オノマトペが属する五感の推定

Which sense does an onomatopoeia belong to?

仲村哲明 [1*]　宮部真衣 [1]　荒牧英治 [1,2]
Tetsuaki Nakamura[1]　Mai Miyabe[1]　Eiji Aramaki[1,2]

[1] 京都大学 学際融合教育研究推進センター デザイン学ユニット
[1] Unit of Design, Center for the Promotion of Interdisciplinary Education and Research, Kyoto University
[2] 科学技術振興機構 さきがけ
[2] JST PRESTO

Abstract: This study aims to develop a system which visualizes subjective information. Focusing on onomatopoeias as such information, we estimate which senses an onomatopoeia belongs to among “touch”, “taste”, “smell”, “hearing”, “sight”, “pleasure (positive)” and “unpleasure (negative)”. For this purpose, we use a machine learning method (Support Vector Machine) which utilizes phonetic symbols and the number of occurrences of them in the onomatopoeia. Then, the experimental result for evaluation demonstrates that (1) the best performance is achieved for “hearing” and “sight”, and (2) the performance of the classifier is similar to that of human. Finally, we propose the system which creates city maps displaying distribution of subjective information for senses.

1　はじめに

世の中には膨大かつ多様な情報が存在し，それらを可視化するシステムが必要とされている．例えば，気候変動による被害状況の可視化システム [Houghton 12]，大気汚染の予測結果を可視化するシステム [菅田 11]，サイバー攻撃の状況を把握するシステム [Suzuki 11]，全国の地中および地表の揺れをリアルタイムで把握するシステム[1] などがある．このように，客観的な情報の可視化技術は，環境がどのような状況であるかを直感的に把握できるようにすることで，我々の生活を支えている．

一方，我々を取り巻く環境には，客観的な情報だけでなく，気温，湿度，食感，環境音など，五感とその快不快に関する主観的な情報も豊富である．このような情報は，一般的に言語によって表現される．特に，日本語においては，オノマトペ（擬音語や擬態語）によって，客観的な情報とその情報に対する主観を効率よく伝えることができる．例えば，気温や湿度に関しては「ポカポカ（暖かくて心地良い）」や「ジメジメ（湿度が高くて不快）」などの表現を使うことができ,「サクサク（歯ごたえが適度で心地良い）」や「モチモチ（粘り気があって心地良い）」などの表現によって食感を表すことができる．環境音に関しては,「ガヤガヤ（音が大きくて不快）」や「ワイワイ（音が大きくて楽しげ）」などの表現がある．

このように，我々の周囲には主観的な評価情報（感性情報）と一体となった情報が満ち溢れている．そのため，これらの情報を可視化できれば，特定の感性情報（例えば,「聴覚に関する不快な情報」）がどの地域に密集しているかなどを捉えることが可能になる．

本研究では，このような技術の確立を目指し，様々なオノマトペに関して，それらが五感（触覚，味覚，嗅覚，聴覚，視覚）のどの感覚カテゴリに属するのか，また，その表現が快・不快のどちらに属するのかを判定する．

まず，本研究では，オノマトペがどの感覚カテゴリに属するかを機械学習によって判定する．学習手法としては，Support Vector Machine（SVM）[Cortes 95] を用いる．これにより，慣習的に使用されるオノマトペや新奇なオノマトペが属する感覚（五感および快・不快）の判定を試みる．次に，開発したシステムの評価を行い，オノマトペが属する感覚カテゴリの判断に関する，人間と機械の違いなどを考察する．

*連絡先：京都大学 学際融合教育研究推進センター
京都市下京区中堂寺粟田町 91
京都リサーチパーク 9 号館 5 階
E-mail: tetsuakinakamura8@gmail.com
[1]強震モニタ（独立行政法人 防災科学技術研究所）
http://www.kyoshin.bosai.go.jp/kyoshin/docs/kyoshinmonitor.html

2　関連研究

オノマトペの印象形成には，そのオノマトペを構成する音が影響すると言われている [田守 99]．また，オノマトペに限らず，ブーバ・キキ効果 [Ramachandran 01] に見られるように，特定の音が特定の印象と結びついている現象は音象徴（sound symbolism）[Hinton 95] として知られている．藤澤ら [藤沢 06] や Ueda et al.[Ueda 12] では，オノマトペの印象を表す複数の感性尺度を想定し，各尺度の値を数量化理論 I 類によって推定している．また，オノマトペではないが，言葉の持つ印象をその音韻的特徴を手がかり情報として解析している研究もある．例えば，ファジィ積分を用いて与えられた言葉の印象を推定する研究 [長町 93] や，キャラクタの名前の音韻的特徴とそのキャラクタの強さの印象の組み合わせを SVM によって学習し，強そうな印象の名前や弱そうな印象の名前を自動生成する研究 [三浦 12, Aramaki 12] がある．

本研究においても，与えられたオノマトペをあらかじめ想定した音韻記号で表現し，SVM の素性として用いる．ただし，印象を表す感性尺度の値を推定するのではなく，オノマトペの属するカテゴリを推定する点が先行研究とは異なる．

3　材料：対象とするオノマトペ

本研究では，ミニログサイト Twitter[2]における，2011 年 7 月 15 日から 2012 年 7 月 31 日までの日本語の投稿（以降，ツイートと表記）を収集し，これらのツイート（24,817,903 ツイート）から抽出されたオノマトペを学習に用いた．

抽出されるオノマトペは，抽出のしやすさを考慮し，片仮名表記で繰り返し構造を持つ表現（“ドキドキ”，“チャリンチャリン” など）で出現回数が 10 回以上のものとした．ただし，機械による抽出の後，3 名の判定による多数決を行い，オノマトペではないと判断されたもの（“ジョジョ”，“ヤバイヤバイ”，“トイレトイレ” など）を除外した．結果的として，本研究で使用するオノマトペの数は 845（異なり数）となった．その一部（延べ出現回数が上位 20 までのもの）を表 1 に示す．

表 1: 本研究で使用するオノマトペの一部（出現回数が上位 20 までのもの）

オノマトペ	出現回数	オノマトペ	出現回数
イライラ	10968	ガリガリ	3450
ギリギリ	9625	ニヤニヤ	2898
ドキドキ	8353	ガンガン	2697
ジメジメ	7966	ダラダラ	2638
ワクワク	7721	キラキラ	2570
ガラガラ	5241	パンパン	2482
ニコニコ	5174	ボロボロ	2456
ジリジリ	4448	フラフラ	2253
バタバタ	3933	ウロウロ	2230
ゴロゴロ	3463	ハァハァ	2217

（注）出現回数は延べ出現回数を意味する．

4　手法

4.1　素性

オノマトペを表現するための音韻記号としては，先行研究 [藤沢 06, Ueda 12] を参考に，表 2 に示す 22 種類の記号を用いる．なお,「ァィゥェォ」（以降では，小書き文字と呼ぶ）に関しては，その小書き文字の使用が発音可能な使い方であれば，その小書き文字を長音，拗音の一部，合拗音の一部のいずれかとして扱う．一方，発音不可能な使い方であれば，その小書き文字は単独の母音として扱う．

