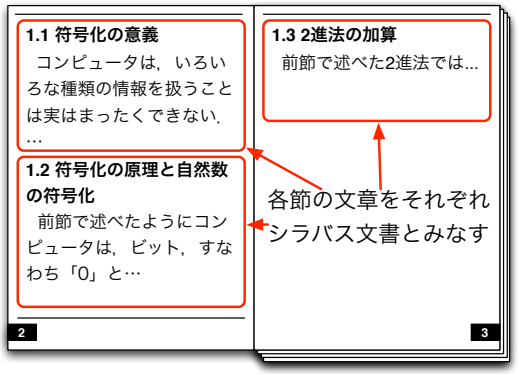


図 4: Reference Curriculum 用シラバス文書の抽出方法の概念図

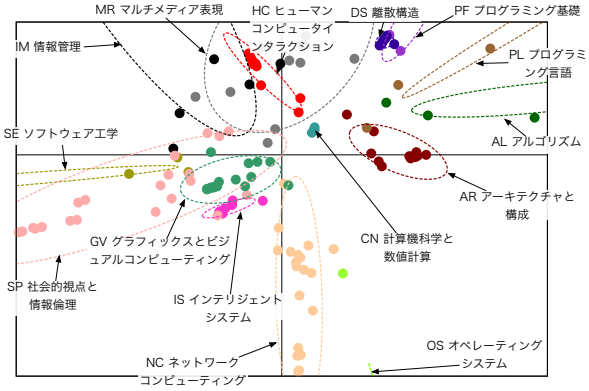
た LDA を用いるとシラバスとトピックとの関係は確率として求められるが，各シラバスに対して最も確率が高いトピックを主たるトピックと呼ぶ．主たるトピックは，そのシラバスの主な内容を示すと考えられる．そこで，表 1 に示すようにトピックの色を決めて，図 5 のマップ上では，それぞれのシラバスの主たるトピックに応じた色の丸記号でシラバスを示している．これによって，Reference Curriculum のマップ上で同じ色の丸記号が固まって配置される部分は，表 1 に示す J07-CS の領域に対応すると考えられる．この丸記号の分布を把握しやすくするために，丸記号の座標値の分散共分散行列から 2 つの軸を求め，$2\sigma(\sigma^2 = \text{分散})$[5] の楕円で囲んでいる．

マップは 15 次元の多様体を Reference Curriculumn のシラバスの配置に基づいて 2 次元に射影しているため，軸の意味は大局的に一定ではないが，図 5-a においては，左は人や社会に，右は理論に関するものが多く，上は管理，下はアーキテクチャに関するものが多い．例えば「SP 社会的視点と情報倫理」の領域は，社会的な視点の内容から左の方に延び，逆に「AL アルゴリズム」や「PL プログラミング言語」の領域は最も右にある．また，「MR マルチメディア表現」は，ディジタル表現のような理論的な内容から，国際的な文字コードなどと，理論と社会の両方を含むことから，左右に跨がっていると考えられる．

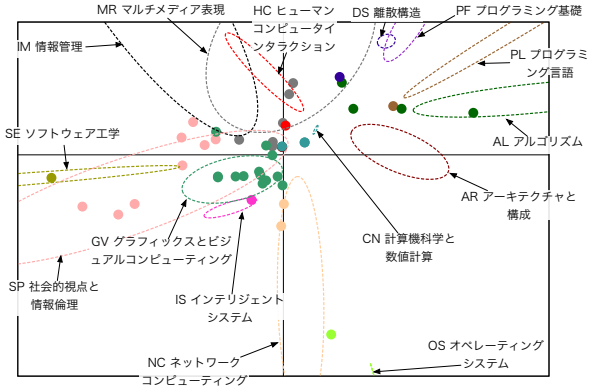
4.3 情報系学科のカリキュラムマップ

本研究では，情報科学や情報工学等の教育・研究を行っている大学の学科を，広く「情報系」の学科として捉え，これらの学科で開講されている講義のシラバスの集合を取得した上で，各学科のカリキュラムを比較分析する．そこで，大学名・学部名・学科名のいずれかに，単語として「情報」「コンピュータ」「ソフト

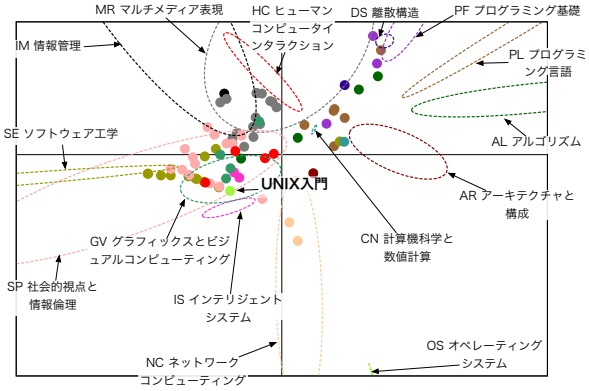
[5]1 次元ガウス分布 (正規分布) では $\pm 2\sigma$ には 95.4%の要素が含まれるが，2 次元ガウス分布では $\pm 2\sigma$ には 86.5%の要素が含まれる．



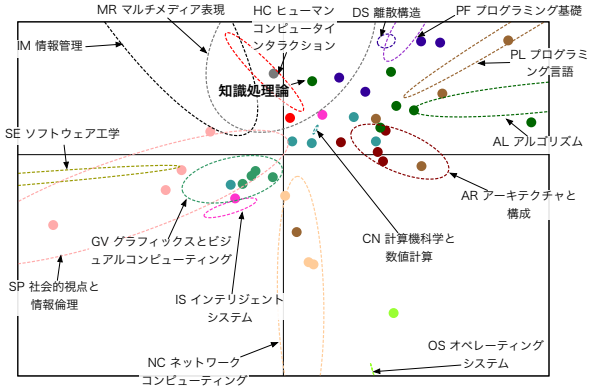
a. J07-CS (Reference Curriculum) のマップ



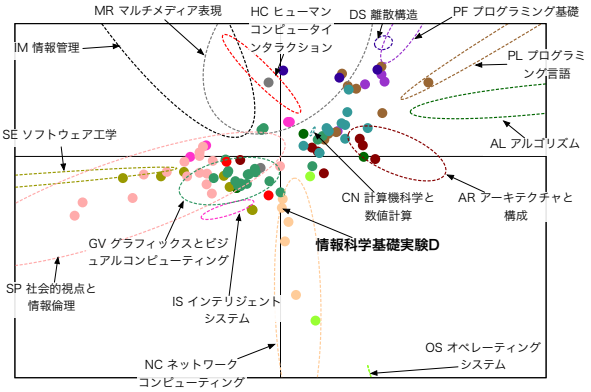
b. A 学科 (国立 工学 その他) のマップ



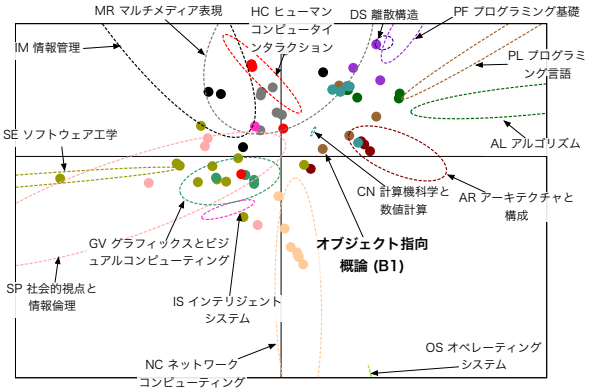
c. B 学科 (私立 工学 その他) のマップ



d. C 学科 (国立 理学 数学関係) のマップ



e. D 学科 (公立 工学 電気通信工学関係) のマップ



f. E 学科 (私立 工学 電気通信工学関係) のマップ

図 5: J07-CS (Reference Curriculum) 及び J07-CS を Reference とする情報系学科のマップ

ウェア」を含む学科を「情報系」の学科とみなし，その中からランダムに学科を 9 個選択した上で，ツール等を用いて比較的容易にシラバスを取得できた 5 つの学科を分析対象とした．

表 2 に 5 つの学科の概要を示す．A–E は以後の説明で学科を指すために独自に設定した記号である．分類は，文部科学省作成の資料に基づいており，対象とした学科が国公立私立のいずれの大学か，またそれぞれの学科の研究及び教育内容がどのように分類されるか (大分類及びその大分類を更に細かく分けた中分類) を示している．学科概要は，各学科の Web サイトで当該学科を説明する文章を抜粋したものである．

表 1: J07-CS の領域とマップ上の色との関係

J07-CS の領域	色
IS インテリジェントシステム	
PF プログラミングの基礎	
AL アルゴリズム	
IM 情報管理	
PL プログラミング言語	
HC ヒューマンコンピュータインタラクション	
SP 社会的視点と情報倫理	
OS オペレーティングシステム	
GV グラフィックスとビジュアル・コンピューティング	
NC ネットワークコンピューティング	
AR アーキテクチャと構成	
SE ソフトウェア工学	
CN 計算科学と数値計算	
DS 離散構造	
MR マルチメディア表現	

表 2: 本研究で分析した情報系の学科 [分類] 及び学科概要

A	[国立 工学 その他] 社会的に新たな価値を創造することのできる想像力と企画力を備えたメディア環境設計の専門家を養成する.
B	[私立 工学 その他] 社会で必要とされるコンピュータを利用する力を重視し，効率的な使い方，便利な使い方を実践的に学ぶ.
C	[国立 理学 数学関係] 情報処理の根本原理の究明・まったく新しい動作原理の計算機の設計・計算機の新しい使い方の提案というように，計算機・情報そのものを研究対象とする.
D	[公立 工学 電気通信工学関係] システム全体の強調と調和を図る幅広い視野と創造的な技術者，研究者を育成.
E	[私立 工学 電気通信工学関係] 情報ネットワーク技術，情報コミュニケーション技術の先駆的分野を開拓する人材を育成.

図 5 の b–f は，図 5-a の Reference Curriculum を用いて，表 2 の各学科のカリキュラムを Target Curriculum として生成したマップである.

最初に全体的な分析を行う．5 つの学科のマップからは，A 学科，C 学科，D 学科，E 学科において，シラバスが全体的に広く分布しており，J07-CS の領域の上では幅広い内容を扱ったカリキュラムであることが推測される．A 学科は，全体の中では「PL プログラミング言語」に関わるシラバスが少なく,「GV グラフィックとビジュアルコンピューティング」や「SP 社会的視点と情報倫理」「MR マルチメディア表現」等に関わる講義が多く，メディア環境設計の専門家を育成するとの特徴を反映している．B 学科は全体的には比較的狭い範囲にシラバスが分布しており，右の理論的な内容や下のアーキテクチャに関する内容は少ない．これは B 学科が「コンピュータの利用」を重視していることが反映されている．C 学科は,「AL アルゴリズム」「PF プログラミング」「AR アーキテクチャと構成」「DS 離散構造」のような，図中では右の方にある領域の内容が多いが,「SE ソフトウェア工学」や「IM 情報管理」のシラバスが少ない.「理学 数学関係」の学科で，情報科学的側面を表している．D 学科はまさに幅広い分野のシラバスを含むし，E 学科も同様である．公立と私立の違いはあるが，いずれも「工学 電気通信工学関係」の学科であることが共通している.

表 3: J07-CS の領域と講義との関係の例

学科	講義名	関連の強い上位 3 つのトピックと関連度 (%)		
B	UNIX 入門	OS: 14.82	SE: 14.80	PL: 13.05
C	知識処理論	AL: 19.77	IM: 17.22	SE: 14.72

次に一部のマップについて個別的な分析を行う．ここで注目するシラバスを表 3 に示す．B 学科の図 5-c には，最も関連が強いトピックは「OS オペレーティングシステム」だが，OS の領域から大きく外れたシラバスがある．このシラバスは「UNIX 入門」で OS，SE，PL と同程度に関わりが強いため，SE などの領域に近付いたことが分かる．シラバスには「ディレクトリの操作」や「プログラミング入門」などの用語が見られ，オペレーティングシステムの技術的な詳細ではなく，OS の使い方を重視していることが分かる．C 学科の図 5-d には，最も関連が強いトピックは「AL アルゴリズム」だが，AL の領域から外れた中央付近に配置されたシラバスがある．このシラバスは「知識処理論」で，AL の次には IM や SE との関わりが強いため，AL と IM の間に配置されたことが分かる．シラバスには「大規模高次元データから知識抽出を行うための統計的データ解析の理論，および，テキスト処理や知識処理の基本的な技法に関する講義を行う」とあり，純粋なアルゴリズム論ではなくデータ解析など IM の内容があることが分かる.

このようなマップによる全体的及び個別的な分析によって，カリキュラムやシラバスの特徴が分かり，学科概要や分類などの特徴はマップに反映されていると言える.

5 終わりに

本発表では，カリキュラムの可視化方法について述べると共に，情報系の分野の教育や研究を行う 5 つの学科を対象として，学科のカリキュラムのマップを作成することでカリキュラムの特徴を分析結果について報告した．それぞれの学科のシラバスは，学科の Web サイト等から取得した．マップの作成に当っては，情報処理学会によるカリキュラム標準 コンピュータ科学領域 J07-CS を用い，その結果，工学の分野で「コンピュータを利用する力を重視する」学科や，理学の分野で「計算機・情報そのものを研究対象とする」学科などの特徴がカリキュラムマップに反映されることが分かった．今後は，マップ上にシラバスの特徴を表す専門用語を示すなど，より分かりやすいマップの生成を検討していきたい．

参考文献

[1] Asuncion, A. and Newman, D.: UCI Machine Learning Repository (2007)

[2] Bishop, C. M.: *Pattern Recognition and Machine Learning*, Springer, New York, NY, USA (2006)

[3] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[4] Chatvichienchai, S.: An XML-Based Syllabus Repository System for Inter-university Credit Exchange Systems, in *International Conference on Computer Technology and Development ICCTD 2009*, Vol. 2, pp. 460–464 (2009)

[5] Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B.: Topics in Semantic Representation, *Psychological Review*, Vol. 114, No. 2, pp. 211–244 (2007)

[6] Ida, M., Nozawa, T., Yoshikane, F., Miyazaki, K., and Kita, H.: Syllabus database and Web service on higher education, in *7th International Conference on Advanced Communication Technology (ICACT2005)*, Vol. 1, pp. 415–418 (2005)

[7] 井田 正明, 野澤 孝之, 芳鐘 冬樹, 宮崎 和光, 喜多 一：シラバスデータベースシステムの構築と専門教育課程の比較分析への応用, 大学評価・学位研究, No. 2, pp. 85–97 (2005)

[8] 石畑 清, 大岩 元, 角田 博保, 清水 謙多郎, 玉井 哲雄, 長崎 等, 中里 秀則, 中谷 多哉子, 疋田 輝雄, 三浦 孝夫, 箕原 辰夫, 和田 耕一, 渡辺 治：理工系情報学科の授業内容分布のシラバスによる調査（中間報告）, 情報教育シンポジウム SSS2010, pp. 139–149 (2010)

[9] 岩田 具治, 山田 武士, 上田 修功：トピックモデルに基づく文書群の可視化, 情報処理学会論文誌, Vol. 50, No. 6, pp. 1649–1659 (2009)

[10] Joorabchi, A. and Mahdi, A. E.: An Automated Syllabus Digital Library System for Higher Education in Ireland, *The Electronic Library*, Vol. 27, No. 4, pp. 640–658 (2009)

[11] Mima, H.: MIMA search: A Structuring Knowledge System towards Innovation for Engineering Education, in *Proceedings of the COLING/ACL on Interactive presentation sessions*, pp. 21–24, Morristown, NJ, USA (2006), Association for Computational Linguistics

[12] Ota, S. and Mima, H.: Machine Learning-based Syllabus Classification toward Automatic Organization of Issue-oriented Interdisciplinary Curricula, *Procedia - Social and Behavioral Sciences*, Vol. 27, pp. 241–247 (2011)

[13] Roweis, S. T. and Saul., L. K.: Nonlinear dimensionality reduction by locally linear embedding, *Science*, Vol. 290, No. 5500, pp. 2323–2326 (2000)

[14] 関谷 貴之, 松田 源立, 山口 和紀：LDA と Isomap を用いた計算機科学関連カリキュラムの分析, 情報処理学会論文誌, Vol. 54, No. 1, pp. 423–434 (2013)

[15] 関谷 貴之, 松田 源立, 山口 和紀：情報系学科のカリキュラムの比較, 情報教育シンポジウム SSS2013, pp. 33–40 (2013)

[16] Tenenbaum, J. B., Silva, de V., and Langford, J. C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science*, Vol. 290, No. 5500, pp. 2319–2323 (2000)

[17] Tissera, W. M. R., Athauda, R., and Fernando, H. C.: Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining, in *International Conference on Information and Automation, ICIA 2006*, pp. 57–62 (2006)

[18] 情報処理学会 コンピュータ科学教育委員会：カリキュラム標準 コンピュータ科学 J07-CS 報告書 (2009 年 1 月 20 日) (2009), `http://www.ipsj.or.jp/12kyoiku/J07/20090407/J07_Report-200902/4/J07-CS_report-20090120.pdf`

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5回)
SIG-AM-05-07

複数文書の相対的特徴可視化による理解支援

A Mulit-Document Relative Characteristics Visualization for Understanding Support

薦田和弘 [1*]　大澤幸生 [1]

Kazuhiro Komoda[1], Yukio Ohsawa[1]

[1] 東京大学工学系研究科
[1]School of Engineering, the University of Tokyo

Abstract: We propose a visualization method showing relative characteristics of multiple documents which have different contexts. We provide users with an interactive graph correlating common words in multiple documents with characteristic words in each document, using co-occurrence and context similarity information. By allowing the users to discover underlying differences between superficially similar documents and to become conscious of them as new viewpoints, we intend to support understanding in terms of cooperative group activities or individual information collections. We examine our method using a document set with context information.