表 2: 本研究で使用する音韻記号

記号	意味	記号	意味
a	ア段の母音	y	ヤ行の子音
i	イ段の母音	r	ラ行の子音
u	ウ段の母音	w	ワ行の子音
e	エ段の母音	N	撥音（ン）
o	オ段の母音	Q	促音（ッ）
k	カ行の子音	R	長音
s	サ行の子音	D	濁音
t	タ行の子音	P	半濁音
n	ナ行の子音	Y	後続母音が拗音
h	ハ行の子音	W	後続母音が合拗音
m	マ行の子音	v	有声唇歯摩擦音（ヴ）

例えば,「ドキドキ」は「toDkitoDki」となり,「キュンキュン」は「kYuNkYuN」となる．また,「きぃきぃ」は「kiRkiR」となり,「ぽぁぽぁ」は「hoPahoPa」となる．

オノマトペの音韻的特徴に関しては，先頭や末尾の音韻の影響を考慮し [田守 99]，便宜的に先頭記号（^）と末尾記号（$）を用いて，音韻記号の組み合わせ（bigram と trigram）も特徴として用いる．

与えられたオノマトペは，片仮名表記にされた後で濁点と半濁点を分離され，文字記号変換表（表 3）を参照しながら，表 2 に示す記号による素性ベクトルを作成する．

[2]https://twitter.com/

表 3: 文字記号変換表

文字	記号	文字	記号	文字	記号	文字	記号	文字	記号
ヮ	Wa								
ヴァ	va	ヴィ	vi	ヴ	vu	ヴェ	ve	ヴォ	vo
ア	a	イ	i	ウ	u	エ	e	オ	o
カ	ka	キ	ki	ク	ku	ケ	ke	コ	ko
サ	sa	シ	si	ス	su	セ	se	ソ	so
タ	ta	チ	ti	ツ	tu	テ	te	ト	to
ナ	na	ニ	ni	ヌ	nu	ネ	ne	ノ	no
ハ	ha	ヒ	hi	フ	hu	ヘ	he	ホ	ho
マ	ma	ミ	mi	ム	mu	メ	me	モ	mo
ヤ	ya			ユ	yu			ヨ	yo
ラ	ra	リ	ri	ル	ru	レ	re	ロ	ro
ワ	wa	ヲ	wo	ン	N	ッ	Q	ー	R
ァ	A	ィ	I	ゥ	U	ェ	E	ォ	O
ャ	Ya			ュ	Yu			ョ	Yo
゛	D	゜	P						

（注）A，I，U，E，O は小書き文字であることを示す.

4.2 SVM による学習

対象とするカテゴリ

本研究では，TinySVM [3] を用いて，与えられた未知のオノマトペが属するカテゴリの学習を行う．オノマトペが属するカテゴリとしては，五感（触覚，味覚，嗅覚，聴覚，視覚）と快不快（快，不快）の 7 つのカテゴリを対象とする.

訓練データの作成

SVM に用いる訓練データを作成するために，3 章で述べた 845 のオノマトペを用いて実験を行った．実験参加者は，3 名（男性 2 名，女性 1 名，平均年齢 31.7 歳）である．参加者には，提示された各オノマトペについて，本節冒頭の 7 カテゴリに関して，2 値判定（属する/属さない）を求めた．各オノマトペが属するカテゴリは，3 名の参加者のうち 2 名以上が属すると回答したものとした．なお，各オノマトペが属するカテゴリに関しては，複数回答を許可した．そのため，オノマトペが属するカテゴリ数は，各オノマトペによって異なる.

この実験によって得られた参加者の回答の一致率を表 4 にまとめる．表 4 における各数値は，3 名の実験参加者の各組み合わせ（A-B，B-C，C-A）の一致率と κ 統計量に関する平均値である．この表より，触覚，聴覚，不快では良好な κ 統計量となっているのに対して，それ以外の感覚では κ 統計量が低い値となっているのが分かる．このことは，触覚，聴覚，不快以外の感覚では，それらの感覚に属するオノマトペが極めて少ない，あるいは，オノマトペが属する感覚を明確に判断することが難しいことを示唆している.

表 4: オノマトペの属性評価実験の結果の平均一致率

	一致率	κ 統計量
触覚	0.75	0.39
味覚	0.96	0.02
嗅覚	0.98	0.27
聴覚	0.78	**0.55**
視覚	0.48	0.10
快	0.77	0.29
不快	0.69	0.34

(注)0.4 以上の κ 統計量は太字で表示.

訓練データには，この実験で得られた回答を用いた．オノマトペ w の感覚 s に関する正例あるいは負例のラベル $l(w,s)$ は，実験参加者 3 名の回答を用いて，式 (1) に従って分類した.

$$l(w,s) = \begin{cases} +1 & (x(w,s) \geq 2) \\ -1 & (x(w,s) = 0) \end{cases} \quad (1)$$

ただし，式 (1) において，$x(w,s)$ は w が s に属すると回答した実験参加者の人数である．なお，$x(w,s) = 1$ の場合は，判断が曖昧なデータであるとして，そのデータは使用しないものとした．したがって，各感覚毎に使用するデータ数は異なる．各感覚において使用されたデータ数を表 5 に示す.

[3] http://chasen.org/ taku/software/TinySVM/

表 5: 実験に使用したデータ数

感覚	正例数	負例数	合計
触覚	238	444	682
味覚	2	792	794
嗅覚	1	821	822
聴覚	399	322	721
視覚	285	112	397
快	125	515	640
不快	269	342	611

表 6: SVM による学習の 10 分割交差検定の結果

	適合率	再現率	F 値	κ	正例数/負例数
触覚	58.04	53.80	54.76	0.39	238/444
聴覚	76.83	79.30	77.57	0.55	399/322
視覚	81.01	80.11	80.01	0.10	285/112
快	48.32	40.49	40.16	0.29	125/515
不快	66.31	65.15	64.87	0.34	269/342

（注 1）表中の「κ」は表 4 の κ 統計量を示し，それ以外の数値は平均値を示す．
（注 2）味覚と嗅覚に関しては，正例の極端な少なさによる不適切な学習により省略．

5　結果

5.1　学習結果

4.2 節の手順で作成された訓練データを用いて SVM による学習を行い，その分類精度を評価した（10 分割交差検定法）．学習には，2 次の多項式カーネルを用いた．その結果を表 6 に示す．

表 6 から，聴覚と視覚に関する SVM の精度が良いことが分かる．これらに比べると劣るものの，触覚，快，不快に関しても，人間の κ 統計量（表 4）を考慮すれば，良好な結果であると言える．全体傾向としては，正例数の多さと精度に何らかの関係が存在する可能性が読み取れる．

5.2　判定誤りの音韻的特徴

5.1 節の評価において，正しく判定できたオノマトペとそうでないオノマトペのうち，出現回数が上位 5 位までのものを表 7 に示す．表 7 より，各感覚について，以下のことが言える．

触覚：正例を負例と判断したもの（以降，正例誤り）には「オ段+オ段」の繰り返しが多く，負例を正例と判断したもの（以降，負例誤り）には「オ段+イ段」の繰り返しが多い傾向が見られる．

聴覚：正例誤りには「ア段+イ段」の繰り返しが多く，負例誤りには傾向らしきものは見られない．

視覚：正例誤りに関しては濁音や半濁音の影響が見られるが，負例誤りに関しては傾向らしきものは見られない．

快　：正例誤りと負例誤りに傾向らしきものは見られない．

不快：快と同様に誤り判定のものには傾向が見られない．

以上から，誤り判定での特徴が見られる感覚とそうでない感覚があると言える．また，誤り判定になったオノマトペの特徴が，その判定カテゴリ（例えば，正例を負例と語判定した場合は負例）に特徴的であるとは言えない．これらを考慮すれば，音韻要素以外にも，感覚の属性を規定する要因（例えば，音韻を発音する際の口腔の形状や動きなどの物理的特徴，文字から受ける印象，など）の影響を受けている可能性が考えられる．