1　はじめに

文書からの知識発見は，学術界や産業界における様々な領域で重要な役割を果たしている．従来，文書内に明示的に出現する事実やその関係性を抽出することで，学術やビジネスにおいて実用上の問題解決が試みられてきた[1]．一方で，文書の著者（情報提供者）の側に立てば，事実関係を示すことに加えて，数ある類似文書との関連性を意識しながら自分の主張の特徴点を把握し，新たな知見を適切に位置づけることによって自分の意図を読者（情報の受け手）に理解してもらう必要がある．例えば，学術論文では，著者独自の問題意識に端を発する数々の試行錯誤をもとに，最新の研究成果に基づく主張が表現される．この主張を他の主張から差別化し，同時に他の主張との関連性を示す特徴点を明確にしてはじめて，学術分野において主張を適切に位置づけ，既存の知見と新たな知見の結合を促すことができる．

文書単独の分析によって得られない相対的な特徴や関連性に対して，2 つ以上の文書集合を組み合わせ，集合間の差に相当する概念を抽出する差異分析が有効である．既存の研究においては，集合間に明確で一方向的な対立関係が想定される場合，あるいは，評価対象とその属性等（評価視点）[2]の様に，集合同士をグループ化して区別する指標が明確な場合が対象とされた．

一方で，本稿では，ユーザの目的に応じて，着目する文書集合D_1と，比較対象としての文書集合D_2を指定する．D_1とD_2を完全に区別することは目的ではなく，両者の間の関連性も考慮したい．この時，D_1をD_2から差別化し，同時にD_2との関連性を示すキーワードの抽出と可視化を行う．例えば，ユーザが特定のテーマを掲げる学会への論文投稿を検討する際，「ユーザが投稿する論文」をD_1，「学会における関連論文集合」をD_2とすることで，ユーザの立場から，他の研究との差分を特徴的に示しつつ当該学会との関連性を失わないキーワードを重要だとみなして抽出することが可能である．また，そのようなキーワードと，D_1とD_2で共通で使用される単語の関係を可視化図で対話的に示すことができれば，投稿する論文において読者に対して強調すべき点を取捨選択できる等の利点がある．

本稿では，着目する文書集合D_1と比較対象D_2の間の相対的特徴を可視化する手法を提案する．本稿における貢献は以下の様にまとめられる．

・D_1内の文書d_kに特徴的なキーワードを抽出する際，D_1をD_2から差別化し，同時にD_2との関連性を示すような単語が重要と考え，単語の特徴量を計算する点．
・上記の特徴量を計算する際，対数尤度比に基づく尺度に，文脈類似度を導入したもの（3 節）を適用し，その有効性を検証する点．
・D_1, D_2の注目語，特徴語（3.3 節）を文共起情報によって関連づけ，両者を可視化する対話的環境を提供する点．

本論文の構成は以下の通りである．まず，2 節で関

* 連絡先：東京大学工学系研究科技術経営戦略学専攻
〒113-8656 東京都文京区本郷 7-3-1
E-Mail: kazuhirokomoda@gmail.com

連研究について述べる．3 節で提案手法を説明し，4 節で評価実験について述べ，5 節で可視化図の観察を行う．6 節で考察を行い，7 節で結論を述べ本稿をまとめる．

2 差異に着目した分析と可視化

文書集合の差異分析は，カテゴリ固有の問題特定や, 経験知の発見・共有につながる点で有益であり，様々な関連研究が行われてきた．文書集合間の差を求める手法として，事前に定義したカテゴリに基づき，各観点に属する概念を辞書として準備する方法[3]がある．例えば，製造業のコールセンターでは部品名，苦情，要望，質問といった観点とそれぞれに該当する表現を事前に集め辞書に登録し，分類された文書集合間の関係を概観する方法が用いられている．この場合，事前に固定された分析観点以外での柔軟な分析が行えないという課題がある．また特定の文書集合に特徴的な表現を抽出する方法[4]がある．これらの手法では，特定の文書集合を他の文書集合から差別化することができる単語を特徴的と判断するため，両者の間の関連性を考慮していない．

大平ら[5]は，議論参加者の共有知識を増やし，相互理解の構築を目指すという目的で，各参加者が作る「人P_iがオブジェクトO_jをI_kと考える」という組を平面に配置し差異を可視化した. この場合には, 個々人の意見の差異が簡潔な外部表現として得られなければならないという制約がある．

3 提案手法

複数文書の相対的特徴を反映する単語の抽出と可視化を行う提案手法の詳細を記述する．

3.1 事前準備

まず，ユーザが着目する文書集合D_1と，比較対象としての文書集合D_2を，重複のないように（$D_1 \cap D_2 = \phi$）用意する．例えば，ユーザが特定のテーマを掲げる学会への論文投稿を検討する際,「ユーザが投稿する論文」をD_1,「学会における関連論文集合」をD_2とすることで，ユーザの立場から，他の研究との差分を特徴的に示しつつ当該学会分野との関連性を失わないキーワードを抽出して可視化する状況を想定する．文書集合全体$D \equiv D_1 \cup D_2$に対して不要語・語尾の除去，品詞の選択（名詞・動詞・形容詞・副詞）を行う．

文書集合D, D_1, D_2で用いられる語彙の集合U, U_1, U_2について，共通部分$S = U_1 \cap U_2$に属する語（「共通語」と定義する）は，D_1とD_2において表層的に共有される主張に関連する単語を含み，一方で差集合$U_1 - S, U_2 - S$に属する語（「特有語」と定義する）は, D_1, D_2それぞれの特徴的な主張に関連する単語を含むと考えられるため，以下で「共通語」と「特有語」の一部の関係を可視化する．

3.2 単語の特徴量の計算

着目する文書集合D_1に含まれる文書d_kに対して，文書集合全体Dの語彙の各単語α_jの特徴量を計算する．今回は，単語の特徴量を決定する方法として，対数尤度比に基づく尺度LLR^+を提案する．これは，特定分野における単語の特徴度を測る尺度[6]を参考に，D_2において単語α_jが出現する際の文脈類似度c_simを新たに考慮したものである．

まず，D_1, D_2内の文書で出現する単語を（重複を許して）順に並べ，単語トークン系列$v_1, \cdots, v_{n_{D_1}}, v_{n_{D_1}+1}, \cdots, v_n$を作成する. $n = n_{D_1} + n_{D_2}$であり, n_{D_1}, n_{D_2}はD_1, D_2の単語トークン数である. 次に, 文書$d_k (k = 1, \cdots, |D_1|)$と単語$\alpha_j (j = 1, \cdots, |U|)$が与えられた時，ある単語トークン$v_i (i = 1, \cdots, n)$に対応する確率変数$W_j^i, T_k^i$の値をそれぞれ$w_j^i, t_k^i$とし，以下の様に定義する．

$$w_j^i = \begin{cases} 1 \ (v_i = \alpha_j) \\ 0 \ (otherwise). \end{cases}$$

$$t_k^i = \begin{cases} 1 \ (v_i \text{は} D_1 \text{内で出現}) \\ 1 \ (v_i \text{は} D_2 \text{内で出現}, v_i = \alpha_j, c\_sim(v_i, d_k) \geq \theta_{sim}) \\ 0 \ (otherwise). \end{cases}$$

$c_sim(v_i, d_k)$は，集合間の Jaccard 係数を用いて以下の様に定義する．

$$c\_sim(v_i, d_k) = \max_l \left(\text{Jaccard}\left(ctx(v_i), ctx(v'_l)\right) \right).$$

ただし，トークンv_iの文脈$ctx(v_i)$を，v_iが出現する文で使用される単語の集合（ただし単語v_i自身は除く）と定義し，$d_k \in D_1$においてv_iと同じ値を持つ単語トークン$v'_l (l = 1, \cdots, L)$の文脈を，$ctx(v'_l)$ $(l = 1,2, \cdots, L)$とする．θ_{sim}は類似度の閾値である．以上で定義された変数の関係を表 1 に示す.

この時，それぞれの単語トークンv_iが確率的に独立であると仮定すると，系列$\Upsilon_{jk} = \langle w_j^1, t_k^2 \rangle, \langle w_j^2, t_k^2 \rangle, \cdots, \langle w_j^n, t_k^n \rangle$の生起確率は以下で定義される．

$$Pr(\Upsilon_{jk}) = \prod_{i=1}^{n} Pr(W_j^i = w_j^i, T_k^i = t_k^i).$$

また，単語トークンv_iについて，確率変数W_j^i, T_k^iの値

表 1: 文書d_k ，単語α_jに関する確率変数W_j^i, T_k^i．?は，$c_sim(v_i, d_k) \geq \theta_{sim}$ならば1

	D_1	D_2
i	$1, \cdots, n_{D_1}$	$n_{D_1}+1, \cdots, n$
v_i	$v_1, \cdots, v_{n_{D_1}}$	$v_{n_{D_1}+1}, \cdots, v_n$
W_j^i	010100010101	0100010000000000
T_k^i	111111111111	0?000?0000000000

表 2: 単語トークンv_iについての集計結果

	$T_k^i = 1$	$T_k^i = 0$
$W_j^i = 1$	a	b
$W_j^i = 0$	c	d

で出現頻度を集計したものが表 2 である．ここで，$a+b+c+d=n$である．文書d_kにおける単語α_jがD_1をD_2から差別化し，同時にD_2との関連性を示すならば，W_j^iとT_k^iの依存度が高まるという意図の下，表 2 の集計を行った．

表 2 に基づき，仮説H_{indep}「確率変数W_j^iとT_k^iとは互いに独立である」に対し，次の対数尤度比$LLR_0^+(d_k, \alpha_j)$を考える．

$$LLR_0^+(d_k, \alpha_j) = \log \frac{Pr(\Upsilon_{jk})}{Pr(\Upsilon_{jk}|H_{indep})}$$

$$= \sum_{i=1}^{n} \log \frac{Pr(W_j^i = w_j^i, T_k^i = t_k^i)}{Pr(W_j^i = w_j^i, T_k^i = t_k^i | H_{indep})}.$$

上式において$Pr(W_j^i = w_j^i, T_k^i = t_k^i | H_{indep})$は$H_{indep}$が成立するとした場合の$Pr(W_j^i = w_j^i, T_k^i = t_k^i)$であり，その値は$Pr(W_j^i = w_j^i) Pr(T_k^i = t_k^i)$に等しい．表 2 に基づく推定により，以下の値が計算できる．

$$Pr(W_j^i = 1, T_k^i = 1) = \frac{a}{n}, Pr(W_j^i = 1, T_k^i = 0) = \frac{b}{n},$$

$$Pr(W_j^i = 0, T_k^i = 1) = \frac{c}{n}, Pr(W_j^i = 0, T_k^i = 0) = \frac{d}{n},$$

$$Pr(W_j^i = 1) = \frac{a+b}{n}, Pr(W_j^i = 0) = \frac{c+d}{n},$$

$$Pr(T_k^i = 1) = \frac{a+c}{n}, Pr(T_k^i = 0) = \frac{b+d}{n}.$$

$LLR_0^+(d_k, \alpha_j)$は，確率変数W_j^iとT_k^iの依存性の度合いが高いほど大きな値を取るため，以下の様に補正した$LLR^+(d_k, \alpha_j)$を用いることで，d_kにおいてD_1をD_2から差別化し，同時にD_2との関連性を示す単語α_jに対して高い特徴量を与えることができる．

$$LLR^+(d_k, \alpha_j) = sign(ad - bc) LLR_0^+(d_k, \alpha_j).$$

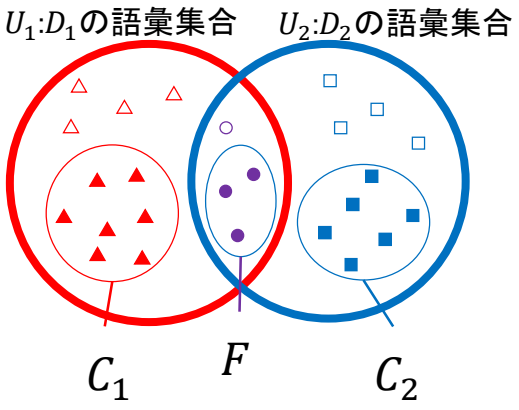


図 1: 「共通語」「特有語」と「注目語」「特徴語」の関係．赤い円は文書集合D_1の語彙集合U_1，青い円は文書集合D_2の語彙集合U_2を表す．共通語から選ばれた特徴量上位の単語が注目語（紫で塗られている），特有語から選ばれた特徴量上位の単語が特徴語（赤あるいは青で塗られている）である．

3.3 可視化する注目語と特徴語の選択

共通語と特有語の中から，可視化に使用する単語を選ぶ．まず，3.2 節により，D_1に含まれる各文書d_kに対して，各単語$\alpha_j \in U$の特徴量を計算する．次に，D_2に含まれる各文書d_kに対しても特徴量を計算するため，D_1とD_2を入れ替えて 3.2 節を再度行う．以上で，全文書が全単語の特徴量情報を持つため，以下用いる単語の特徴量は，各文書における特徴量の和とする．共通語から特徴量上位のものを指定個数選んだものを「注目語」，D_1, D_2の特有語から特徴量上位のものをそれぞれ指定個数選んだものを「特徴語」と定義する．以下，注目語集合をF，文書集合D_1, D_2の特徴語集合をそれぞれC_1, C_2とする．以上の関係を図 1 に示す．

3.4 注目語と特徴語の文共起情報

ある注目語$f \in F$が文書集合D_1において出現したものをfのトークン(token)と呼び，f'とする．f'が出現する同一文中に，特徴語$c_1 \in C_1$のトークンc_1'が存在すれば，$(f, c_1 n_1(f, c_1))$の組を作成する．$n_1(f, c_1)$の値は 3.3 節で求めた特徴語c_1の特徴量とする．注目語fと特徴語c_1に関するグラフを作成するため，f

とc_1の間のリンクの重みを以下で計算し，重みに応じた太さを持つリンクを描画する．

$$w(f,c_1)=\frac{|c_1|n_1(f,c_1)}{\sum_{\widehat{c_1}}|\widehat{c_1}|n_1(f,\widehat{c_1})}.$$

ただし，$|c_1|$は特徴語c_1のD_1におけるトークン数，$\sum_{\widehat{c_1}}|\widehat{c_1}|n_1(f,\widehat{c_1})$は$f$を含む全ての組の重みの総和である．以上を文書集合$D_2$，特徴語集合$C_2$についても同様に行い，合わせて描画することで，ある注目語$f\in F$について，注目語fと特徴語の文共起情報に基づくグラフが得られる．このグラフは，注目語が，それぞれの文書集合においてどのような特徴語と共に用いられるかを示す．文書集合間の共通語のみに着目した従来の要約・集約において看過されやすい，配慮すべき重要な論点の展開の違いに焦点を当てている．

4 実験

4.1 実験仕様

3.2 節の特徴量計算を評価するため，論文タイトルに含まれるべき単語を論文要旨から推測する評価実験を行った．論文のタイトルは，他の研究者が最初に注目する重要な部分であり，既存研究との差分を特徴的に示しつつ，分野における一般性や関連性を失わない簡潔な言語表現が求められる．このような言語表現は，論文要旨を，関連する他の論文要旨と比較した際に抽出できる場合が多いと考えた．今回は，2013 年度人工知能学会全国大会「自然言語」セッションに投稿された 33 論文要旨d_k $(k=1,\cdots,33)$を利用し，前処理後の要旨情報から「他の論文との差分を特徴的に示しつつ，セッションとの関連性を示す(*)」単語を 3.2 節の手法で抽出した．3.1 節において，ある 1 本の論文要旨d_kをD_1，残りの 32 本の論文要旨をD_2とすることで，d_kについて単語を特徴量によってソートしたリストを得て(retrieved)，対応する論文タイトルから(*)を満たす単語を人手で選択した(relevant)．θ_{sim}は経験的に 0.22 と設定した．なお，比較手法LLRは，t_k^iの定義式において右辺 2 行目を考慮しない．