5.3　正例数と精度の関係

5.1 節では，正例数と精度の間に何らかの関係が存在する可能性を述べた．そこで，表 6 に示した結果に関して，正例数と精度（F 値）に関する Pearson の積率相関係数を求めた．すなわち，各感覚において，ある交差検定時の正例数を x，そのときの精度（F 値）を y として，x と y に関する相関係数を求めた．その結果を表 8 に示す．

表 8 より，聴覚と快に関しては有意な相関が見られたが，触覚と視覚と不快に関しては有意な相関は見られなかった．ただし，視覚と不快に関しては，相関係数としてはある程度の数値が得られている．そのため，正例数と精度の間には相関関係が存在する可能性があり，正例数を充実させることで精度を向上できると考えられる．

6　考察

6.1　感覚間の類似度

表 6 に示す精度が十分に高くない理由として，ある感覚と別の感覚が類似していることで，感覚の弁別が人間にとって難しいことが考えられる．そこで，4.2 節の回答に関して，各感覚カテゴリ間にどの程度の一致があるのかを分析した．これは，ある 2 つの感覚を $s1$,

表 7: オノマトペの属性に関する人間の判定と SVM の判定の結果例

	人間の判定/SVM の判定							
	正/正		正/負		負/負		負/正	
触覚	ジリジリ	(4448)	ジメジメ	(7966)	ワクワク	(7721)	ニコニコ	(5174)
	ゴロゴロ	(3463)	トロトロ	(740)	バタバタ	(3933)	ボチボチ	(1303)
	ガリガリ	(3450)	ヒリヒリ	(711)	ニヤニヤ	(2898)	ノリノリ	(1248)
	バリバリ	(1986)	ボコボコ	(700)	ガンガン	(2697)	モリモリ	(989)
	ポカポカ	(1890)	ドロドロ	(655)	ダラダラ	(2638)	コミコミ	(748)
聴覚	ガラガラ	(5241)	サクサク	(1806)	ワクワク	(7721)	ジリジリ	(4448)
	バタバタ	(3933)	ワイワイ	(1528)	ニコニコ	(5174)	ポカポカ	(1890)
	ゴロゴロ	(3463)	パリパリ	(704)	ニヤニヤ	(2898)	ボチボチ	(1303)
	ガンガン	(2697)	パチパチ	(599)	フラフラ	(2253)	キュンキュン	(900)
	パンパン	(2482)	トントン	(498)	ウロウロ	(2230)	プンプン	(816)
視覚	ニヤニヤ	(2898)	ニコニコ	(5174)	ジメジメ	(7966)	ガラガラ	(5241)
	ガンガン	(2697)	ジリジリ	(4448)	タンタン	(1100)	ツルツル	(1044)
	キラキラ	(2570)	ガリガリ	(3450)	コツコツ	(775)	ヘロヘロ	(1038)
	ボロボロ	(2456)	パンパン	(2482)	コミコミ	(748)	モグモグ	(918)
	フラフラ	(2253)	バンバン	(1086)	ウトウト	(717)	ウマウマ	(649)
快	サクサク	(1806)	ドキドキ	(8353)	イライラ	(10968)	フラフラ	(2253)
	ツルツル	(1044)	ワクワク	(7721)	ギリギリ	(9625)	モクモク	(540)
	フワフワ	(719)	ニコニコ	(5174)	ジメジメ	(7966)	スカスカ	(469)
	スイスイ	(542)	キラキラ	(2570)	ジリジリ	(4448)	ピキピキ	(420)
	サラサラ	(528)	ポカポカ	(1890)	ガリガリ	(3450)	トコトコ	(397)
不快	イライラ	(10968)	ニヤニヤ	(2898)	ワクワク	(7721)	キラキラ	(2570)
	ギリギリ	(9625)	ガンガン	(2697)	ニコニコ	(5174)	ペロペロ	(1665)
	ジメジメ	(7966)	フラフラ	(2253)	パンパン	(2482)	ボチボチ	(1303)
	ガラガラ	(5241)	バンバン	(1086)	ハァハァ	(2217)	ピカピカ	(1181)
	バタバタ	(3933)	ドンドン	(1077)	ポカポカ	(1890)	モテモテ	(668)

（注 1）表上部の「正/負」は「属する/属さない」と判定したことを表す.
（注 2）味覚と嗅覚に関しては，正例の極端な少なさによる不適切な学習により省略.
（注 3）括弧の数字は Twitter での出現回数を示す.

表 8: 正例数と精度に関する Pearson の積率相関係数

感覚カテゴリ	相関係数	p 値
触覚	0.14	0.71
聴覚	0.75	0.01
視覚	0.41	0.24
快	0.71	0.02
不快	0.32	0.37

（注）味覚と嗅覚に関しては，正例数が極端に少ないために相関係数を算出できないので省略.

$s2$ として，各オノマトペに対する属性の有無（属する/属さない）について，$s1$ と $s2$ の κ 統計量を求めることで行った．その結果を表 9 に示す.

表 9 を見ると，全体的に大きい κ 統計量は得られていないことが分かる．この中で比較的大きな値になっている感覚の組み合わせは，以下の A~D であった.

A. 触覚と視覚

B. 触覚と快

C. 触覚と不快

D. 視覚と不快

我々は，物体に触れる際にその物体を見ている経験から，表面を見れば触り心地が分かる．これにより，触覚と視覚を同じような意味合いとして我々が認識して

表 9: オノマトペが属する感覚の κ 統計量（人間）

	触覚	味覚	嗅覚	聴覚	視覚	快	不快
触覚		0.03	0.00	0.01	0.17	0.15	0.16
味覚			0.04	0.01	0.01	0.01	0.02
嗅覚				0	0.00	0.00	0.00
聴覚					0.01	0.00	0
視覚						0.06	0.15
快							0.02
不快							

（注）表の数値は実験参加者の平均値を示す.

表 10: オノマトペが属する感覚の κ 統計量（SVM）

	触覚	味覚	嗅覚	聴覚	視覚	快	不快
触覚		−	−	0	0.15	0.11	0.14
味覚			−	−	−	−	−
嗅覚				−	−	−	−
聴覚					0	0	0
視覚						0.12	0.23
快							0
不快							

（注）味覚と嗅覚は学習不能のため，κ 統計量は算出していない.

いると考えれば，A の結果は予想できる．また，こうした経験で得られる心的印象（i.e., 快 · 不快）が触覚や視覚の感覚的意味と結びつきやすくなった（B，C，D）とするのは自然である.

オノマトペは音韻によって様々な感覚を伝える言語表現であり，この多感覚知覚という特徴において，共感覚比喩（synesthetic metaphor / synaesthetic metaphor）[Ullmann 51, Williams 76, Yu 03, Werning 06, 楠見 94] と呼ばれる現象に近いと言える．近年，このような言語表現の解釈には，我々が経験によって蓄積した具体的な場面の知識が仲介している可能性が示唆されている [仲村 12]．そのため，今後の更なる言語認知の研究成果から，オノマトペに関する様々な計算処理に必要な手がかりを得られることが期待できる.

次に，表 9 と同様の分析を，システムの判定結果に対しても行った．その結果を表 10 に示す．表 10 を見ると，上述の A∼D の組み合わせにおいて比較的大きな κ 統計量が得られていることが分かる.

以上から，人間の判定と機械の判定の両方で同じ感覚間の類似性が見られる．この結果は，表 6 の精度が低い原因の 1 つとして，感覚間類似度による弁別の難しさが挙げられることを示唆している.

6.2　本研究の応用可能性

1 章で述べたように，我々の生活環境における感性が関与する情報を可視化できれば，我々の主観的な環境の情報を直感的に捉えられる．その 1 つとして，著者らが現在開発中である，五感に関する市街地の快 · 不快マップの画面例を図 1 に示す．この図は，Twitter の投稿から抽出されたオノマトペを投稿位置と属する感覚に基づいて集計し，新宿駅周辺における聴覚と視覚の快 · 不快マップとして表示したものであり，赤に近いほど投稿が多いことを示す．図 1 から，新宿駅周辺では，聴覚に関して不快な印象が多い一方，視覚に関して快な印象が多いことが読み取れる．このシステムを使えば，どの場所でどのような感覚が得られるかをを把握することができる．このようなマップをリアルタイムあるいは一定の期間で生成することは，街歩きでの経路設定や，都市開発計画の支援など，様々な分野に応用できる．ただし，このマッピングを精度良く行うには，本章で述べた様々な課題を解決しなければならず，本研究のさらなる発展が必要である.

7　むすび

本研究では，我々の環境を取り巻く主観的な情報の可視化のためにオノマトペに着目し，様々なオノマトペに関して，それらが属する感覚カテゴリ（触覚，味覚，嗅覚，聴覚，視覚，快，不快）を SVM によって判定するシステムを実装した．SVM に用いる素性としては，あらかじめ想定した音韻記号とそれらの記号の出現順序を用いた.

実装したシステムの分類精度の評価および考察の結果は以下の通りである.

1. 分類精度の評価では，聴覚と視覚で良好な結果が得られた．その他の感覚に関しては，人間の回答に関する一致率が十分ではなかった.

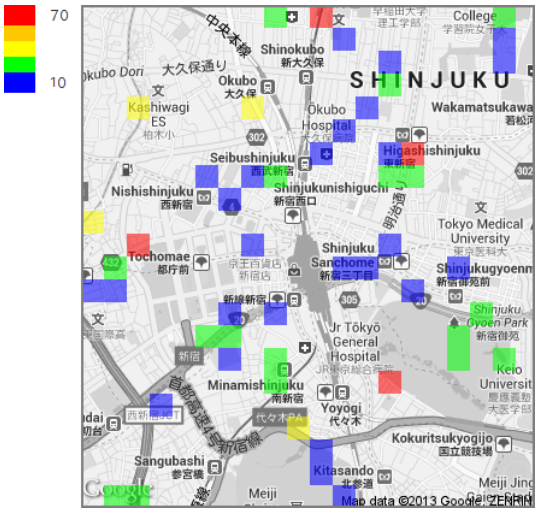
2. 実験参加者の回答の感覚間類似度，および，SVM の判定結果の感覚間類似度の分析から，幾つかの感覚の弁別は困難であり，このことが分類精度の低下を招いている可能性が示された.

今後の課題としては，解析誤りの特徴を適切に捉える手法の検討による精度向上や本研究の具体的な応用が挙げられる.

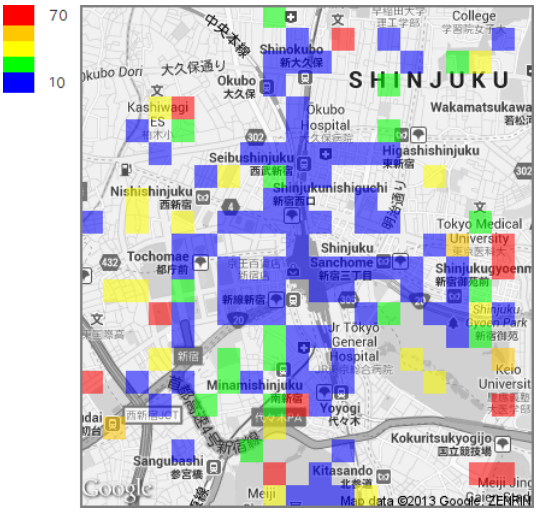
謝辞

本研究は JST さきがけ「自然言語処理による診断支援技術の開発」プロジェクトの助成を受けた.

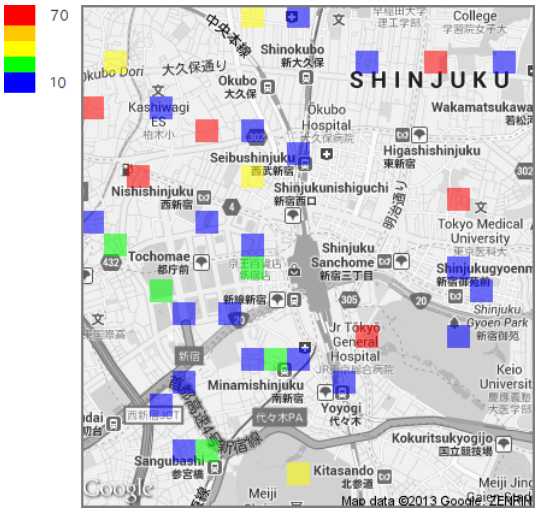
（a）聴覚 + 快



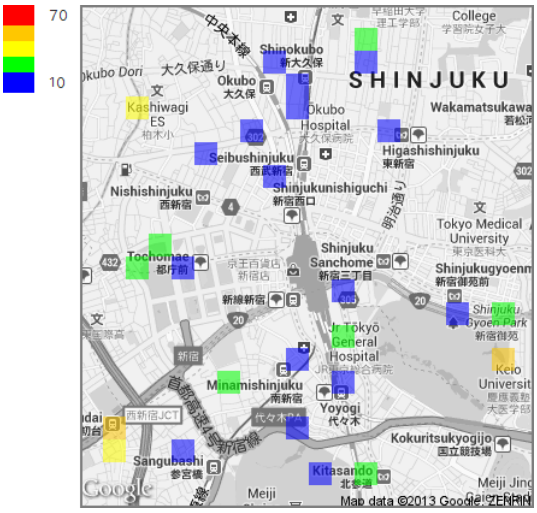
（b）聴覚 + 不快



（c）視覚 + 快



（d）視覚 + 不快



（注）マップ上の色彩ブロックは，そのブロックの範囲内で Twitter に投稿されたオノマトペのうち，指定した感覚に属するものの占める割合（%）を示す．この図では，青（下限）が 10% であり，赤が 70% 以上を示す．

図 1: 五感に関する市街地（新宿駅周辺）の快 · 不快マップ

参考文献

[Aramaki 12] Aramaki, E., Yasuda, S., Miyabe, M., Miura, S., and Murata, M.: Which is Stronger? : Discriminative Learning of Sound Symbolism, in *Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci2012)*, p. 2627 (2012)

[Cortes 95] Cortes, C. and Vapnik, V.: Support-Vector Networks, *Machine Learning*, Vol. 20, No. 3, pp. 273–297 (1995)