評価指標として，$k=1,\cdots,33$の 33 論文の Mean Average Precision(MAP)を算出した．MAPは以下の式で定義される．

$$MAP=\frac{1}{|D|}\sum_{d_k\in D}\frac{1}{|\Omega_k|}\sum_{\alpha_j'\in\Omega_k}Precision\left(rank\left(d_k,\alpha_j'\right)\right).$$

ここで，Ω_kは論文要旨d_kのタイトルから人手で抽出した(*)を満たす語の集合（正解データ）であり，

表 3：実験結果（人手による(*)の選定）

手法	MAP
LLR^+	0.281
LLR	0.278

$rank(d_k,\alpha_j')$は文書$d_k\in D$において単語$\alpha_j'\in\Omega_k$が手法によって得た順位，$Precision\left(rank\left(d_k,\alpha_j'\right)\right)$はその順位までの出力結果における適合率である．この評価指標では，各文書d_kについて，手法による順位づけ結果の上位$|\Omega_k|$単語の中にΩ_k内の単語が全て含まれる場合が最も望ましいとされ，この時 Average Precision の値は 1 である．

4.2 実験結果

実験の結果を表 3 に示す．MAP値に関して，全体的な性能の向上が確認できた．実際，1 つの論文を除いて Average Precision 値が不変または増加している．以下，特徴的な個別事例の詳細を確認する．

表 4 は，LLR^+とLLRにおけるAP値の差が最大$(0.372-0.268=0.104)$となった事例である$(k=3)$．LLR^+では，文書の潜在トピックを抽出する手法である Latent Dirichlet Allocation(LDA)に関する単語をより上位で抽出でき，「自然言語」セッションへの関連性を示したと言える．また，LLRで上位に来ている tag や put という単語は，当該論文における具体的な処理に関する単語である．他の論文との関連性が低い単語の順位を下げることができたために，比較的良い結果が得られたと考えられる．

表 5 はLLR^+とLLRにおけるAP値の差がないが，値が最も高い(0.603)事例である$(k=19)$．非常に少数の文によって提案内容のみが的確に表現されており，かつ「完備束表現」という表現自体が「自然言語」セッションにおいては非常に特徴的な表現だったため，文脈類似度が設定した閾値θ_{sim}を上回りにくく，LLR^+とLLRで同じランキング結果となった．タイトルに含まれるべき単語が上位に抽出されており，LLRによって得られる Average Precision が最も高かった．

表 6 は，LLR^+とLLRにおけるAP値の差が最小$(0.250-0.293=-0.043)$となった事例である$(k=9)$．33 論文要旨の中で唯一 Average Precision が低下した．LLRでは$|\Omega_k|=4$のうち 3 単語が上位 4 単語に含まれているが，LLR^+ではΩ_kのうち 2 単語が低い順位を得ているため，僅かではあるが Average Precision が低下した．

表 4：$k = 3$における手法ごとの特徴量上位語を順に並べたリスト．下線は$\Omega_{k=3}$に含まれる単語．

$\Omega_{k=3}$	{supervis, alloc, dirichlet, latent, pseudo, label}
LLR^+	latent, document, llda, alloc, dirichlet, label, topic, lda, tag, put, estim, limit, pseudo, abil, exce, supervis
LLR	document, llda, label, latent, tag, put, topic, estim, limit, pseudo, abil, exce, supervis, applic, alloc, dirichlet

表 5：$k = 19$における手法ごとの特徴量上位語を順に並べたリスト．下線は$\Omega_{k=19}$に含まれる単語．

$\Omega_{k=19}$	{represent, languag, lattic, natur, data, structur, complet}
LLR^+, LLR	represent, complet, bag-of-word, add, data, structur, lattic, mathematic, defin, order, …

表 6：$k = 9$における手法ごとの特徴量上位語を順に並べたリスト．下線は$\Omega_{k=9}$に含まれる単語．

$\Omega_{k=9}$	{topic, domain, user, knowledge}
LLR^+	latent, user, knowledge, alloc, dirichlet, difficulti, domain, term, lda, captur, technic, topic
LLR	user, knowledge, difficulti, domain, term, captur, technic, depend, key, background, level, belong, adapt, focu, experiment, show, previou, person, lot, determin, inform, topic

5　可視化図の観察

公開報告書の様な文書は，組織内外の人間に広く読まれることを意図しているにも関わらず，分野の専門知識を前提として記述され，読み手にとっては複雑な構造を持つ傾向にある．専門家が，難解な報告書を非専門家である読み手に分かりやすく伝えることを考える際，非専門家にとってより身近な資料を用意して両者の位置づけを比較することで，有益な指針が得られると考えられる．日本原子力安全部会「福島第一原子力発電所事故に関するセミナー第 8 回」議事メモの一部（東北地方太平洋沖地震の概況及び福島第一原子力発電所の事故の概要）をD_1，「福島第一原発　原子力安全」というキーワードで検索して得られた Yahoo!ニュース記事数件をD_2とした時，3 節の処理を経た可視化図を図 2 として掲載する．特徴量計算によって選択された注目語，特徴語が上から順に並んでいる．

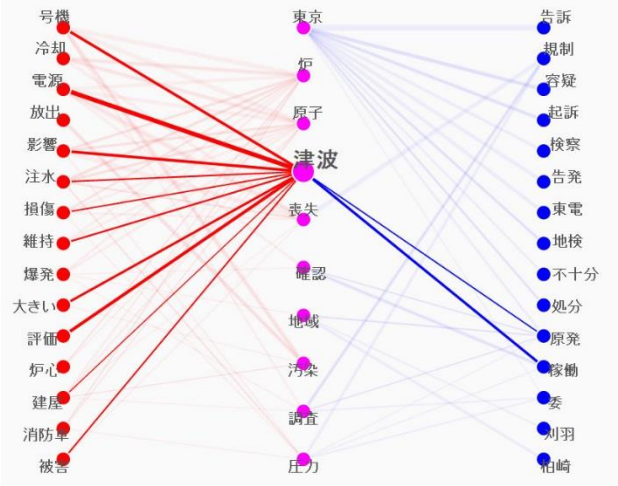


図 2:「津波」にマウスを載せてフォーカスする．

なお，図 2 は Web ブラウザ上での閲覧・操作を意図している．一度に表示されるリンクの数を重みの閾値によって変更でき，最も重要なリンクから優先して確認することができる．また，ノード（＝単語）の上にマウスを載せると，そのノードがハイライトされ，関連するリンクのみが表示されるため，特定の単語に容易に着目できる．

図 2 では，注目語，特徴語自身に加え，それぞれの単語が他のどのような単語と共起しているかが読み取れる．例えば，D_1に出現する特徴語の多くが「炉」や「原子」といった単語と共起しているが，一方でD_2においてそれらの単語は特徴語との共起は多くなく，リンクは「東京」に集中している．D_1とD_2を直接結びつけている「津波」に関しては，注目語であってもその使われ方が大きく異なることが推測できる．D_1を執筆する立場としては，D_2の特徴語との関連性が強い話題を強調して説明し，一方で，D_2を収集する立場としては，例えば，刑事告発に関する記事を減らして原子炉や汚染に関する記事を増やすなど，よりD_1との関連性が強くなるような関連情報を収集して再度可視化を行うような，チャンス発見[7]の二重らせんプロセスを行うことが可能である．

6　考察

4 節の論文タイトル課題において，提案手法LLR^+は比較手法LLRに対して，全体的な性能の向上が確認された．4 節の個別事例に留まらずより一般的に，提案手法と比較手法の両方における性能，そして提案手法に対する比較手法の性能を向上させるために考慮すべき要素を検討する．

課題全体の性能を向上させるため，課題に沿った適切な前処理を行うことが望ましい．まず，複合語・

類義語を考慮に入れる必要がある．具体的には”Latent Dirichlet Allocation”の様な単語列を複合語として扱い，さらに”LDA”の様な省略形と類義語であるという情報を与えることに相当する．また，現在は，名詞・動詞・形容詞・副詞を取り扱っている．これは，文中で強い意味内容を持ち他の単語との関連性が高い可能性のある単語を網羅するという意図の下行った処理であるので，今後，不要な単語を取り除くことが可能である．

提案手法を比較手法に対して向上させるためには，タイトルに含まれるべき単語の候補を絞ることが望ましい．論文のタイトルに含まれる単語は全てが同じ重要度を持つわけではなく，単語自体の機能，著者が命名において重視する観点に応じて，タイトルに含まれるべき単語は変化するため，単語の重み情報を取り入れた評価を行うことができる．実際，今回は，前処理では取り除かなかったが，「他の論文との差分を特徴的に示しつつ，セッションとの関連性を示す(*)」を満たさない単語を人手で除去した．これは(*)を満たす単語に1，満たさない単語に0の重みを与えることに相当する．今後は，タイトルに不適切な語の除去だけでなく重要語の強調も行い，また一人の評価者ではなく，論文の著者に依頼して得られた重み情報による実験を行うことが考えられる．

5 節の可視化図においては，まず全語彙を特徴量で順位づけし，共通語集合，特有語集合に属する単語から特徴量上位のものを選んで可視化図に出力し（注目語と特徴語），その後に注目語と特徴語の文共起情報を用いてリンクを結んだ．したがって，現状では単語の順位は全て，3 節で求めた特徴量に依存している．また，文書集合間の特徴語は，必ず図の中心にある注目語を媒介として結びついているように可視化されている．したがって，注目語については，文書集合間の特徴語（ピンクノード）を関連付ける度合いによって選定する方が望ましい可能性がある．あるいは，3 節で特徴的かつ関連性を示す単語を抽出し，その際に文脈類似度を導入したことを考えると，その結果を生かし，文書集合間の特徴語同士（赤ノードと青ノード）をより直接的に結びつけるような可視化によって，ユーザが意外な関連性に気づくことができる可能性がある．その実現のため，文脈類似度c_simの計算方法と閾値設定も課題である．今回は Jaccard 係数を用いて定義したが，cosine 類似度や自己相互情報量など別の方法との比較も検討したい．

7 結論

本稿では，着目する文書集合D_1と比較対象D_2の間の相対的特徴を可視化する手法を提案した．D_1に特徴的な単語を計算する際，対数尤度比に基づく尺度に文脈類似度を導入してD_2との関連性を示す単語を高く評価する点，注目語と特徴語の文共起情報に基づく対話的可視化環境を提供する点が特徴である．ある論文要旨に特徴的で，他の論文要旨集合との関連性を示すような単語を抽出して論文タイトルを再現する実験を行い，提案手法の有効性を示した．また，2 つの文書集合における「注目語」と「特徴語」の間の文共起情報を用いた対話的インタフェースを提供し，専門的な報告書と関連するニュース記事を題材に，観察結果を報告した．

参考文献

[1] Mack, R. and Hehenberger, M.: Text-based knowledge discovery: search and mining of life-sciences documents, Drug Discovery, Vol. 7, No. 11, pp. S89-S98, (2002)

[2] 乾孝司, 板谷悠人, 山本幹雄, 新里圭司, 平手勇宇, 山田薫: 意見集約における相対的特徴を考慮した評価視点の構造化, 自然言語処理, Vol. 20, No. 1, pp. 3-26, (2013)

[3] 那須川哲哉: コールセンターにおけるテキストマイニング, 人工知能学会学会誌, Vol. 16, No. 2, pp. 219-225, (2001)

[4] Hisamitsu, T. and Niwa, Y.: Topic-Word Selection Based on Combinatorial Probability, NLPRS, Vol. 1 (2001)

[5] 大平雅雄, 山本恭裕, 蔵川圭, 中小路久美代: EVIDII: 差異の可視化による相互理解支援システム, 情報処理学会, Vol. 41, No. 10, pp. 2814-2826, (2000)

[6] 内山将夫, 中條清美, 山本英子, 井佐原均: 英語教育のための分野特徴単語の選定尺度の比較, 自然言語処理, Vol. 11, No. 3, pp. 165-197, (2004)

[7] Ohsawa, Y. and McBurney, P.: Chance Discovery, Berlin-Heidelberg, Springer Verlag, (2003)

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5回)
SIG-AM-05-08

無意味スケッチ図形を命名する

Naming of meaningless sketch image

荒牧英治 [1, 2]　仲村哲明 [1]　臼田泰如 [3]　久保圭 [4]　宮部真衣 [1]

Eiji ARAMAKI[1, 2]　Tetsuaki NAKAMURA[1] Yasuyuki USUDA [3]　Kay KUBO[4]　Mai MIYABE [4]

[1] 京都大学 デザイン学ユニット
[1]Design Unit, Kyoto University
[2] 科学技術振興機構 さきがけ
[2] JST PRESTO
[3] 京都大学 人間・環境学研究科
[3] Graduate School of Human and Environmental Studies, Kyoto University
[4] 大阪大学 日本語日本文化教育センター
[4] Center for Japanese Language and Culture, Osaka University

Abstract: We basically could give any name to an object, because the relation between an object and its name is generally arbitrary. However, we find that some names fit the object, and other do not. This indicates that names are restricted by the objects to some extent. To investigate this restriction, this study operated naming test of meaningless images. The experimental results revealed that voiced sounds and vowel /a/ matched among the subjects, indicating that these sounds are bounded by the images. In contrast, a vowel /e/ is less restricted. These results would contribute to name generation, which is a unexplored research field.

はじめに

名前とその指示物との関係は恣意的である. つまり，一般的に名前をつけるという行為はまったく自由に行うことができると考えられる．しかし，しっくりとくる名前というものはある．例えば，命名を行った実験ではないが，ブーバ・キキ効果（図 1）においては丸みのある物体が「ブーバ」だと判定する傾向が強い．これはある種の画像と音または文字（ここでは，これらをあわせて言語形式と呼ぶ）の間に関連性（有縁性，有契性）があることを示している[1]．しかし，この関連性は言語形式を唯一物に限定するほど強いものではない．例えば，先の例では，「ブーバ」が最も強い連想を喚起する言語形式というわけではなく，「ブーバ」と「ブーボ」のどちらがよいかと問われれば，明確な答えは得られないであろう[1]．このように，画像とその言語形式は，なんらかの関連性を持つ場合と持たない場合があり，その関連性について，音象徴をはじめとして多くの研究が行われてきた．

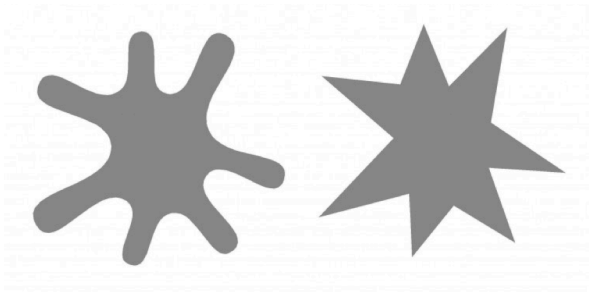


図 1: ブーバ・キキ効果
どちらかの図形が「ブーバ」で他方が「キキ」だとする．どちらがどちらの印象として適切ですか？との問いに実験協力者の 95%が右の図形をキキ，左の図形をブーバだと判定した．

例えば，Ohara は母音/a/が大きなもの，柔らかさ，鈍重な動きを表すと報告している．逆に，母音/i/は小ささを示すとされている[2]．日本語においても，金田一[3]や牧野[4]らは，/b/が大きさを連想させるとしている．これらの実験では，ブーバ・キキ効果のような言語形式の選択課題や，言語形式の印象に対する形容詞選択課題など選択課題をベースにしており，「大きさ／小ささ」「強さ／弱さ」「鈍重さ」など物体や現象の性質に着目して知見が集積されつつあ

[1] 「ブーバ」「ブーボ」については，実験を行ったわけではなく，母語話者の直感による

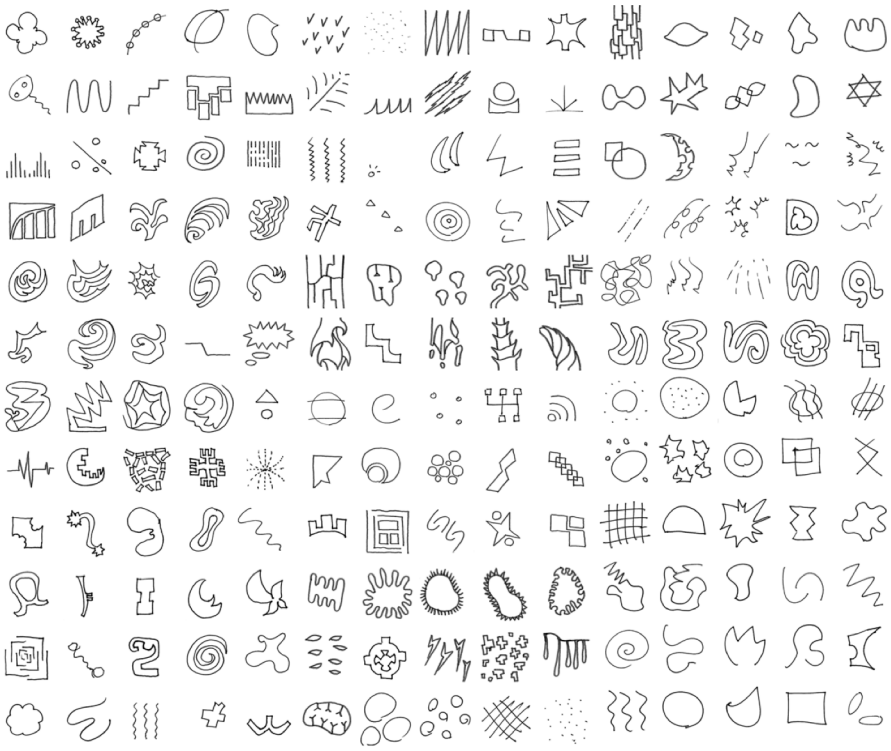


図 2: 実験に用いた全図形
実験に用いた無意味図形はサイト(http://mednlp.jp)にて公開している.