[藤沢 06] 藤沢 望, 尾畑 文野, 高田 正幸, 岩宮 眞一郎：2 モーラの擬音語からイメージされる音の印象, 日本音響学会誌, Vol. 62, No. 11, pp. 774–783 (2006)

[Hinton 95] Hinton, L., Nichols, J., and Ohala, J. J. eds.: *Sound Symbolism*, Cambridge University Press, Cambridge (1995)

[Houghton 12] Houghton, A., Prudent, N., Scott III, J. E., Wade, R., and Luber, G.: Climate Change-Related Vulnerabilities and Local Environmental Public Health Tracking through GEMSS: A Web-Based Visualization Tool, *Applied Geography*, Vol. 33, pp. 36–44 (2012)

[楠見 94] 楠見 孝：比喩の処理過程と意味構造, 風間書房, 東京 (1994)

[三浦 12] 三浦 智, 村田 真樹, 保田 祥, 宮部 真衣, 荒牧 英治：音象徴の機械学習による再現：最強のポケモンの生成, 言語処理学会第 18 回年次大会 発表論文集, pp. 65–68 (2012)

[長町 93] 長町 三生：言葉の響きに関する感性工学, 日本音響学会誌, Vol. 49, No. 9, pp. 638–644 (1993)

[仲村 12] 仲村 哲明, 坂本 真樹, 内海 彰：具体的な場面想起の仲介に基づく異感覚間形容詞比喩の解釈, 認知科学, Vol. 19, No. 3, pp. 314–336 (2012)

[Ramachandran 01] Ramachandran, V. S. and Hubbard, E. M.: Synaesthesia: A window into perception, thought and language, *Journal of Consciousness Studies*, Vol. 8, No. 12, pp. 3–34 (2001)

[菅田 11] 菅田 誠治, 大原 利眞, 黒川 純一, 早崎 将光：大気汚染予測システム (VENUS) の構築と検証, 大気環境学会誌, Vol. 46, No. 1, pp. 49–59 (2011)

[Suzuki 11] Suzuki, M. and Inoue, D.: DAEDALUS: Practical Alert System Based on Large-scale Darknet Monitoring for Protecting Live Networks, *Journal of the National Institute of Information and Communications Technology*, Vol. 58, No. 3, pp. 51–60 (2011)

[田守 99] 田守 育啓, ローレンス・スコウラップ：オノマトペ –形態と意味–, くろしお出版, 東京 (1999)

[Ueda 12] Ueda, Y., Shimizu, Y., and Sakamoto, M.: System Construction Supporting Communication with Foreign Doctors Using Onomatopoeia Expressing Pains, in *Proceedings of the 6th International Conference of Soft Computing and Intelligent System*, pp. 508–512 (2012)

[Ullmann 51] Ullmann, S.: *The principles of semantics*, Blackwell, Oxford (1951)

[Werning 06] Werning, M., Fleischhauer, J., and Beseoglu, H.: The cognitive accessibility of synaesthetic metaphors, in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp. 2365–2370 (2006)

[Williams 76] Williams, J. M.: Synaesthetic adjectives: A possible law of semantic change, *Language*, Vol. 52, No. 2, pp. 461–478 (1976)

[Yu 03] Yu, N.: Synesthetic metaphor: a cognitive perspective, *Journal of literaty semantics*, Vol. 32, No. 1, pp. 19–34 (2003)

検索エンジンを用いた情報検索におけるユーザ行動の分析

Analysis of User's Behavior in Information Retrieval Using Search Engine

桑折 章吾[1*]　加藤 優[1]　高間 康史[1]
Shogo Kori[1] , Yu Kato[1] , Yasufumi Takama[1]

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

Abstract: 本稿では，検索エンジンを用いた情報検索におけるユーザ行動を分析した結果について報告する． 我々は,「動向に関する問い」を対象とした検索エンジン構築を目指し，その基本的検索機能について検討を進めている．既存検索エンジンを用いた検索でも，ユーザは異なる意図に基づく基本的検索を組み合わせて目的を達成しているとの考えに基づき，本稿では検索意図の観点からユーザの情報検索行動を分析し，得られた結果に基づき動向に関する基本的検索機能について考察する．

1　はじめに

本稿では，検索エンジンを用いた情報検索におけるユーザ行動を分析した結果について報告し，得られた結果に基づき我々が目指す「動向に関する問い」を対象タスクとした検索エンジンの基本検索機能について考察する．Web の魅力の一つとして，世界中のリアルタイムな情報が収集可能である事が挙げられる．近年では，ソーシャルメディアの普及によりリアルタイムな情報がますます注目されている．その一方で，Web が利用されるようになってから 20 年弱が既に経過し，Web 上には膨大な量の情報が蓄積されている．このように蓄積された情報に着目し，過去の情報を知るためのリソースとして Web を有効的に活用していくことも検討すべきであると考えるが，既存検索エンジンが提供する機能と，ユーザの情報収集目的との乖離が大きいという問題がある．すなわち，既存検索エンジンが提供するのは，キーワードベースの検索要求指定，ページ単位での結果出力といった低機能にとどまったままであり，情報要求をキーワードに分解するのに要するユーザの負担が大きいと考える．

検索エンジンの知的化・高機能化に関するアプローチとしては，対象ドメインを限定することが考えられるが [8][10]，本稿で検討している検索エンジンでは，ドメインに依存せず，広く一般的に利用可能であることを目指している．提案する検索エンジンでは，対象タスクに特化したいくつかの基本検索機能を検討しているが，それらは組み合わせて用いることで多様かつ高度な検索目的を達成可能である必要がある．既存検索エンジンでも，ユーザは異なる意図に基づく基本的検索を組み合わせて目的を達成しているとの考えに基づき，本稿では検索意図の観点からユーザの情報検索行動を分析する．

本稿では，ユーザの情報検索行動を調査するために行った実験について述べ，得られた結果に基づき動向に関する基本検索機能を考察する．実験では既存検索エンジンを使い Web から答えを見つける問題を実験協力者に出題した．入力されたクエリを検索意図毎に分類して分析した結果，ユーザは自らの情報要求を満たすために異なる意図に基づく基本検索機能を多様に組み合わせて検索を行っていることを示す．分析結果に基づき，構築中の検索エンジンに必要な基本検索機能について考察する．

2　関連研究

2.1　次世代検索エンジン

Web が普及してから 20 年弱が経ち，Web 上は情報過多となってきている．現在，Web 上に蓄積された情報を探す方法としては，検索エンジンが最も用いられている．しかし，既存の検索エンジンは指定したキーワードを含む Web ページを返すという汎用的ではあるが低機能なものにとどまっているため必要とする情報

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

に辿り着くまでに何度も検索を繰り返す必要がある場合が発生する．このような手間を省くために，対象を絞ることでより効率的な検索を実現することを目指す次世代検索エンジンの研究・開発がなされている．

亀井ら [4] は，WWW に存在するソフトウェア開発に関する知見や情報を検索するための検索エンジンを提案している．過去に多くのソフトウェアが開発され，それらに関するノウハウや関連情報などが Web 上で多数公開されている．しかし，それらは体系化されず Web 上に点在しており，網羅的・効率的に情報収集することは困難である．そのため現状では，似たようなソフトウェアが開発されていたり，同じようなミスでソフトウェア開発が滞ることがある．亀井らの提案する検索エンジンは，ソースコードそのものやそのコメント，開発日記，Tips などソフトウェアの知見に関する情報にドメインを特化することで，既存検索エンジンよりも効率的な検索を目指している．

対象ドメインを限定しない検索エンジンとして，動向情報を対象としたコンテクスト検索エンジンが提案されている [6][7]．動向情報とはある商品の価格や売上の状況，ある会社の業績状況，内閣や政党の支持状況などの事であり，幾つかの統計量の時系列データを基として，その変化を通時的にとらえつつ，それらを総合的にまとめ上げることで得られるものである [5]．動向情報は検索エンジンの検索数やヒット数などの主観的動向情報と，アイテムの価格や販売量，生産量などの客観的動向情報に分けられる [6]．文献 [6] では，主観的動向情報として Google Trends[1]で公開されている検索数や Yahoo!検索ランキング[2]で公開されている急上昇ワードなどを収集対象としている．客観的動向情報としては，ベジ探[3]で公開されている野菜の価格や自転車産業振興協会[4]で公開されている自転車の生産台数などの統計データを収集対象としている．収集した動向情報はデータベース（MySQL）に格納し，Web アプリケーションフレームワークに Ruby on Rails を用いてシステムを実装している．プロトタイプシステムのインタフェースを図 1 に示す．このシステムでは「指定アイテムに関する動向情報のピーク（最大値）時期の検索」,「指定期間に動向情報の最大値を持つアイテムの検索」の 2 つの基本検索機能を実装している．

図 1: コンテクスト検索エンジンのインタフェース

2.2　情報検索におけるユーザ行動

既存検索エンジンを用いた情報検索では，ユーザは異なる意図に基づき基本検索機能を組み合わせて目的を達成している．ユーザの検索意図はクエリとして表現されるが，うまく表現できない場合もある．そのような場合，検索を繰り返しても，膨大な検索結果の中に必要な情報を含むページが埋没してしまい，必要とする情報にたどり着く事は難しい．

藤田ら [3] らは，ユーザの連続した検索からクエリ変更意図を推測することで，ユーザの検索の先を読み，自動で検索する先読み検索を提案している．提案手法では，クエリログからユーザのクエリ変更意図について分析し，その結果に基づき SVM によるクエリ変更意図の自動分類を行う．クエリ変更意図毎に先読み検索を行っている．

南ら [9] は，ユーザが問題解決を目的に複数の検索結果を確認しながら情報を集めて行く際の作業効率向上を目的とし，検索結果のフィルタリングを行っている．ユーザの Web ページ閲覧時の行動をモニタリングして検索タスクにおけるユーザ意図を動的に抽出する手法を提案し，検索結果のフィルタリングシステムを実装している．

旭ら [1] は,「iPod を買う」→「iPod を使う」→「iPod が壊れて修理する」のようにある話題の中で行われる一連の行動の流れを行動連鎖と呼び，ブログのエントリ内，エントリ間という 2 つの観点からシーケンシャルパターンマイニングを用いて行動連鎖の抽出を行っている．抽出した結果に基づきユーザが目的とする行動に応じて必要な Web ページをランキングしてユーザに提示するシステムを提案している．順序だてて行動連鎖をユーザに提示することで，ユーザは自分にとって必要な情報を効率よく調べることが可能となる．また，行動プロセス提示によりユーザは問題解決のためにどのような事を調べれれば良いのかを把握することができる．

[1] http://www.google.co.jp/trends/
[2] http://searchranking.yahoo.co.jp
[3] http://vegetan.alic.go.jp
[4] http://www.jbpi.or.jp

3　ユーザ行動の分析

検索意図の観点からユーザの情報検索行動を調査する実験を行った．3 節で行った実験の概要および，その結果を分析し検索意図を分類した結果について示す．4 節では 3 節で定めた分析意図によりログデータにラベル付けを行い分析を行った結果を示すと共に，構築中の検索エンジンに必要な基本検索機能について考察する．

3.1　ユーザ行動調査のための実験

実験で用いた問題を図 2 に示す．実験では，二枚の画像から検索エンジン（Google）を用いて画像の撮影場所を特定する問題を出題した．図 2 の問題のように答えをどのような視点から探せばよいか，画像をクエリとしてどの様に表現すればよいかが明確ではない場合，実験協力者は自ら解答への道筋を考えなければならない．答えを見つけるアプローチの仕方が様々であり，多様な検索行動が生じることが期待できるためこの問題を選択した．実験は実験協力者 3 名を対象に行い，各協力者は平均して約 20 分で正解を出すことができた．実験協力者がどのような検索を行い，どのようなページを開いているかを正確に調査するため実験中に oCam[5]を用いて画面のキャプチャを行った．

図 2: 実験で用いた問題

3.2　検索意図の分析

入力されたクエリを分析し，検索意図を図 3 の様に分類した．実験協力者の検索意図は Verify（検証）と Discover（発見）の二つのタイプに大別される．また，何かに関する情報を探す際には，対象ページを限定しない Informational と，特定の Web ページの発見を目的とする Navigational に分類できる [2]．さらに Discover-Informational には，正確に目標を定めた検索（Pinpoint）と幅広い検索結果を期待した検索（Broad）が存在する．Discover-Informational-Pinpoint には条件を満たす情報を一つだけ探す検索（Single）と複数の情報を探す検索（Multi）があり，Multi にはそれらが一覧のようにまとめられている Web ページを期待した検索（List）と一つずつ別ページに存在することを期待した検索（Item）が存在する．同じクエリであっても検索される段階によって検索意図が異なると考えられる場合があった．

今回の実験で入力されたクエリの例を以下に示す．今回の実験では Item に該当する検索は行われなかった．

- Verify-Informational
  「市場　スペイン　バルセロナ」
  …写真がバルセロナ（スペイン）の市場で撮影したものであることを確認
- Verify-Navigational
  「サン・ジョセップ市場　Google マップ」
  …サン・ジョセップ市場の場所を Google マップで確認
- Discover-Navigational
  「バルセロナ　wiki」
  …バルセロナについて書かれている Wikipedia のページを期待
- Broad
  「ヨーロッパ　市場」
  …ヨーロッパにある市場について幅広い情報を期待
- Single
  「サン・ジョセップ市場　住所」
  …サン・ジョセップ市場の住所が書かれている Web ページを期待
- List
  「スペイン　市場　一覧」
  …スペインの市場が一覧のようにまとめられている Web ページを期待