	実験参加者 1	実験参加者 2	実験参加者 3	実験参加者 4	実験参加者 5
	グルグル	グルグル	グルルル	グルグル	グルグル
	グシャグシャ	グジャグジャ	モジャモジャ	グシャグシャ	グチャグチャ
	ギザギザ	ガギガギ	グワシャ	ガラガラ	ギザギザ
	モゾモゾ	ズモモモ	ズモモモ	ウゾウゾ	モジャモジャ
	ガチガチ	ボロボロ	ピシッ	ガチャガチャ	バサバサ
	グルグル	モケモケ	ハナハナ	ホワホワ	フワフワ
	ボコボコ	ガジガジ	(Skip)	ムシャムシャ	パクパク

表 1：実験に用いた刺激と人間による命名の例.

る.

しかし，先行研究の調査法は，一部の性質に着目しているため，ある画像に対して，どれくらいの言語形式が許容されているか，逆に言語形式のどの部分が制約されているのか，といった可能な言語形式の可能性に関する知見を得ることは困難である.

そこで，本研究では，画像刺激に対して，自由に言語形式をあてはめるという課題を行い，作成された名前の類似性を考察する．実験の結果，濁音と母音/a/の選択に一定の一致度が見られた．これは，画

像が濁音と母音/a/の使用に影響をあたえていることを示している．

関連研究

音象徴の観点から

言語とその指示物との関係としては，音象徴[5-9]がよく知られている．音象徴は，音そのものがある特定のイメージを喚起する事象を指し，ブーバ・キキ効果はその代表的な例である．さらに，音がない現象についても，他の感覚を媒介として音で示すことが可能で，舌の位置の高い母音（狭母音、例えば/i/）が小さいもの，鋭いもの，すばやい動きと関連する．逆に，舌の位置の低い音（広母音、例えば/a/）が大きなもの，柔らかさ，鈍重な動きと関連することなどが指摘されている[2]．

同様の研究は，日本語についても行われ，大きさ，力強さに関連する音として，/g/,/z/,/d/や/b/があげられ，静かさや柔らかさに関連する音として/s/ や /j/があげられている[10]．このように言語や文化にかかわらず音象徴に関する報告は多く，音象徴は普遍的な現象だとする見方もある．その一方，一部の報告[11]では普遍性をもたないと主張している．

このように音象徴の普遍性に関する議論が困難である理由の一つは，多くの音象徴研究が一部の音と指示物の一部の特徴のみを個別に扱っており，可能な名前すべてを扱っていないことにある．一方，本研究は画像から自由に言語表現を割り当てる実験を行う点が新しい．

マーケティングの観点から

毎年，市場には膨大な数のブランドネームが導入されている．例えば，特許庁の発表によると日本では 2010 年に 167,326 件の商標の出願が行われている ．アメリカでも 2005 年の 1 年間で 320,000 件の商標が新たに登録されている．こうしたなか，多くの企業はブランドネームに巨額の資金を投じている．例えば，ネーミングを行う会社は 1 ネームあたり 5 万から 18 万までの費用を請求する．これは，ネーミングの巧拙がブランド認知や製品評価において重要な要素となりうるからである[12]．近年のネーミングに関する研究では，単語や音楽が持つ意味だけでなく，それらから引き起こされる感情に対しても注目が集まっている．本研究の知見は，マーケティングにおけるネーミングにも貢献すると考えられる．

材料と方法

イメージがどの程度名前と関連性を持つのかを調査するために，まず入力となるイメージを作成した．

分類	例	分類	例
母音	/a/ /i/ /u/	濁音	ガ, ギ
子音	/k/ /s/ /t/	半濁音	パ, ピ
撥音	ン	拗音	キャ
促音	ッ	合拗音	クヮ
長音	ー		

表 2: 実験で用いた音素カテゴリ．

母音平均 0.175	/a/ 0.262	/i/ 0.208	/u/ 0.173	/e/ 0.038	/o/ 0.192
子音平均 0.132	/k/ 0.207	/s/ 0.213	/t/ 0.169	/n/ 0.187	/h/ 0.109
	/m/ 0.102	/y/ 0.087	/r/ 0.156	/w/ 0.087	
撥/促/長平均 0.056	**撥音** 0.058	**促音** 0.074	**長音** 0.037		
濁/半濁平均 0.221	**濁音** 0.248	**半濁音** 0.191			
拗/合拗平均 0.055	**拗音** 0.082	**合拗音** 0.027			

表 3: 音素カテゴリでの Kappa 値.
Kappa 値が 0.2 を超えた値を太字で表示している．

図 3: 2 名以上が命名を skip した画像.

図 4: OCR ライブラリを用いた命名アプリケーション.

これは実験参加者 6 名が 30 個ずつ合計 180 個の無意味画像を描画した．描画にあたっては，先行研究[13]と同様に，形態的に無意味で物体を容易に連想できないことを制約とした．

次に，これらに対して，実験参加者 5 名が命名を行った．命名は，具体物と関連を持たない，かつ，オノマトペ（擬音語，擬声語，擬態語）によって名前をつけることを制約とした．命名が困難な画像については，命名せずにスキップを許容した．命名された結果例を表 1 に示す．結果，スキップされた 45 個を除き，855 個（=180×5−45）の名前が得られた．

最後に，このデータを用いて，無意味図形と名前の音素の関係を調査した. 調査は次の 2 つを行った．

まず，名前を音素（表 2）に変換し[14]，どの程度音素が一致しているかを調査した．さらに，音素を母音群，子音群，濁音／半濁音などカテゴリ化して一致の調査も行った．

結果

実験結果を表 3 に示す．母音平均は各母音 /a/, /i/, /u/, /e/, /o/を平均した値である．子音平均など他の音素カテゴリについても同様である．

一致する音素について

高い一致をみた音素は，/a/，濁音，/s/, /i/, /k/ であった. 特に/a/と濁音の Kappa 値は 0.25 と他と比較して高い．この結果は，本来，自由に音素を使えるはずであるのに，母音の/a/と濁音の使用に関しては束縛される傾向があることを示す．なお，この知見は，/a/が大きさを示す，逆に/i/が小ささを示す，濁音が力強い印象を持つ，など音素が画像と関連の深い連想をおこさせるという先行研究の知見と矛盾するものではない.

一致しない音素について

高い一致をもつ音素がある一方，チャンスレベルでしか一致しない音素もある．例えば，母音の/e/，撥音, 長音, 合拗音は Kappa 値が 0.05 を下回っており，人間間の相関はほぼみられない．つまり，これらの音素は恣意的に用いられていると言える．

特に，母音の/e/に関しては他の母音が高い 0.17 以上という比較的高い Kappa 値をもつなかで，0.038 と極端に低い一致率となっている．この理由の推定は困難であるが，音韻的には/e/は非円唇前舌半狭母音であり，発音時の口のサイズ，舌の位置などにおいて，他の母音ほどの特徴を持たないことに起因する可能性がある．今後のさらなる考察が望まれる．

名状しがたい画像について

名前をつけることができずスキップした割合は，平均 9 個（5%）であった．本来，名前は自由につけてよいはずであるのに，スキップが一定の割合で行われることは，人は名前と指示物になんらかの関連性を必要とし，それが不可能である場合に不適切な名前をつけてしまうことに，ストレスを感じる可能性がある． 実験で，2 名以上が名前をつけることができなかった図形（6 個）の例を図 2 に示す．これはある種の画像は共通して名状することが困難であることを示している．

画像類似度による名前生成

本研究のデータを用い，入力画像ともっとも類似した画像を検索し，その音素を出力することで，名前生成システムを構築することができる．実装例を図 3 に示す.

おわりに

本研究では，画像刺激に対して，自由に言語形式をあてはめるという命名課題を行い，作成された名前の類似性を考察した．実験の結果，濁音と母音/a/の選択に一定の一致度が見られ，画像が濁音と母音/a/の使用に影響をあたえていることが分かった．本研究の成果により，命名という創造的な行為における制約の一部が明らかになったといえる. 今後, なぜ, これらの音素が影響を受けるのか考察が待たれる.

謝辞

本研究の一部は JST 戦略的創造研究推進事業(さきがけタイプ)「情報環境と人」および科研費補助金(若手研究 A)による.

参考文献

[1] Ramachandran, V. and E.M. Hubbard, Synaesthesia - A window into perception, thought and language. Journal of Consciousness Studies, 2001. 8(12): p. 3-34.

[2] Ohala, J.J., The frequency codes underlies the sound symbolic use of voice pitch. Sound symbolism. Cambridge. 1994: Cambridge University Press. 325-347.

[3] 金田一春彦, 擬音語・擬態語概説 「擬音語・擬態語辞典」. 1978: 角川書店.

[4] 牧野成一, 「音と意味の関係は日本語では有縁か 鼻音 vs. 口蓋音と文法形式のケーススタディ」. 言語学と日本語教育実用的言語理論の構築を目指して』. 1999: くろしお出版.

[5] Doyle, J.R. and P.A. Bottomley, Mixed Messages in

Brand Names: Separating the Impacts of Letter Shape from Sound Symbolism. Psychology & Marketing, 2011. 28: p. 749-762.

[6] Ibarretxe-Antunano, I., Sound Symbolism and Motion in Basque. Muenchen: LINCOM EUROPA. 2006.

[7] Klank, L.J.K., H. Y.H, and R.C. Johnson, Determinants of success in matching word pairs in tests of phonetic symbolism. Journal of Verbal Learning and Verbal Behavior, 1971. 10(2): p. 140-148.

[8] Westbury, C., Implicit sound symbolism in lexical access: Evidence from an interference task. Brain and Language, 2005. 93: p. 10-19.

[9] Y.H, H., S. Pratoomraj, and R.C. Johnson, Universal magnitude symbolism. Journal of Verbal Learning and Verbal Behavior, 1969. 8: p. 155-156.

[10] 田守育哲 and スコウラップ、ローレンス, 『オノマトペ―形態と意味―』. 日英語対照研究シリーズ（6）. 1999: くろしお出版.

[11] Diffloth, G., i: big, a: small, in Leanne Hinton. Sound Symbolism, ed. J. Nicols and J.J. Ohala. 1994: Cambridge University Press.

[12] Argo, J., M. Popa, and M. Smith, The Sound of Brands. Journal of Marketing, 2010. 74(4): p. 97–109.

[13] 信貴遠藤, et al., 無意味輪郭図形の階層的特徴記述に基づく知覚判断特性の分析 心理学研究.

[14] 仲村哲明, 宮部真衣, and 荒牧英治. オノマトペが属する五感の推定. in インタラクティブ情報アクセスと可視化マイニング研究会（SIG-AM）第４回研究会. 2013.

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第5回)
SIG-AM-05-09

NTCIR-11の概要

Outline of NTCIR-11

岸田和明[1*]　上保秀夫[2,3]　神門典子[3]
Kazuaki KISHIDA[1]　Hideo JOHO[2,3]　Noriko KANDO[3]

[1] 慶應義塾大学文学部図書館・情報学専攻
[1] School of Library and Information Science, Keio University
[2] 筑波大学図書館情報メディア系
[2] Faculty of Library, Information and Media Science, University of Tsukuba
[3] 国立情報学研究所
[3] National Institute of Informatics

Abstract: NTCIR (NII Testbeds and Community for Information access Research) is one of the international workshops for research on information retrieval, question answering, natural language processing and so on, which is organized by the NII (National Institute of Informatics) in Japan. From September 2013, the eleventh NTCIR workshop (NTCIR-11) has started. This paper describes a framework, research tasks and schedule of the NTCIR-11 briefly.

1　はじめに

情報検索技術の向上を主目的として，米国においてTREC（Text REtrieval Conference）が始まって以来，約20年が経過した。この間，インターネットのサーチエンジンをはじめとして，情報検索およびその関連技術は飛躍的に向上したが，これには，TRECをひとつの端緒とするいわゆる「評価型ワークショップ」が大きな影響を与えたことは周知のとおりである。

「評価型ワークショップ」の基本的な目的は，数多くの研究者の参加・協力の下に，開発された諸技術に対する妥当かつ信頼性の高い評価を可能とする，テストコレクション（test collection）を構築することにある。このテストコレクションが，新たに開発された手法の検証に確実な基盤を与え，既存の他の手法との比較評価を実現することにより，科学的な知識の着実な蓄積がもたらされ，技術の発展へとつながっていく。

日本においても，国際的なプロジェクトとして，当時のNACSIS（学術情報センター）がこの事業に着手し，1998年に最初の成果報告会が開催された（NTCIR-1）。それ以来，1年半の間隔で着実に回を重ね，2012年12月に，NTCIR-10の成果報告会（カンファレンス）が盛況のうちに終了した。この間，NACSISは，NII（国立情報学研究所）となり，現在では，NTCIRは「NII Testbeds and Community for Information access Research」の略称である。

この名称が示すとおり，研究対象が情報検索から「情報アクセス」へと拡大した。これは，NTCIRプロジェクトが，質問応答や自動要約，機械翻訳などの研究テーマを幅広く網羅するようになったためである。また，名称中の「Community」は，NTCIRが，これらのテーマに関する研究者の共同体として機能することを示唆している。一種の「ビッグサイエンス」として，多くの研究者が協調して問題に取り組むことにより，全体として大きな成果を生み出すことを，NTCIRは企図しているわけである。

実際，NTCIR-10には，世界の16の国と地域から100を超える機関が参加し，成果報告会では，研究者の間で活発な議論が交わされた。この議論の結果を各自が持ち帰り，それぞれの研究に活かすことにより，次の回に向けての新たな創意工夫が生まれる可能性が高い。このサイクルこそが，技術の着実な向上をもたらす原動力（エンジン）である。本稿では，以下，次の段階として2013年9月にスタートしたNTCIR-11の内容や日程について述べる。

2　NTCIR-11の内容

その時点での研究状況に合わせて，各回ごとに，異なる研究テーマが複数設定される。NTCIRでは，この研究テーマをタスク（task）と呼んでいる。NTCIRへ参加する研究者は，自分の研究テーマに応じて，関心

*連絡先：慶應義塾大学文学部
〒 108-8345 東京都港区三田 2-15-45
E-mail: kz_kishida@z8.keio.jp

あるタスクを選び，登録しなければならない。そして，各タスクの規定に従って自分のシステム（手法）で実験を試み，その結果を提出する。それらの結果に基づいて，テストコレクションが構築されるとともに，各研究者は，自分のシステムの有効性を，他のシステムと比較しつつ，検証することが可能となる。

各タスクの実行は，タスクオーガナイザ（task organizer: TO）により管理される。TO の仕事は多彩であり，NII による支援の下に，1) タスクへの参加者（participants）を募集，2) 実験用コーパス（文書データ）を準備し，参加者に配布，3) 参加者から提出された結果を集約し，評価，4) 参加者が用いた技術やその有効性をレビューする必要がある。通常，TO はそのテーマに関する第一線の研究者であり，タスクの遂行によって，当該領域の研究を牽引する重要な役割を担っている。

したがって，どのようなタスクを設計し，採用するかは，たいへん重要であると同時に，困難な作業でもある。今回も前回に引き続き，タスク提案を公募し，プログラム委員会（Program Committee：PC）の審査およびその結果に関する General Co-chairs による議論を経て[1]，現時点で（2013 年 9 月），以下の 8 つのタスクを採択し，NTCIR-11 をスタートさせた[2]。

- [IMine] Search Task and Intent Mining
- [Math-2] Mathematical Information Access
- [MedNLP-2] Medical Natural Language Processing
- [MobileClick] Mobile Information Access
- [RITE-VAL] Recognizing Inference in Texts and Validation
- [SpokenQuery&Doc] Spoken Query and Spoken Document Retrieval
- [QALab] QA Lab for Entrance Exam
- [Temporalia] Temporal Information Access

このうち，最後の QALab と Temporalia は「パイロットタスク」である（他は「コアタスク」）。

この一覧が示すように，情報検索や自然言語処理，自然言語理解，音声認識，質問応答，推論などの諸技術を中核とする，多様なタスクを今回も設定することができた。また，処理対象となるメディアについても，時代を反映して，ニュース記事，ウェブ文書，発話，医学関連テキスト，入試問題など多岐に渡っている。

3 NTCIR-11 の日程

キックオフイベントはすでに 2013 年 9 月 2 日に NII にて開催済みであるが，その後のおおよその日程は表 1 のとおりである。参加登録は今年度末までに締め切られ，本格的な実験は，2014 年 1 月から開始される予定である。この実験に関する日程は，タスクによって，異なる可能性がある。

表 1: NTCIR-11 の主要日程（予定）.

1	2013.09.02	キックオフイベント
2	2013.12.30	タスク参加登録締切
3	2014.01.05	文書データ配布開始
4	2014.01-05	予備テスト（dry run）
5	2014.03-07	本テスト（formal run）
6	2014.08.01	評価結果返送
7	2014.08.01	タスク概要報告一部公開
8	2014.09.01	成果報告会用論文提出締切
9	2014.11.01	成果報告会用論文最終原稿締切
10	2014.12.09-12	NTCIR-11 成果報告会（東京）

注：4，5，6 はタスクによって異なる点に注意。

最終的な成果報告会は，2014 年 12 月 9 日から 4 日間，東京にて開催される予定である。この成果報告会には，参加者は必ず論文を投稿することになっており，その最初の提出は 9 月 1 日と早めに設定されている。これは，TO がタスクを総括し，まとめるのに各参加者の論文を必要とするためであり，実際にプロシーディングとして掲載される最終稿の提出は，11 月 1 日である。

なお，例年，NTCIR の成果報告会に付随して，Evaluating Information Access（EVIA）ワークショップが開催されている[3]。これは，情報アクセス技術の評価方法に関する国際ワークショップであり（査読付き），NTCIR などの評価実験の遂行に対して重要な役割を果たしている。NTCIR-11 にも，この EVIA を併設する方向で検討が進んでいる。

4 むすび

インターネットの発展に伴い，様々な形式・様態の文書データを高度に処理することが一層重要性を増している。このような状況において，基礎・応用の両面に渡って，その技術を向上させていくには，多くの研究者が協調して問題に取り組むことが不可欠である。NTCIR がそのひとつの基盤となり，研究者と社会の両方に恩恵をもたらすことが望まれる。関心のある方はぜひ，NTCIR のウェブサイト[4] を訪れていただくようお願いいたします（過去のテストコレクションの利用も可能です）。

[1]本稿の著者である岸田と上保が PC Co-chairs を務めている。また神門は General Co-chairs の一人である。

[2]各タスクの詳細は http://research.nii.ac.jp/ntcir/ntcir-11/tasks.html を参照。

[3]前回については，http://research.nii.ac.jp/ntcir/evia2013/ を参照。

[4]http://research.nii.ac.jp/ntcir/index-ja.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5回)
SIG-AM-05-10

検索意図とタスクマイニング: NTCIR-11 IMineタスクへの誘い

Intent and Search Task Mining: Invitation to the NTCIR-11 IMine Task

山本岳洋[1*] Yiqun Liu[2] Ruihua Song[3] Min Zhang[2]
Zhicheng Dou[3] 加藤誠[1] 大島裕明[1] Ke Zhou[4]

[1] 京都大学
[1] Kyoto Unviersity
[2] 清華大学
[2] Tsinghua University
[3] マイクロソフトリサーチアジア
[3] Microsoft Research Asia
[4] エディンバラ大学
[4] University of Edinburgh

Abstract: This paper introduces the NTCIR-11 IMine task, whose aim is to explore and evaluate technologies of mining and satisfying different user intents behind a Web search query. The NTCIR-11 IMine task comprises three subtasks: (1) Subtopic Mining, (2) Document Ranking, and (3) Search Task Mining. We briefly introduce these subtasks and how to participate in them. Japanese researchers in information retrieval, natural language processing, interactive information access, or machine learning are highly welcome to participate in the IMine task to accelerate their research.

1 はじめに

NTCIR [1]（NII Testbeds and Community for Information access Research）は国立情報学研究所が主催する，情報アクセスに関する評価型ワークショップである．NTCIR は 1999 年にスタートし，米国の TREC [2]（Text REtrieval Conference）や欧州の CREF [3]（Cross-Language Evaluation Forum）などと同様に，複数の参加者が，同じ目的，テストコレクションのもと，互いに技術を競い合う，もしくは協力することを通じて情報アクセス技術を進歩させることを目的とした国際的なワークショップである．2013 年 6 月に開催された NTCIR-10 では，7 つのタスクが開催され，13 の国と地域から計 103 チームが参加した [3]．NTCIR は 1 年半の周期で開催されており，次回の NTCIR-11 は 2014 年 12 月 9 日から 12 日かけて開催される．

我々は NTCIR-11 にむけて，**IMine**（Intent and Search Task Mining）というタスクをオーガナイズしている．IMine タスクは，NTCIR-9 および NTCIR-10 でそれぞれオーガナイズされた，INTENT [6] と INTENT-2 [4, 5] タスクがその前身となっており，INTENT-2 では，日本，中国，韓国，フランスの 4 カ国から計 11 チームの参加があった．INTENT，INTENT-2，そして IMine タスクの目的は，ウェブ検索ユーザの入力したクエリから，その背後にある**多様な検索意図**を発見し，異なる検索意図を持ったユーザを満足させる方法，およびその適切な評価手法を確立することである．

本稿では，2014 年 12 月にかけて開催される IMine タスクの概要とその参加方法について述べる．これまでの INTENT および INTENT-2 では，日本からの参加チームが多いとはいえなかった．そこで，日本における，情報検索，対話的情報アクセス，自然言語処理，機械学習など，さまざまな分野の研究者の方々に，本稿をご参考にしていただき，是非 IMine タスクに参加していただきたい．

*連絡先：京都大学大学院情報学研究科
〒 606-8501 京都府京都市左京区吉田本町
E-mail: tyamamot@dl.kuis.kyoto-u.ac.jp
[1]http://research.nii.ac.jp/ntcir/index-ja.html
[2]http://trec.nist.gov/
[3]http://www.clef-initiative.eu/

2 IMine: 検索意図とタスクマイニングに関するタスク

2.1 タスクの概要

ウェブ検索エンジンを利用するユーザの検索意図は多様であるにもかかわらず，検索エンジンに投入されるクエリは短かったり，曖昧であったりすることが多い [7]．そのため，クエリがユーザの情報要求を必要十分な形で表現できていることは稀である．たとえば,「ハリーポッター」というクエリは，ハリーポッターの映画に関する情報を求めている場合もあれば，ハリーポッターの書籍，あるいは主人公に関する情報を求めている場合もあり，適合する文書はユーザの検索意図によって異なる．

このような，異なる検索意図を持ったユーザを満足させるための技術として，情報検索の分野では近年，検索結果多様化（search result diversification）に関する研究が盛んである [1]．検索結果多様化に関する研究で重要となる技術は，与えられたクエリ（トピック）に対して，**どのような検索意図が存在するのかを発見**し，それにもとづき，**多様性を考慮しながら文書をランキングする**ことで，上位の検索結果だけでできるだけ多くのユーザを満足させることである．

今回行う IMine タスクでは，以下の 3 つのサブタスクを設けている．

- サブトピックマイニング
- ドキュメントランキング
- 検索タスクマイニング

このうち，サブトピックマイニングとドキュメントランキングサブタスクに関しては，INTENT および INTENT-2 でも開催していたサブタスクであり，検索タスクマイニングサブタスクは今回 IMine で初めて取り組むサブタスクである．IMine タスクでは，表 1 に示すように日・英・中の 3 つの言語でのタスクを用意しており，参加者はどの言語のサブタスクでも参加できる．

表 1: IMine で実施されるサブタスク

	実施される言語		
サブタスク	**日本語**	**英語**	**中国語**
サブトピックマイニング	x	x	x
ドキュメントランキング		x	x
検索タスクマイニング	x		

2.2 サブトピックマイニング

サブトピックマイニング（Subtopic Mining）サブタスクは，与えられたクエリ（**トピック**）に対し，考えられる意図を文字列で表現したもの（**サブトピック**）を重要度でランクづけして出力するタスクである．たとえば,「ハリーポッター」というクエリに対し,「ハリーポッター 本」,「ハリーポッター　映画」,「ハリーポッター 主人公」といったサブトピックが出力として考えられる．

どのような情報源を用い，どのようなアプローチでサブトピックを抽出するかは参加者に委ねられている．たとえば，これまで行ってきた INTENT および INTENT-2 タスクでは，複数の検索エンジンのクエリ推薦や Wikipedia の曖昧性解消ページを用いて単語をクラスタリングすることで，重要なサブトピックを発見する手法が良い成績を収めている．また，我々オーガナイザも，参加者に対して，商用検索エンジンのクリックスルーデータ（中国語）やクエリ推薦（日本語，中国語，英語）を情報源として提供することを予定している．

今回 IMine タスクで新しく取り組む試みとして，これまで，システムはサブトピックを一次元のリストとして出力することが求められていたのに対して，IMine タスクでは，2 次元の**階層的**なリストを出力することを予定している．サブトピックを階層的に抽出することで，多様な検索意図をより詳細に理解することができると考えている．

2.3 ドキュメントランキング

ドキュメントランキング（Document Ranking）サブタスクは，TREC のタスクと同様，ウェブ検索結果を多様化するタスクである．このサブタスクでは，サブトピックマイニングと同一のクエリに対して，文書を適合性と多様性に基づいてランクづけして出力することが求められる．この際，ランキング対象となる文書集合として，参加者に対してあらかじめクロールされたウェブページのデータセットが提供される．そのため，参加者自らがウェブページをクローリングする必要はなく，与えられた文書集合をランキングするだけで良い．

これまで行ってきた INTENT および INTENT-2 タスクでは，クエリ推薦やウェブページのドメイン，アンカーテキストなどの異種情報源を統合して検索結果を多様化する手法 [2] やクエリの意図を考慮し，選択的に検索結果の多様化を行う手法 [8] などが好成績を収めている．

サブトピックランキングおよびドキュメントランキングサブタスクの，各参加チームの具体的なアプローチについては，NTCIR-10 のオンライン論文集[4]より閲覧できる．

2.4 検索タスクマイニング

検索タスクマイニング（Search Task Mining）サブタスクは，IMine で今回新しく取り組むサブタスクである．本サブタスクは，2.2 節で述べたサブトピックマイニングと類似のタスクである．サブトピックマイニングが，与えられたトピックのサブトピックを抽出するのに対して，検索タスクマイニングでは，与えられたクエリに対し，クエリの背後にある目的を達成するために必要な行動（**タスク**）を，文字列で表現したものを出力するタスクである．たとえば,「花粉症対策」というクエリに対し,「マスクをつける」,「アレロック錠剤を摂取する」,「レーザ治療を受ける」,「花粉症に効くツボを押す」といった出力が考えられる．

検索タスクマイニングでは，サブトピックマイニングが重要度でランクづけされたリストのみを考慮するのに対して，検索タスクマイニングでは，重要度だけでなく，得られたタスク間の順序関係についても出力する必要がある．たとえば,「引っ越し手続きをする」というクエリに対しては,「転入届けを提出する」,「転出届けを提出する」などのタスクが出力として考えられるが，このとき,「転入届けを提出する」というタスクは「転出届けを提出する」というタスクを実行する前に実行する必要がある．

このように，サブトピックマイニングが検索意図の「トピック」に注目していたのに対して，検索タスクマイニングでは，検索ユーザの「行動」に着目したタスクである．なお，本サブタスクは京都大学の 3 名によりオーガナイズされ，日本語で実施される予定である．

2.5 IMine@NTCIR-11 にご参加ください

IMine タスクに参加するためには，まず，2013 年 12 月 20 日までに，NII の NTCIR-11 のホームページ上から参加登録をする必要がある．その後は，表 2 に示した予定で，参加者はシステムの作成，結果の出力，論文執筆，会議への参加などを進めることとなる．また，検索タスクマイニングサブタスクに関しては，一部スケジュールが異なっており，表 3 に従って進められる予定である．IMine および検索タスクマイニングサブタスクに関する，より詳細な情報は，それぞれのウェブページ [5][6]にアクセスいただきたい．

表 2: IMine のスケジュール（2013 年 10 月時のものです）.

2013 年 12 月 20 日	NTCIR-11 への参加登録
2014 年 1 月	評価用クエリを参加者に配布
2014 年 3 月	システム出力結果提出
2014 年 8 月	評価結果を参加者に配布
2014 年 9 月	論文初稿締め切り
2014 年 10 月	論文最終稿締め切り
2014 年 12 月 9 日〜12 日	NTCIR-11@NII

表 3: 検索タスクマイニングサブタスクのスケジュール（2013 年 10 月時のものです）.

2013 年 12 月 20 日	NTCIR-11 への参加登録
2014 年 2 月	サンプルクエリを参加者に配布
2014 年 3 月	評価用クエリを参加者に配布
2014 年 4 月	システム出力結果提出
2014 年 8 月	評価結果を参加者に配布
2014 年 9 月	論文初稿締め切り
2014 年 10 月	論文最終稿締め切り
2014 年 12 月 9 日〜12 日	NTCIR-11@NII

IMine タスクでは，日本語で実施されるサブタスクはサブトピックマイングと検索タスクマイングの 2 種類であるが，両サブタスクとも大規模なクロールやインデックスを必要としないため，手軽に参加していただけると考えている．情報検索，ウェブマイニング，対話的情報アクセス，自然言語処理や機械学習といった，さまざまな分野の研究者に是非参加していただき，情報アクセスに関する研究を，ともに協力して進めていきたい．

3 むすび

本稿では，NTCIR-11 で行われる IMine タスクについて紹介した．さまざまな分野の研究者，特に日本の研究者の参加には是非本タスクに参加していただき，研究に役立てていただきたい．

謝辞

INTENT，INTENT-2 の各オーガナイザおよび参加チーム，NTCIR の各チェア・事務局の日頃のご尽力に感謝いたします．

[4] http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/NTCIR/toc_ntcir.html

[5] IMine タスク, http://www.thuir.org/IMine/

[6] 検索タスクマイニングサブタスク, http://www.dl.kuis.kyoto-u.ac.jp/ntcir-11/taskmine/

参考文献

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14, 2009.

[2] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 475–484, 2011.

[3] H. Joho and T. Sakai. Overview of NTCIR-10. In *Proceedings of NTCIR-10*, 2013.

[4] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, M. P. Kato, R. Song, and M. Iwata. Summary of the NTCIR-10 INTENT-2 task: Subtopic mining and search result diversification. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 761–764, 2013.

[5] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the NTCIR-10 INTENT-2 task. In *Proceedings of NTCIR-10*, 2013.

[6] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 INTENT task. In *Proceedings of NTCIR-9*, 2011.

[7] J. Teevan, S. Dumais, and E. Horvitz. Potential for personalization. *ACM Transactions on Computer-Human Interaction*, 17(1):4:1–4:31, 2010.

[8] Y. Xue, F. Chen, A. Damien, C. Luo, X. Li, S. Huo, M. Zhang, Y. Liu, and S. Ma. THUIR at NTCIR-10 INTENT-2 task. In *Proceedings of NTCIR-10*, 2013.

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第5回)
SIG-AM-05-11

数式検索タスク NTCIR-11 Math-2

Math Retrieval Task : NTCIR-11 Math-2

Akiko Aizawa[1*]　Michael Kohlhase [2]　Iadh Ounis [3]

[1] National Institute of Informatics
[2] Jacobs University Bremen
[3] University of Glasgow

Abstract: NTCIR-11 Math-2 aims at promoting researches in mathematical content access. Based on the past experience in NTCIR-10 Math Pilot Task, we settled two major goals in our task: construction of a reusable test collection, and establishment of a research community in the field. In this paper, we introduce an outline of the task and briefly discuss possible challenges in mathematical content access.