[5]http://ohsoft.net/product-oCam.php

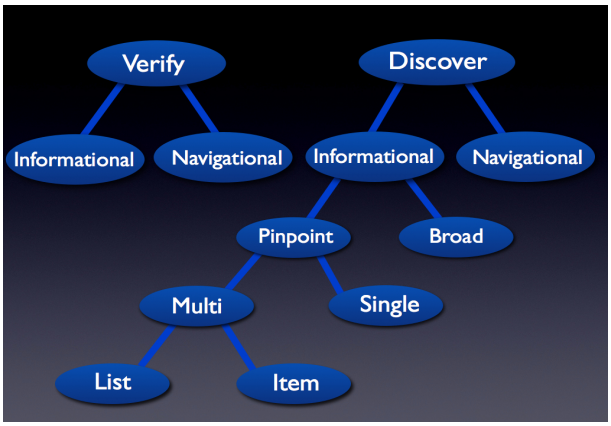


図 3: 検索意図の分類

4　実験結果と基本検索機能の考察

4.1　ログデータへのラベル付け

ユーザ行動をさらに分析するために図 2 と同じ様な問題を用いて再び実験を行い，収集したログデータに図 3 に示したラベルを付けた．実験は計 3 問行い，問題 1 では 3 人，問題 2 では 3 人，問題 3 では 5 人の実験協力者を対象とした．問題 1〜3 の全実験協力者が答えを導くことができたが，解答に要した時間にはばらつきが見られた．表 1 に，全実験協力者についてログデータに付与された各ラベルの数を示す．なお実験協力者 A〜H は解答が早かった順に上から並べている．表より，Discover の Navigational，Broad，List は利用者が少ない一方，Verify の Navigational と Informational，Discover の Pinpoint（Single，Multi）はほぼ全員が利用していることがわかる．また，3 問全てに共通して前半では Pinpoint（Single/Multi/List），中盤では Verify-Informational，後半では Verify に該当する検索のパターンが多く見られた．問題 3 で最も解答が早かった実験協力者 F のラベル付け結果を表 2 に，最も解答が遅かった実験協力者 H のラベル付け結果を表 3 に示す．ここで，前半，中盤，後半はラベルの総数を 3 分割したものである．表 2，表 3 より両者とも前半では Pinpoint（Single/Multi/List），中盤では Verify-Informational，後半では Verify に該当するクエリが比較的多く入力されていることがわかる．また，最も解答が遅かった実験協力者は中盤で Multi に該当する検索の回数が多かったり，後半で Single に該当する検索の回数が多いなど他のラベルに該当する検索が多く見られ，欲する情報を見つけるのにてまどっていることがわかる．

表 1: 全実験協力者のラベルの数

		Verify		Discover					総数
		Navigational	Informational	Navigational	Broad	Single	Multi	List	
実験1	A	1	3	1	0	5	0	0	10
	B	3	2	0	0	2	1	0	8
	C	1	3	1	3	5	1	0	14
実験2	D	2	4	0	0	2	1	0	9
	E	1	5	0	0	1	8	4	19
	F	5	26	0	2	2	6	0	41
実験3	F	2	5	0	0	0	0	3	10
	C	0	4	1	0	6	3	0	14
	B	1	6	0	1	0	1	1	10
	G	0	16	0	1	1	3	1	22
	H	3	5	0	1	10	14	0	38

表 2: 実験協力者 F のラベル付結果（実験 3）

		前半	中盤	後半
Verify	Navigational	0	1	1
	Informational	0	3	2
Discover	Navigational	0	0	0
	Broad	0	0	0
	Single	0	0	0
	Multi	0	0	0
	List	2	1	0

表 3: 実験協力者 H のラベル付結果（実験 3）

		前半	中盤	後半
Verify	Navigational	0	0	3
	Informational	0	5	0
Discover	Navigational	0	0	0
	Broad	1	0	0
	Single	0	0	10
	Multi	10	4	0
	List	0	0	0

4.2　基本検索機能の考察

我々が構築中のコンテクスト検索エンジンで想定する検索タスクは 3 節で示したものとは異なるが，既存検索エンジンと同様にユーザの検索意図を満たす機能が必要であるとの考えに基づき，3.2，4.1 節に示した結果に基づきコンテクスト検索エンジンが備えるべき基本検索機能について考察する．

構築中の動向情報を対象としたコンテクスト検索エンジンでは，入力されるクエリとしてアイテムや期間，

変動タイプが考えられる．ここでいう変動タイプとはアイテムのピーク，急激に値が変わった時，最大ピーク，最小ピーク，最初に訪れたピークなど，特徴的な動向の変化を指す．図 3 に示した既存検索エンジンにおける検索意図のラベル分類を元に，コンテクスト検索エンジンの検索意図を考慮して体系化しなおしたものを図 4 に示す．コンテクスト検索エンジンでは検索対象が Web ページではないため，既存検索における Navigational に直接対応するものは存在しない．そこで本稿では Informational はアイテムを指定しない場合，Navigational は指定した場合の検索としている．変動タイプを指定した場合は Pinpoint，指定しなかった場合は Broad とし，最大ピークの様に各動向情報に一つしか存在しない変動タイプを指定した場合は Pinpoint-Single，複数存在するものを指定した場合は Pinpoint-Multi に分類している．以下に上記のラベルに該当するコンテクスト検索エンジンでの検索例を示す．

- Informational-Pinpoint-Multi
  「2013 年に売れ始めたアイテムは？」
  入力：期間
  出力：アイテム
  変動タイプ：上昇傾向

- Informational-Pinpoint-Single
  「自転車が最も売れた時期に同様に売れたアイテムは？」
  入力：アイテム
  出力：アイテム
  変動タイプ：最大ピーク

- Informational-Broad
  「2013 年に特徴的な変動を示したアイテムは？」
  入力：期間
  出力：アイテム
  変動タイプ：指定なし

- Navigational-Pinpoint-Multi
  「自転車が急激に売れた年は？」
  入力：アイテム
  出力：期間
  変動タイプ：急激な上昇

- Navigational-Pinpoint-Single
  「自転車の生産台数が一番少なかった年は？」
  入力：アイテム
  出力：期間
  変動タイプ：最小値

- Navigational-Broad
  「自転車の生産台数の動向について知りたい」
  入力：アイテム
  出力：期間，変動タイプ
  変動タイプ：指定なし

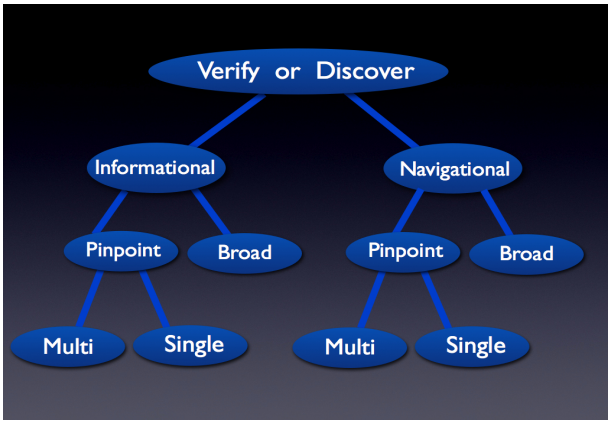


図 4: コンテクス検索エンジンにおける検索意図の分類

5　終わりに

本稿では，既存検索エンジンを用いた情報検索におけるユーザ行動を分析した結果について報告し，「動向に関する問い」を対象とした検索エンジンの基本検索機能について考察した．実験結果に基づきユーザの検索意図を分類し，該当するラベルをログデータに付与することで実験協力者の検索行動を分析した．また，ラベルを「動向に関する問い」を対象とした検索エンジンの場合に置き換えることによって，必要な基本検索機能について考察した．今後は，考察結果に基づき基本検索機能の実装を進める予定である．統計局が平成 25 年 6 月に API を公開するなど情報公開の流れもあり，今後は官公庁も含めて公開されるデータは増える事が期待されるため，動向情報を対象としたコンテクスト検索エンジンがより幅広い分野に対して有効になることが期待できる．

参考文献

[1] 旭 直人，山本 岳洋，中村 聡史，田中 克己：行動連鎖を用いた情報検索支援と Web からの行動連鎖の抽出，DEIM Forum，A7-2, 2009

[2] C. D. Manning, P. Raghavan, H. Schutze: Ch. 19: Web search basics, Introduction to Information Retrieval, Cambridge University Press, 2008.