1　はじめに

「数式」は、数や記号、およびそれらの関係を表すための記法であり、科学に不可欠な知識表現法である。数式は、演算や推論を必要とするあらゆる分野で用いられるが、その普遍性にもかかわらず、現在の検索システムでは、数式の構造や意味を適切に扱うことは困難である。数式の検索を可能にすることは、単に検索システムで扱える対象を拡大するというだけではなく、数学の知識基盤構築に向けた第一歩ともなる。

NTCIR-11 Math-2 は、数式の検索手法を研究・開発するための評価基盤の構築を通して、数式検索の研究に貢献し、研究コミュニティを活性化することを目的としている。数式検索は従来から、数学電子図書館や計算機による数学知識処理などの分野で検討されてきた [2]。また、木構造を扱うことから、データベース分野からのアプローチもされている。しかしながら、情報検索の観点からのタスク設計やの評価の試みはこれまでほとんどなかった。Math-2 では、評価基盤を共有することで、情報検索や自然言語解析の先端的な手法の適用を可能にし、数学知識アクセスの新たな展開を目指す。

本稿では以下、NTCIR-11 における Math-2 タスクの概要を紹介し、数式検索の研究課題を概観する。

2　Math-2 タスクの概要

数式検索とは、クエリまたは対象文書に数式を含む検索である。NTCIR-11 Math-2 では特に、クエリと検索対象の両方に数式を含む場合を想定してタスクを設計している。以下、Math-2 のタスクの概要について紹介する。

(1) タスクの目標

NTCIR-10 で行ったパイロットタスクの経験を踏まえ、Math-2 では、２つの大きな目標をタスク設計の指針としている。第一は、再利用可能なテストコレクションの構築である。具体的には、なるべく多様な検索手法を導入するとともに、適合性判定を複数の評価者により行うことで、タスク終了後も活用できる資源の構築を目指す。第二は、数式検索の研究コミュニティの形成と支援である。具体的には、数式検索で検討するべきタスク設計や評価手法について、数学の電子図書館や数学知識処理を専門とする研究者と、情報検索や自然言語処理を専門とする研究者が意見交換をする場を提供する。

(2) 検索対象文書

Math-2 では、MathML 形式の数式を含む英語文書を検索対象とする。MathML は、数式を計算機で処理するための表記法として、World Wide Web コンソーシアムが勧告する XML アプリケーションで [3]、数式の標準記法として広く普及しつつある。既存の数式検索システムの多くは、MathML 形式の数式を想定して設計されている。

Math-2 の検索対象文書は、LaTeX ソースの形で公開されている ArXiv.org [4] の論文から選んだ 100,000 文書である。これらの latex ソースは、arXMLiv プロジェクト [5] において MathML を含む XML 形式に機械的に変換されており、NTCIR Math の参加者は arXMLiv プロジェクトから変換済の文書を入手可能である。これらの文書中には、約 35M（１論文あたり平均 350）

*連絡先：国立情報学研究所コンテンツ科学研究系
〒 101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: aizawa@nii.ac.jp

個の数式が含まれている。ただし、この場合の「数式」は、LaTeX ソース中の数式モードの文字列を MathML 形式に機械的に変換して ID を付与したもので、変換に伴うノイズを含んでいる。オリジナルの LaTeX 表記も併記されているので、LaTeX ベースの検索システムでの参加も可能である。

なお、Math-2 の検索対象文書は、NTCIR-10 の Math パイロットタスクで用いたものと同じである。ただし、NTCIR-10 では検索の単位を「数式」に限定していたのに対して、今回のタスクでは、通常の情報検索システムによる参加を容易にするため、論文をセクションなど最小の検索単位に分割して、テキストを含む検索単位のランキング結果を評価する枠組みを検討予定である。

(3) トピック

数式検索は情報検索分野では新しい課題であることから、Math-2 では伝統的な Ad hoc 検索タスクのスタイルを踏襲している。すなわち、文書集合の中から与えられたトピックに適合する文書を検索し、スコア順にソートした検索結果をプーリングして人手による適合性判定を行う。

Math-2 のトピックは、高度な検索や信頼性の高い適合性判定を可能にするための詳細説明を含み、以下の形式をとる予定である。トピックの数は 50 を予定している。また、すべてのトピックは、文書コレクション中に複数の適合文書を持つものとする。

表 1: トピックの例.

Topic ID	トピックの識別子
Query	数式とキーワードで表されるクエリ
Description	ユーザの検索要求の記述
Narrative	適合性判定の手がかりとするべき、ユーザの検索状況や検索意図に関する詳細な説明

クエリは数式とキーワードの両方を含み、数式については、MathML および LaTeX 形式で表現されている[1]。NTCIR-10 の Math Pilot Task におけるクエリの例を以下に示す。

表 2: クエリの例.

Math	$\sum_{n=1}^{\infty} \frac{sin(n)}{n}$
Text	infinite series conditionally convergent

[1]NTCIR-10 では、数式中で任意の記号とマッチングできる「ワイルドカード」変数を記述できるよう、タスク固有の拡張を行っている

(4) 評価

Math-2 では、Query フィールドによる検索のランキング結果の提出が必須である。また、オプションとして、人手により生成したクエリに基づく結果の提出なども歓迎する。提出結果には、根拠となる数式 ID などの情報を追加可能である。

オーガナイザによる評価では、参加システムからのランキング結果をプーリングし、上位 100 件について、2 名の判定者が適合性判定を行う。判定は、*Relevant*（適合）、*Partially relevant*（部分適合）、 *Non relevant*（不適合）、*Cannot be assessed*（判定不能）の 4 通りで行う。

3 数式検索における研究課題

数式検索には大きく分けて次の 2 つの課題がある。1 つは数式が持つ構造情報の扱いであり、もう 1 つは数式の意味の考慮である。まず前者について、数式は明示的な木構造を持つため、類似数式の検索に適した木構造の索引付けや、クエリが変数を含む場合への対応などの検討が必要となる。また後者について、数式の意味を考慮するために、文書のドメインや数式周辺のテキストなどの検索コンテクストを、いかに獲得して利用するかの検討が必要となる。

(1) 数式の構造情報の扱い

文書中の「数式」は、改行を伴う関係式だけではなく、行の中に埋め込まれた変数記号なども含む。MathML において数式は木構造の形で表現されることから、数式検索における検索対象は、大規模かつ不均一な木構造データ集合となる。また、数式の木構造表現と数式の持つ意味は 1 対 1 対応とはならないので、木構造データを検索する際には、あいまいさを許すマッチングが必要となる。

上記を踏まえると、数式検索システムの設計では、以下の課題を考慮することが必要である。第一は、数千万規模の木構造データを検索するための索引づけ手法の検討である。効率的な検索のために、ルートからのパスや部分木などの部分構造を索引として用いる実装が考えられるが、これらの部分構造に対する重みづけや、ランキングのためのスコア関数、Top-k 検索などの効率的な検索手法を検討する必要がある。

第二は、数式の特性を考慮した木構造データの扱いである。たとえば、子ノード間の順番を考慮するかどうかなどを、検索効率や性能を踏まえて考える必要がある。また、数式中の変数は他の任意の記号と置き換え可能な場合が多い。このような変数の扱いは、現状では検索システムによって様々であり、評価タスクを通して必要性や有効性を確認することが重要である。一方で、

クエリの中では一般に、どの変数が書き換え可能であるかは明示されないので、変数と解釈するべき記号をどのようにして特定するかという検索意図の推測も課題として残る。さらに、より一般的に、数式は意味を変えずに書き換えることが可能であるが（「$(-1) \times x$」と「$-x$」など)、どの範囲の書き換えを許すのかは検索システムの判断に委ねられているため、タスクを通した検証が必要である。

数式検索の第三の課題は、テキストと数式の統合である。クエリや文書には数式以外の通常のテキストも含まれるので、木構造検索の結果と通常テキストの検索結果をどのように統合してスコアを計算するか、数式とテキストの意味的な対応づけをどのように獲得して検索に利用するかなどを検討する必要がある。

(2) 数式の意味の考慮

数式を使えば、ものごとの厳密で形式的な記述が可能であると思われがちであるが、実際には文書中に出現する数式の解釈には、多くのあいまい性が存在する。たとえば「w^2」は「w の 2 乗」と読むのが通常であるが、実は「w^2」が表現しているのは、2つの文字「x」と「2」の間に存在する特定の位置とサイズの関係に過ぎない[2]。また、「$f(x)$」の「f」は関数であるが「$1/f$ ゆらぎ」の「f」は周波数である、「$y = ax^2$」は x の二次方程式であるが「$E = mc^2$」を二次方程式と呼ぶ人は少ないなど、解釈に必要な知識も多様である。

このように、数式それ自体は抽象的な構造の表現に過ぎず、文脈により与えられる解釈なしには、検索において必要となる同一性や類似度を定義することが困難である。具体的な数式解釈の手法としては、(1) 数式のレイアウト構造の意味構造への変換、(2) 関数や変数の型、束縛変数のスコープの決定、(3) 物理的な単位系や概念クラスへの対応づけ、などの意味処理が考えられるが、多様なアプローチが想定され、現時点では検討がはじまったばかりである。

4 おわりに

本稿では NTCIR-11 における Math-2 タスクの概要について紹介した。数式検索は、数式という独特の構造を扱うものであるが、検索システムの中心になるのは汎用的な木構造検索およびテキスト言語解析の技術である。実装上は、汎用の検索エンジンにテキストと数式の両方を索引付して読み込むことで、手軽にベースライン・システムを構築することが可能である。このようなベースライン・システムを出発点として、数学知識の抽出・体系化と利用、大規模 XML 木構造の検索、深い言語解析に基づく意味解析、テキストと非テキスト要素の統合検索、などの幅広い課題を検討することができる。

タスクの詳細設計やオプショナルなサブタスクについては現在検討中であり、参加者からのコメントやフィードバックを歓迎している。また、スケジュール等を含めた詳細については、順次、タスクのメーリングリストやウェブサイト等でアナウンス予定である。参加問合せや最新情報については以下を参考にして頂ければ幸いである。

表 3: NTCIR-11 Math-2 関連情報.

Organizers ML:	ntcadm-math@nii.ac.jp
Community ML:	ntcir-math@nii.ac.jp
Task Webpage:	http://ntcir-math.nii.ac.jp/
Community Site:	http://ntcir.mathweb.org/

謝辞

本研究の一部は、科学研究費補助金基盤研究 (B) 課題番号 24300062 の助成を受けた。

参考文献

[1] A. Aizawa, M. Kohlhase, and I. Ounis: “ NTCIR-10 Math Pilot Task Overview, ” Proceedings of the 10th NTCIR Conference (2013).

[2] J. Carette, et. al. (Eds.): Intelligent Computer Mathematics - MKM, Calculemus, DML, and Systems and Projects 2013, Part of CICM 2013 Proceedings, Lecture Notes in Computer Science, Springer (2013).

[3] W3C: Mathematical Markup Language (MathML) Version 3.0, Word Wide Web Consortium Recommendation 21 October 2010, http://www.w3.org/TR/MathML3/ .

[4] ArXiv.org e-Print archive: http://arxiv.org/ .

[5] H. Stamerjohanns, M. Kohlhase, D. Ginev, C. David and B. Miller: ”Transforming large collections of scientific publications to XML,” Mathematics in Computer Science. 3(3):299-307 (2010) (Also, see https://trac.kwarc.info/arXMLiv/).

[2]「w^i」とあれば、「i」を添え字だと考える人も多いのではないだろうか

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5回)
SIG-AM-05-12

NTCIR MedNLP-2: 医療分野の言語処理

NTCIR MedNLP-2: Natural Language Processing for Medixal Field

荒牧英治 [1, 2]　大熊智子 [3]　狩野芳伸 [2]　森田瑞樹 [4]

Eiji ARAMAKI [1, 2]　Tomoko OHKUMA [3]　Yoshinobu KANO [2]　Mizuki MORITA [4]

[1] 京都大学デザイン学ユニット
[1]Design Unit Kyoto University
[2] 科学技術振興機構 さきがけ
[2] JST PRESTO
[3] 富士ゼロックス（株）
[3] Fuji Xerox Co., Ltd.
[4] 東京大学
[4] The University of Tokyo

Abstract: Recently, medical records are increasingly written on electronic media instead of on paper, thereby increasing the importance of information processing in medical fields. We have organized an NTCIR-11 MedNLP task for medical records. Our pilot task, MedNLP, comprises three tasks: (1) term extraction for complaint and diagnosis, and (2) term normalization. These tasks represent elemental technologies used to develop computational systems supporting widely diverse medical services. This report presents the objective and data of this shared task.

はじめに

近年，急速に紙カルテから電子カルテへの移行がはじまっている．新規に開業する診療所で電子カルテの導入率は 7〜8 割に上るとされ[1]，今後さらに電子カルテの普及が進んでいくことは確実である．カルテの電子化に伴い，紙カルテの時代には事実上不可能であった大規模な医療情報の利活用が進むと期待されている．この活用先として，例えば次のような例が想定されている[2]:

- 新たな医学的知見の抽出
- 診療行為の結果評価
- 類似症例の検索
- まれな副作用や疾患の頻度の正確な把握

これらの実現のためには，病名や医薬品名などの専門用語やそれに対応するコードの標準化，データ記録形式の標準化などが必須となり，それを効率よく行うために言語処理技術の利用が注目されている．このため, 欧米では 1960 年代から医療分野の言語処理研究が盛んになっている．しかし，同時に他の分野と比べて進歩が遅いことも指摘されている[3]．その理由として，主なものは次の 3 つが指摘されている：

- 入手できるコーパスが不足している
- アノテーション済みのコーパスはさらに不足している
- アノテーションされる場合もその方針が統一されていない

我が国においても, こうした問題意識は同様である．したがって，我が国の医療分野の言語処理を高度化するためには，日本語で書かれたアノテーション済みの文書を研究者が共有できる仕組みが望まれる．

以上のような問題意識から，私たちは研究利用が可能な日本語のアノテーション済み医療文書を構築し，これを用いた解析タスクと共に研究コミュニティに提供する活動を 2011 年より続けている. この活動により，解析技術の客観的な評価を行うことが可能となる．また，医療文書の解析技術の開発に興味はあるがコーパスを持っていない研究者や解析技術を適用する場を模索している企業を集めて産学連携のコミュニティを形成し，課題共有と技術の発展・向上も図ることもできる．

関連するシェアドタスク

さまざまな分野において，実験材料を共有して解析手法の評価を行う，ということが行われている．こうした催しの呼び方はいろいろであるが（sharedtask, contest，competitio など），ここでは「**シェアドタスク**」に統一する．シェアドタスクに参加するグループには実験材料が配られるため，研究者がまず直面する実験材料の入手という壁が取り払われる．また，複数のグループで同じ実験材料を共有することで，手法ごとの解析精度の特徴を評価や議論が可能となる．このような利点から，言語処理分野では，TREC[1]をはじめとして CoNLL[2]や CLEFInitiative[3]，国立情報学研究所による NTCIR[4]，生命科学分野の BioNLP-ST[5]，BioCreative[6]など多くのシェアドタスクが開かれている．

医療分野でも，医療分野の言語処理シェアドタスクとして 2006 年から米国国立衛生研究所(NIH;National Institutesof Health)の主導による i2b2[7]が開催されている．また，TREC は 2011 年から Medical Records Track を開始した．これらのシェアドタスクは英語圏の文書解析技術の向上に貢献している．日本語での医療文書のシェアドタスクはこれまでのところ MedNLP が唯一のものとなる．

MedNLP-2 の概要

タスクの概要

先に挙げた目的を達成するためには，我々は日本語の医療文書を用いた言語処理シェアドタスクを NTCIR-10[8]のパイロット・タスクとして開催した．本年度は以下の２つのタスクを課題とする．

- **病名・症状抽出タスク**：文章から症状や診断病名を抽出する．
- **病名・症状正規化タスク**：抽出された病名と症状に ICD-10 コードを付与する．

[1] http://trec.nist.gov/
[2] http://conll.cemantix.org/
[3] http://www.clef-initiative.eu/
[4] http://research.nii.ac.jp/ntcir/
[5] http://www.bionlp-st.org/
[6] http://www.biocreative.org/
[7] http://www.i2b2.org/
[8] http://mednlp.jp/medistj-ja/

コーパスの概要

本パイロット・タスクのためのコーパスとして，医師によって書かれた患者の病歴要約(medical history)を用意している．ただし，個人情報保護の観点から現実の要約をそのまま研究に利用することは困難である．そこで，疑似的な病歴要約を書き起こしたものを用いる．この際，作成にあたっては，医師免許を持った医師に依頼した．

アノテーションの概要

コーパスに対し，次の２つのタグを付与している

- <t>：日時(time)
- <c>：症状と診断表現(complaint&diagnosis)

さらに，症状に診断表現に関しては，次の２つの属性を付与している．

- **モダリティ**：その症状が実際にあったのか（positive)，または，認められなかったのか(例；胸水なし（negation)），または，家族歴としての症状なのか（family)，単なる疑いであったのか（suspicion）を区別する．
- **ICD コード（疾病分類コード）**： WHO による病名分類コード．アルファベット１桁と数字３桁からなる．病名正規化タスクは，この属性を出力するタスクとなる．

図１にアノテーションされたコーパス例を示す．

本シェアドタスクを達成するために実装されるツールは，様々な医療システムの構築に必要となる基礎的な要素技術である．たとえば，患者にいつ，どのような症状が実際に出現したのかを自動集計できるシステムが構築される．これは医療システムの検索機能を実現するための基盤となるシステムである．

こうした応用アプリケーションの構築，また要素技術のさらなる性能向上のためには，ツールが互換性や再利用性を考慮した形で利用可能であることが理想である．そこで，コーパスデータとその評価器に加え，再配布可能なツールを国際標準 UIMA[9]準拠のコンポーネントとして実装し，Kachako プラットフォーム[10]に統合して大規模処理も含め自動実行できるように配布することを予定している．

[9] http://uima.apache.org/
[10] http://kachako.org/

図 1: コーパス例.

□ **入力データ例**：原文

工場に勤めている64歳の男性。
2025年月8月2日（来院5日前）頃から腹痛が生じるとともに，食欲不振，嘔気・嘔吐出現した。
体幹は温かいが，末梢は湿潤冷汗でショック状態。
明らかな運動麻痺はみられず。
翌日，意識障害出現し，腎機能障害の増悪を認めて徐々に尿量低下し，8月9日18時10分に心肺停止。
8月9日21時44分死亡確認。

□ **出力データ例**：タグ付き

工場に勤めている64歳の男性。
2025年月8月2日（来院5日前）頃から<c icd="R104">腹痛</c>が生じるとともに，<c icd="R630">食欲不振</c>，<c icd="R11">嘔気</c>・<c icd="R11">嘔吐出現</c>した。
体幹は温かいが，末梢は<c icd="unk">湿潤冷汗</c>で<c icd="R579">ショック状態</c>。
明らかな<c icd="G839" modality="negation">運動麻痺</c>はみられず。
翌日，<c icd="R402">意識障害出現</c>し，<c icd="N289">腎機能障害</c>の増悪を認めて徐々に<c icd="unk">尿量低下</c>し，8月9日18時10分に<c icd="I469">心肺停止</c>。
8月9日21時44分<c icd="R99">死亡確認</c>。

おわりに

本稿では，医療分野の言語処理シェアドタスク MedNLP2 について述べた．他の分野と異なり，材料であるコーパスが入手困難である医療分野において，シェアドタスクは，産学連携をすすめるよい装置となる可能性がある．これをきっかけとして，我が国において医療分野の言語処理を担う研究者が増加することを信じている．
また，このような試みは継続的に開催をすることでコミュニティが形成され，さらに開発が促進される．今後の継続開催に向け，様々な方の協力を得ながら努力を続ける予定である．

謝辞

本研究の一部は JST 戦略的創造研究推進事業(さきがけタイプ)「情報環境と人」および科研費補助金(若手研究 A)による．本シェアドテスクの開催にご協力して頂いた NTCIR 事務局および医師，アノテーター，参加者の皆様に感謝いたします．

参考文献

1. 株式会社シード・プランニング, *2011-2012 年版電子カルテの市場動向調査*. 2012.
2. 大江和彦 and 今井健, *臨床医学知識処理を目指した医療オントロジー開発*, in *オントロジーの普及と応用* 2012. p. 131-148.
3. Chapman, W.W., et al., *Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions.* J Am Med Inform Assoc, 2011. **18**: p. 540-543.

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5回)
SIG-AM-05-13

モバイル「情報」検索に向けて: NTCIR-11 MobileClick タスクへの誘い

Towards Mobile "Information" Retrieval: Invitation to the NTCIR-11 MobileClick Task

加藤 誠 [1]　Matthew Ekstrand-Abueg[2]　Virgil Pavlu[2]　酒井 哲也 [3]　山本 岳洋 [1]　岩田 麻佑 [4]

[1] 京都大学　[2] Northeastern University　[3] 早稲田大学　[4] KDDI 株式会社

Abstract: The One Click Access Task (1CLICK) is one of the tasks of NTCIR that requires systems to return a concise multi-document summary of web pages in response to a query which is assumed to have been submitted in a mobile context. Our goal is to retrieve "information" rather than documents to directly and immediately satisfy a user's information need. We report the result of the second 1CLICK task in NTCIR-10 (1CLICK-2), and describe participants' approaches to discuss who can benefit from the participation in the 1CLICK task. Furthermore, we introduce the next round of the 1CLICK task called *MobileClick*, in which participants are required to submit a two-layered summarization suitable for mobile information access.

1　はじめに

NTCIR (NII Testbeds and Community for Information access Research) はアジア言語に焦点を当てた，1 年半ごとに開催される情報アクセスシステムの評価フォーラムである．本報告では，NTCIR の一タスクである，1CLICK (One Click Access) タスク [Sakai 11, Kato 13] を紹介する．1CLICK タスクは，与えられたクエリに対して簡潔な複数文書要約を検索結果として出力するタスクであり，特に，モバイルユーザの情報アクセスを想定したタスクとなっている．例えば，テレビを店で買う際にプラズマと液晶の違いを知りたくなったユーザが,「プラズマ 液晶 違い」というクエリを携帯端末で入力したとする．検索ボタンがクリックされた後，従来の検索システムであれば，順位付けされた文書のリストを出力する．もしユーザが事実に基づく解説や利用者の使用感など多角的な視点から違いを調べたい場合には，これらのリストを順に閲覧し満足するまで何度も検索結果をクリックをしなければならない．1CLICK タスクは，検索ボタンがクリックされた後，それ以上クリックをせずに，即座に必要な情報が得られるようになることを目的として提案された．従来の検索システムが提示する 10 件の URL リストの代わりに，1CLICK システムは複数の文書の重要な「情報」を要約してユーザに提示するのである．

参加システムは，iUnit と呼ばれる最小の情報単位をどれくらい多く含むか，より重要な iUnit を要約の最初の方に出力することができるか，また，いかにユーザが読む文章量を少なくできているか，という観点から評価される．表 1 はクエリ「北川景子」に対して用意された iUnit の例を示す．各 iUnit はその内容を示す semantics，semantics の最小のテキスト表現である vital string，そして，重要度を表す重み w によって構成されている．iUnit 間には含意関係がある場合があり，含意する iUnits の ID が entails 列に書かれている．例えば，iUnit I049 は I050 を含意している．

表 1: クエリ「北川景子」の iUnit 例.

ID	entails	semantics	vital string	w
I001		身長： 160cm	160cm	11
I004		1986 年生まれ	1986 年生まれ	18
I049	I050	2009 年明治大学卒	2009 年	15
I050		明治大学卒	明治大学卒	11

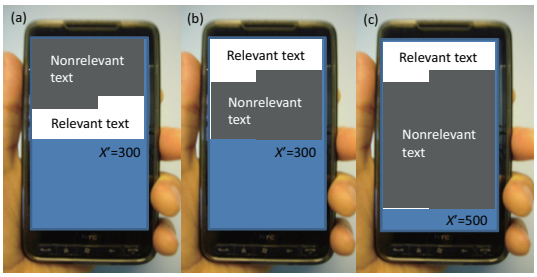


図 1: 1CLICK システムの評価.

2 回目の 1CLICK タスク (1CLICK-2) では，S-measure および T-measure，両評価指標を組み合わせた S♯-measure

という評価指標を利用することで，参加システムの性能を定量化した [Sakai 12]．S-measure はより重要な情報を要約の先頭に出力するシステムを高く評価する．例えば，図 1 のシステム出力のうち，S-measure は正解情報を先頭に出力する (b) を (a) よりも高く評価する．一方で，T-measure は正解情報の割合が高い結果を高く評価する指標であり，図 1 の例では (b) を (c) よりも高く評価する．S♯-measure は S-measure と T-measure の調和平均であり，1CLICK-2 ではこの指標を主たる評価指標としている．

本論文では，1CLICK-2 における取り組みと参加チームの提案システムを紹介することによって，1CLICK タスクが目指すモバイル「情報」検索を説明し，1CLICK タスクに参加するメリットについて述べる．また，1CLICK タスクを引き継ぎ，現在進行している MobileClick タスクについても紹介する．

2　1CLICK-2 の概要

1CLICK-2 タスクは 10 回目の NTCIR にて企画され，2012 年 8 月に本番クエリの公開，同 10 月に結果の提出，2013 年 6 月に NTCIR Conference にて結果の報告がされた．対象言語は日本語と英語であり，世界 5ヵ国から 10 チームの参加があった．参加チームは 1CLICK システムを構築してその優劣を競い，本論文の著者によって構成されるオーガナイザはタスクを定義し参加チームの評価を担当している．

1CLICK-2 のメインタスクは冒頭でも述べたとおり，与えられたクエリに対して複数文書を要約した結果 (日本語 : 500 文字もしくは 140 文字，英語 : 1000 文字もしくは 280 文字) を出力するというものである．参加者は 3 種類のランを選ぶことができ，それぞれ，Mandatory ラン（提出必須．配布された文書集合の中から要約を作成する），Oracle ラン（配布された文書集合とクエリへの適合文書 ID リストから予約を作成する），Open ラン（任意の文書集合から要約を作成できる）となっている．

参加者にはクエリと文書集合が配布され，各クエリに対して決められた文字数以内の要約を出力することが求められる．クエリは ARTIST（例 :「倉木麻衣」），ATHLETE（例 :「イチロー」），POLITICIAN（例 :「小池百合子 キャスター」），ACTOR（例 :「栗山千明 カーネーション」），FACILITY（例 :「京都真如堂」），GEO（例 :「京都市 スーパー銭湯」），DEFINITION（例 :「gps」），QA（例 :「なぜ猫はのどを鳴らすのか」）の 8 つのカテゴリのうちのいずれかに属しており，合計 100 クエリが参加者に配布された．

参加者が 1CLICK システムを構築する一方で，オーガナイザー側は，参加者に配布したものと同じ文書集合から，各クエリについて iUnit と呼ばれる適合情報の抽出を行った．iUnit の例は冒頭にて述べたとおり，表 1 に示されている．

図 2: 提出されたランと iUnit のマッチング.

参加者からランを受け取った後，オーガナイザによって 1CLICK システムが評価された．まず，提出されたランのどの部分が用意された iUnit に合致するのかを人手で判定した．この作業の様子を図 2 に示す．評価者は左側に表示されたランの内容から，右側に表示された iUnit に合致する部分をマウスで選択し，ランのどの箇所が各 iUnit に対応しているのかを簡単に記録することができるようになっている．

提出されたランと iUnit のマッチングの後，各システムは S-measure, T-measure, および，S♯-measure によって評価され，この結果は参加チームへと公開された．次節では参加チームが利用したアプローチについて簡単に紹介し，次々節では各システムの結果について触れる．

3　1CLICK-2 の参加チーム

本節では 1CLICK-2 のメインタスクに参加したチームがとったアプローチについて簡単に紹介する．詳細については NTCIR-10 会議論文集 [1] を参照して欲しい．

TTOKU (東京工業大学) チーム は 1CLICK タスクを要約問題として取り組んだ [Morita 13]．要約には Query SnowBall というクエリ指向要約手法が用いられ，元のテキストをそのまま抜粋するだけではなく文圧縮も行っている．MSRA (マイクロソフトリサーチアジア) チームは，クエリタイプごとに重要な属性を抽出し，それらの属性を元にして文をランキングすることで要約を作成している [Narita 13]．NSTDB (奈良先端科学技術大学院大学) チームは，配布された HTML 文書を XML 文書に変換することによって，1CLICK タスクを XML 要素検索の問題と捕らえ，他のチームとは違う視点から本タスクに取り組んでいる [Keyaki 13]．

[1] http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/NTCIR/toc_ntcir.html

ut (トゥウェンテ大学) チームは様々な API（Freebase TEXT API や Yahoo! GeoPlanet API やなど）を組み合わせることによって，各クエリタイプに特化した結果を出力している [Ionita 13]．udem (モントリオール大学) チームは質問応答技術に基づいた手法によって，異なるクエリタイプに対し統一的な方法をもって取り組んでいる [Duboue 13]．KUIDL (京都大学) チームは Web 文書から属性と属性値のペアを抽出し，それをランキングすることによって構造を意識した要約に挑戦している [Manabe 13]．

要約，属性抽出，XML 文書検索，マッシュアップ，質問応答，情報抽出と，上述の 6 チームのキーワードのみを拾ってみても，そのアプローチは多岐にわたっていることがわかる．したがって，1CLICK タスクは多様な興味を持つ研究者が参加できるタスクであると言える．ぜひ，少しでも関連のある研究をしている場合には，参加を検討していただきたい．

4 1CLICK-2 の結果

図 3 および図 4 は 1CLICK-2 日本語ランと英語ランの結果をそれぞれ表している．両グラフとも横軸には提出されたランが S♯-measure の高い順に並んでおり，縦軸は S♯-measure を表している．詳細は割愛するが，今回の 1CLICK タスクではベースライン（図中の ORG および BASELINE）が意外にも良い結果を出しており，1CLICK システムには改善の余地が大いにあると言える．ベースラインには Wikipedia の最初の文章や検索結果の上位数件のスニペットを結合したものが用いられており，Wikipedia にあるようなエンティティ名をクエリにした場合にはこれらの手法で十分に良い結果が得られている．一方で，あるエンティティに関する概要的な情報だけではなく，より特化した情報が正解となるようなクエリ（「中村紀洋 ホームラン」や「ロバート ケネディ キューバ」など）に対しては，単純なベースライン手法はそれほど有効でないことがわかっている．

いくつかのクエリタイプに対しては，参加チームのシステムがベースラインに勝っている．例えば，FACILITY クエリに対しては，KUIDL チームが最も良い結果を残しており，これは同チームの情報抽出に基づく手法が FACILITY クエリの iUnit（住所や電話番号など）を発見するのに適していたと考えられる．QA クエリでは，TTOKU チームのランが最も良い結果を残している．同チームの要約に基づく手法が特に質問応答タスクに似た QA クエリの設定に効果的であったと思われる．

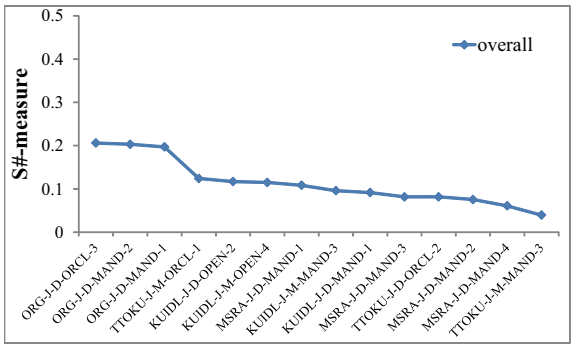


図 3: 1CLICK-2 日本語ランの結果

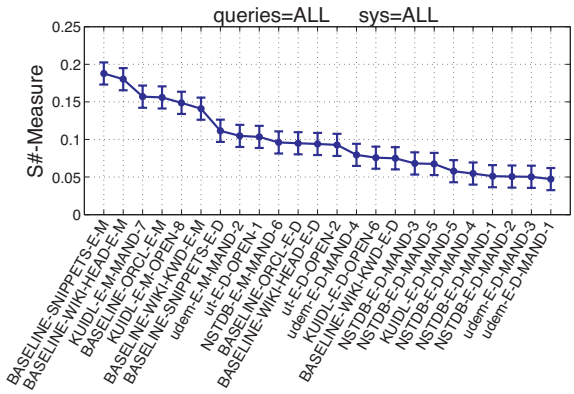


図 4: 1CLICK-2 英語ランの結果

5 MobileClick タスク

11 回目の NTCIR では，1CLICK タスクは MobileClick タスクとその名前を変えて継続される予定である．MobileClick タスクではよりモバイル環境の検索に焦点を当て，与えられたクエリに対して 2 層の要約を出力するタスクとなっている．図 5 に 2 層の要約の例を示す．MobileClick システムの第 1 層の出力は同図の中央のスクリーンショットのように，与えられたクエリの概要的な情報とより詳細な情報へのリンクを含んでいる．ユーザは表示されたリンクをクリックすることで，同図の右側にあるような画面へと遷移することができる．この出力は 1CLICK タスクよりも現実的な，また，たくさんの情報を一画面で見ることができないモバイルユーザにとって効果的な提示方法となっている．

一方で，このようなタスクは参加者にとってもより挑戦的な内容となっている．参加チームは取得した情報を第 1 層に表示するべきか，もしくは，第 2 層に表示するかの選択を行わなくてはならない．取得された情報がどのユーザにとっても有益（概要的）であれば第 1 層に，ある特定のユーザにとって有用であれば第 2 層に提示した方が，1CLICK タスクでの評価基準であった「即座に重要な情報」,「必要最低限の閲覧文章

図 5: MobileClick タスクの出力例

表 2: MobileClick のスケジュール.

2013 年	8 月	Web ページ公開
	10 月	サンプルクエリ・iUnit 公開
2014 年	3 月	テストクエリ公開
	4 月	ラン提出締切
	8 月	評価結果公開
	12 月	NTCIR-11 Conference

量」をより良く満たすことになる.

上述の変更点に加え，MobileClick タスクでは iUnit 検索サブタスク（重要な情報の断片を文書から抽出）と iUnit 要約サブタスク（与えられた重要な情報の断片から 2 層の要約を生成）の 2 つのサブタスクを用意しており，例えば，情報抽出にだけ興味がある参加者は前者のサブタスクのみに，要約にだけ興味がある参加チームは後者のサブタスクのみに参加することができるため，今回のタスクではより多くの方々に広く門戸を開いている.

6 まとめ

本論文では，1CLICK-2 における取り組みと参加チームの提案システムを紹介することによって，1CLICK タスクが目指すモバイル「情報」検索を説明し，1CLICK タスクに参加するメリットについて述べた．また，1CLICK タスクを引き継ぎ，現在進行している MobileClick タスクについても紹介した.

現在のところ，MobileClick のスケジュールは表 2 を予定している．情報検索，自然言語処理，データベースなどに関する研究をしている方は，一度 MobileClick のホームページ [2] を訪れていただきたい.

[2] http://www.dl.kuis.kyoto-u.ac.jp/ntcir-11/mobileclick/

謝辞

1CLICK-2 の参加チーム，および，NTCIR のチェア・事務局の皆様のご尽力に感謝申し上げる.

参考文献

[Duboue 13] Duboue, P., He, J., and Nie, J.-Y.: Hunter Gatherer: UdeM at 1Click-2, in *Proc. of the 10th NTCIR Conference*, pp. 233–236 (2013)

[Ionita 13] Ionita, D., Tax, N., and Hiemstra, D.: API-based Information Extraction System for NTCIR-1Click, in *Proc. of the 10th NTCIR Conference*, pp. 225–232 (2013)

[Kato 13] Kato, M. P., Ekstrand-Abueg, M., Pavlu, V., Sakai, T., Yamamoto, T., and Iwata, M.: Overview of the NTCIR-10 1CLICK-2 Task, in *Proc. of the 10th NTCIR Conference*, pp. 183–211 (2013)

[Keyaki 13] Keyaki, A., Miyazaki, J., and Hatano, K.: XML Element Retrieval@1CLICK-2, in *Proc. of the 10th NTCIR Conference*, pp. 237–242 (2013)

[Manabe 13] Manabe, T., Tsukuda, K., Umemoto, K., Shoji, Y., Kato, M. P., Yamamoto, T., Zhao, M., Yoon, S., Ohshima, H., and Tanaka, K.: Information Extraction based Approach for the NTCIR-10 1CLICK-2 Task, in *Proc. of the 10th NTCIR Conference*, pp. 243–249 (2013)

[Morita 13] Morita, H., Sasano, R., Takamura, H., and Okumura, M.: TTOKU Summarization Based Systems at NTCIR-10 1CLICK-2 task, in *Proc. of the 10th NTCIR Conference*, pp. 212–217 (2013)

[Narita 13] Narita, K., Sakai, T., Dou, Z., and Song, Y.-I.: MSRA at NTCIR-10 1CLICK-2, in *Proc. of the 10th NTCIR Conference*, pp. 218–224 (2013)

[Sakai 11] Sakai, T., Kato, M. P., and Song, Y.-I.: Overview of NTCIR-9 1CLICK, in *Proc. of NTCIR-9*, pp. 180–201 (2011)

[Sakai 12] Sakai, T. and Kato, M. P.: One Click One Revisited: Enhancing Evaluation Based on Information Units, in *Proc. of AIRS 2012*, pp. 39–51 (2012)

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5回)
SIG-AM-05-14

SpokenQuery&Doc: 自由発話音声クエリからの情報アクセス

SpokenQuery&Doc: Information Access from Spontaneously Spoken Query

秋葉友良 [1*]　西崎博光 [2]
南條浩輝 [3]　Gareth Jones[4]

[1] 豊橋技術科学大学
[1] Toyohashi University of Technology
[2] 山梨大学
[2] Yamanashi University
[3] 龍谷大学
[3] Ryukoku University
[4] Dublin City University

Abstract: This paper introduces the SpokenQuery&Doc task, which will be conducted in the next NTCIR evaluation. The SpokenQuery&Doc is a successor to the previous SpokenDoc and SpokenDoc-2 tasks evaluated at the past NTCIR workshops, which are also presented in this paper.

1 はじめに

国立情報学研究所が主催する情報アクセス技術の評価型ワークショップ NTCIR(NII Testbets and Communities for Information access Research) では、これまで様々な情報アクセスタスクの評価が行われてきたが、2011 年の NTCIR-9[7] より音声ドキュメント検索タスク SpokenDoc[1] が評価タスクとして採択され、最新の NTICIR-10[6] では 2 回目の評価タスク SpokenDoc-2[3] が実施された。

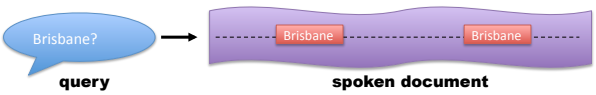
NTCIR-11 では、SpokenDoc の後継タスクとして SpokenQuery&Doc タスクを実施する。SpokenQuery&Doc では、これまでテキストで与えた検索クエリを自由発話音声で与えることに評価の焦点を当てる。一方、これまでのテキスト入力 STD および SCR タスクもサブタスクとして継続する。

2 NTCIR-9 SpokenDoc と NTCIR-10 SpokenDoc-2

多くの情報検索タスクが対象とするテキストと同様に、ラジオやテレビなどの放送や動画データなどに付随する音声にも豊富な言語情報が含まれている。動画配信サイト等を通して、音声を含むコンテンツは増加の一途を辿りつつあるが、その言語情報へのアクセスはテキストのように容易ではない。音声データを、その文書としての側面に注目して「音声ドキュメント」と呼び、音声データを対象とした検索を「音声ドキュメント検索」[9] と呼ぶ。

*連絡先：豊橋技術科学大学 情報・知能工学系
〒 441-8580 愛知県県豊橋市天伯町雲雀ヶ丘 1-1
E-mail: akiba@cs.tut.ac.jp

- 音声検索語検出(Spoken Term Detection)



- 音声内容検索(Spoken Content Retrieval)

図 1: STD タスクと SCR タスク

第 9 回 NTCIR で実施した “IR for Spoken Documents (SpokenDoc)” タスク [1, 2] は、NTCIR 初の音声ドキュメント検索タスクであり、講演音声を対象とした 2 つの音声ドキュメント検索タスクが設定された (図 1)。

Spoken Term Detection (STD) 語をクエリとして与え、音声ドキュメント中からクエリが現れる位置を特定するタスク。計算効率 (索引に必要な空

間コスト、検索時間コスト、など) と検索性能 (精度と再現率) の 2 つの観点から評価を行った。

Spoken Document Retrieval (SDR) 質問文で表現した比較的長いクエリを与え、クエリと関連する音声セグメントを見つけるタスク。テキストを対象とした検索における内容検索に相当する。音声セグメントとして、講演全体 (講演検索タスク) と、講演中の数秒程度の音声区間 (パッセージ検索タスク)、の 2 種類の粒度を設定した。

SCR における講演検索タスクは、テキストを対象とした検索における文書検索に相当する。しかし、音声ドキュメント検索では講演のような大きな単位が検索されたとしても、検索結果を確認するためには音声の再生が必要となり、テキストのように全体をざっと一覧することができない。したがって、よりピンポイントに検索の適合箇所 (音声区間) を見つける技術が必要となる。これがパッセージ検索タスクを設定した理由である。

続く NTCIR-10 で実施した SpokenDoc-2 タスク [3, 8, 10] では、上の 2 つのタスクに加えてクエリが音声ドキュメント中に出現しないことを確認する iSTD タスクの新規設定や、より低い認識率の環境下での音声ドキュメント検索の評価を行った。

音声ドキュメント検索の性能は音声認識の精度に依存するため、タスクオーガナイザはタスク参加グループ共通で利用できる音声認識結果を用意した。これにより、参加グループの検索結果を共通の土台の上で比較することが可能になるとともに、音声認識システムを保有しない参加グループや、音声認識よりも検索手法に興味を持つ参加グループに対し、参加を容易にする環境を提供した。

3 NTCIR-11 SpokenQuery&Doc

SpokenDoc および SpokenDoc-2 では、テキストで与えたクエリから、音声データを対象とした情報検索タスクの評価を行った。NTCIR-11 SpokenQuery&Doc では、音声を扱う検索のもう一つの側面として、クエリを音声で与える検索にも焦点を当てる。

タスク設定の背景には、テキスト入力による情報検索の限界を音声入力によって克服したいという展望がある [4]。人はしばしば複雑な情報要求を抱くが、現状の情報検索システムでは入力インタフェース (テキスト入力) の制限によりユーザの情報要求を十分に汲み取ることができない。一方、音声入力には，ユーザへの負荷が少ない、入力速度が速い，思考を即時に表現できる，といった利点があり、ユーザが音声で思い付くまま情報要求を表現することで、これをそのままシステムに入力する手段を提供する。一方、ユーザがその場で制限無く自由に発話した音声 (自発音声) は、あらかじめ発話内容を調整してから発話した読み上げ音声に比べて、話し言葉的特徴 (間投詞の頻出などの非流暢性、発話内容の非文法性、など) が多く現れるため、音声認識が難しい。図 3 に、予備実験 [11] で収集した自由発話音声クエリを示す。

> 今年の夏休みに始めて山に登山に行くことになったんですけれども、あの登山は結構事故があのやはり夏になるとよくニュースで聞きますし、まあ誰々が行方不明になったとか遭難したとかそういう話がいろいろあると思うんですけれども、あの山に登るときにはまあどういった心構えというか、あの装備こういう装備が必要だとかこういうものがあるといいよとかそういったあの山登りに関しての、山に登るときについての心構えについて知りたいです。

図 3: 自由発話音声クエリの例

NTCIR-11 SpokenQuery&Doc では、以下の 3 つのタスクを設定する。(図 2)

SQ-SCR 音声クエリからの音声内容検索 (Spoken-query-driven Spoken Content Retrieval) タスク。SpokenQuery&Doc のメインタスクである。自由発話音声で表現した比較的長いクエリをシステムへの入力として、音声内容検索 (SCR) を行うタスク。このタスク実現のためのサブタスクとして、次の 2 つのタスクを実施する。

SQ-STD 音声クエリからの音声検索語検出 (Spoken-query-driven Spoken Term Detection) タスク。自由発話音声クエリ中に現れる語の音声区間を入力として、音声検索語検出 (SCR) を行うタスク。音声クエリから語の音声区間の切り出し結果は、タスクオーガナイザから提供する。

STD-SCR 音声検索語検出結果からの音声内容検索 (Spoken Content Retrieval using STD results) タスク。SQ-STD タスク参加者による検出結果を入力として、音声内容検索 (SCR) を行うタスク。

SQ-SCR および SQ-STD タスクでは、音声クエリに加えて、人手で書き起したテキストクエリも提供する。タスク参加者は、このテキストクエリを使って結果を提出してもよい。すなわち、SpokenDoc、SpokenDoc-2 と同様の STD タスクおよび SCR タスクも実施する。また、音声から音声を検索するタスクでは、必ずしも音声認識を用いたテキスト表現を介した検索アプロー

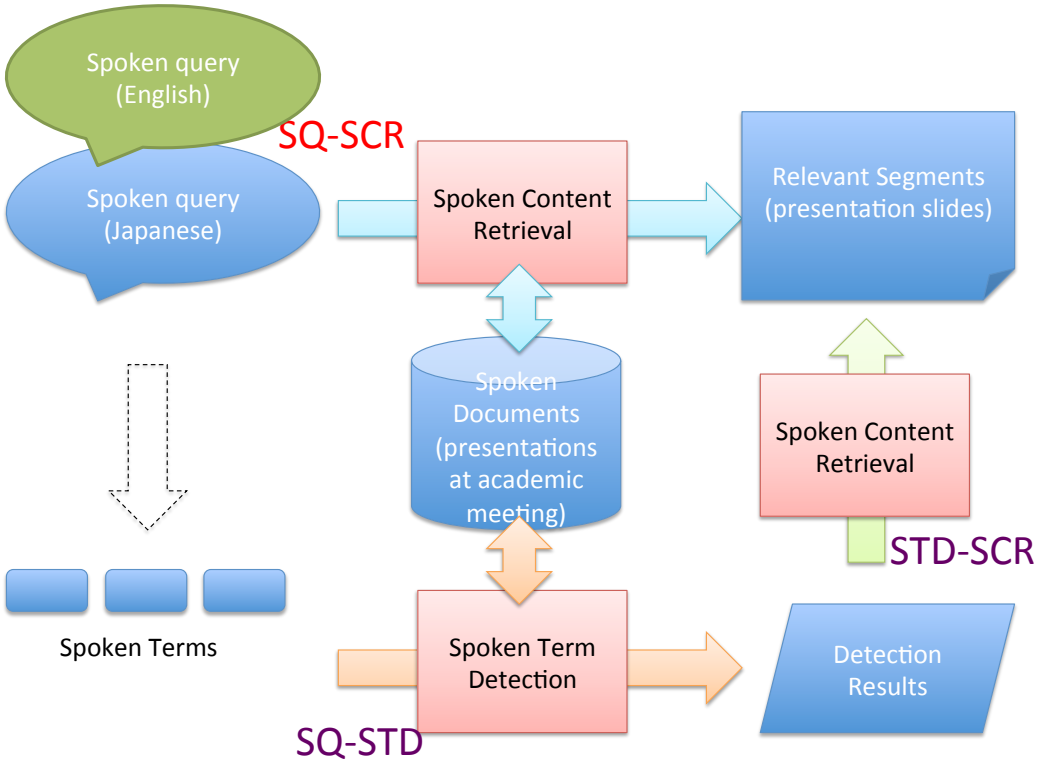


図 2: SpokenQuery&Doc で実施する 3 つのタスク

チを採用する必要はなく、音声を直接比較することも可能である [5]。タスクオーガナイザは、このようなアプローチによるタスク参加も期待している。

4 むすび

本稿では、過去 2 回の NTCIR で実施した SpokenDoc タスクと、次回 NTCIR-11 で実施する SpokenQuery&Doc タスクを紹介した。SpokenQuery&Doc のより詳しい情報は、以下の Web ページを参照されたい。

`http://www.nlp.cs.tut.ac.jp/ntcir11/`

参考文献

[1] T. Akiba, et al. Overview of the IR for spoken documents task in NTCIR-9 workshop. In *Proceedings of The Ninth NTCIR Workshop Meeting*, pp. 223–235, 2011.

[2] T. Akiba, et al. Designing an evaluation framework for spoken term detection and spoken document retrieval at the NTCIR-9 SpokenDoc task. In *Proceedings of International Conference on Language Resources and Evaluation*, 2012.

[3] T. Akiba, et al. Overview of the NTCIR-10 SpokenDoc-2 task. In *Proceedings of The 10th NTCIR Conference*, pp. 573–587, 2013.

[4] T. Akiba, A. Fujii, and K. Itou. Collecting spontatneously spoken queries for information retrieval. In *Proceedings of International Conference on Language Resources and Evaluation*, pp. 1439–1442, 2004.

[5] T. K. Chia, K. C. Sim, H. Li, and H. T. Ng. A lattice-based approach to query-by-example spoken document retrieval. In *Proceedings of Annual International ACM SIGIR Conference on Research and development in information retrieval*, pp. 363–370, 2008.

[6] H. Joho and T. Sakai. Overvew of NTCIR-10. In *Proceedings of The 10th NTCIR Conference*, pp. 1–7, 2013.

[7] T. Sakai and H. Joho. Overview of NTCIR-9. In *Proceedings of The Ninth NTCIR Workshop Meeting*, pp. 1–7, 2011.

[8] 西崎, 秋葉, 相川, 伊藤, 河原, 胡, 中川, 南條, 山下. Ntcir-10 spokendoc-2 spoken term detection タスクの結果と知見. 日本音響学会秋季研究発表会講演論文集, pp. 107–110, 2013.

[9] 秋葉. 音声ドキュメント検索: マルチメディアデータを対象とした音声言語情報検索. 情報の科学と技術, 63(1):21–27, 2013.

[10] 秋葉, 西崎, 相川, 伊藤, 河原, 胡, 中川, 南條, 山下. Ntcir-10 spokendoc-2 spoken content retrieval タスクの結果と知見. 日本音響学会秋季研究発表会講演論文集, pp. 111–114, 2013.

[11] 大島, 秋葉. 自発音声による情報要求の収集実験とその検索性能評価. 日本音響学会春季研究発表会講演論文集, pp. 233–234, 2013.