[3] 藤田 遼治，太田 学，徳永 徹郎：ユーザのクエリ変更意図に基づく先読み検索，DEIM Forum 2012, A4-4, 2012

[4] 亀井 俊之，門田 暁人，松本 健一：WWW を対象としたソフトウェア検索エンジンの構築，電子情報通信学会技術研究報告 ソフトウェアサイエンス，Vol.102, No617, pp.59-64, 2007

[5] 加藤 恒昭，松下 光載，平尾 努：動向情報の要約と可視化に関するワークショップの提案，情報処理学会研究報告／自然言語処理研究会報告，2004(108), pp.88-94, 2004

[6] 加藤 優，桑折 章吾，高間 康史：「動向に関する問い」を対象タスクとしたコンテクスト検索の提案，人口知能学会，インタラクティブ情報アクセスと可視化マイニング研究会（第 3 回），pp.7-12, 2013

[7] 加藤 優，高間 康史：Web コンテクスト情報に基づく同時期流行アイテム検索手法の提案，FSS2012, pp.115-118, 2012

[8] 小久保 卓，小山 聡，山田 晃弘，北村 泰彦，石田 亨：検索隠し味を用いた専門検索エンジンの構築，情報処理学会論文誌，Vol.43, No.6, pp.1804-1813, 2002

[9] 南 翔太郎，岡 誠：閲覧行動モニタリングに基づく検索意図の抽出と検索結果の分類，情報処理学会報告, HCI-142(8) , pp.1-6, 2011

[10] 山田 泰寛，廣川 左千男：専門検索サイトの動的統合による次世代検索システム DAISEN における検索サイトエディタの開発，第 1 回情報科学技術フォーラム，一般講演論文集第 2 分冊，pp.11-12, 2002

対話的情報アクセスのログデータ分析

Log Data Analysis on Interacitve Information Access

加藤 恒昭 [1*]
Tsuneaki Kato[1]

[1] 東京大学
[1] The University of Tokyo

Abstract: The characteristics of user behaviors in explorative information access are reported, which reflect the differences of the environments she uses and the tasks she engages in. Using a model of information access behaviors and a log data coding based on that model, the analysis was conducted on the log data obtained in VisEx, an experiment for evaluating interactive and explorative information access environments. It shows that introduced retrieval methods, narrowing-down and similarity-based retrieval, are used as a substitute of sequential document checking, and those effectiveness differs depending on task characteristics.

1　はじめに

利用者が必要とする情報を獲得するために行う情報アクセスは，多くの場合，一連の行為からなる対話的・探索的な過程となる．優れた情報アクセス環境は，利用者が最初に持つ情報要求に適切に答えるだけでなく，その後の過程の中の様々な場面で利用者を支援できる必要がある．情報アクセスの過程の中で利用者がどのように振る舞ったかを記録した情報アクセス行為のログデータは，そのような情報アクセス環境を構築するヒントを与えてくれると期待できる．

本稿では，探索的情報アクセス環境の評価実験である VisEx[3] を通じて得られたログデータの分析を行い，対話的・探索的な情報アクセスにおける利用者の振る舞いと環境や課題との関係について報告する．環境との関係では，キーワード検索を主たる情報アクセス方法とするシステムと，それに加えて簡単なファセット検索と類似検索の機能を持つシステムとで，利用者の情報アスセス行動にどのような差が現れるかを示す[1]．課題との関係では，VisEx で実施された２種類の課題，イベント収集課題とトレンド要約課題，の違いが利用者の振る舞いにどのように影響しているかを報告する．

これらの具体的な分析とあわせて，分析の方法論として，情報アクセス行為のモデルと，それに従ったログデータのコーディングを提案する．これらのモデルやコーディングは，利用者実験で得られるログデータ分析のケーススタディであり，多くのログデータ分析の参考になることを期待している．

本稿の構成は以下の通り．まず２章で情報アクセス環境評価実験 VisEx と，そこに参加し本稿の分析対象となった情報アクセス環境について概説する．３章ではログデータ分析の方針として，情報アクセス行動のモデルとそれに基づいたコーディングを説明する．４章で得られた分析結果を報告し，５章でそれについて議論を行う．６章で全体をまとめる．

2　情報アクセスの設定と環境

2.1　情報アクセス環境評価実験 VisEx

VisEx は，探索的な情報アクセスの環境を評価する枠組みを検討する試みである．情報アクセス技術に関する評価ワークショップ NTCIR-9[2]のパイロットタスクとして 2011 年に実施された利用者実験では，探索的な情報アクセス課題を参加者が提出した様々な環境の下で利用者に実施させ，その成果として得られるレポートや，利用者の情報アクセス行動のログデータの収集を行っている．

VisEx の特徴のひとつは，評価される情報アクセス環境に図 1 に示した構成を仮定し，そのうち，核部のみを参加者が作成・提出し，それ以外は環境間で共通化するという枠組みにある．実際の検索を行う情報検

*連絡先：東京大学大学院総合文化研究科言語情報科学
〒 153-8902 東京都目黒区駒場 3-8-1
E-mail: kato@boz.c.u-tokyo.ac.jp

[1]本稿では情報アクセス環境と情報アクセスシステムをほぼ同じ意味で用いる．利用者の情報アクセスを様々に支援するということで「環境」という用語を用いるが,「環境への入力」等の表現は違和感があるので必要に応じて「システム」と呼ぶようにする．

[2]http://research.nii.ac.jp/ntcir/index-ja.html

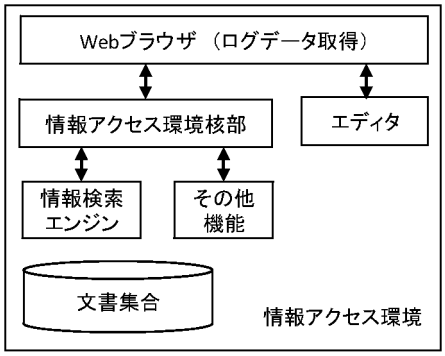


図 1: VisEx における情報アクセス環境

索エンジンや集められた情報の編集や記録に用いられるエディタ部分が共通となる．核部を含めてこれらすべてが Web ブラウザの下で動作し，利用者はすべてのインタラクションをブラウザ経由で行う．エディタは，Web ブラウザの機能拡張（アドオン）として開発されたもので，これが Web ブラウザとエディタに対する操作のログデータ取得機能を持っている [4]．

このような枠組みとすることで．(1) 情報アクセス行為全体，つまり，情報を集めるだけでなくそれを知識としてまとめ上げる部分までを観察しつつ，そのような広い行為の観察に伴う揺れをできるだけ少なくする，(2) 情報アクセス環境を用いる利用者の振る舞いに加えて，その中の構成要素間のやりとりについても統一的かつ詳細なデータを取得する，ことが目指されている．

2.2　課題

VisEx では，利用者は与えられた環境で，与えられたトピックについての情報を文書集合の中から探し出して，収集し，それをレポートにまとめる．文書集合は毎日新聞の 1998-2001 年の記事を用いている．課題は，TREC の Interactive track[2] を参考に作成された．以下の 2 種類の課題が実施された．

イベント収集課題　トピックとして与えられた出来事の特徴，発生日時や発生場所等，を収集して，まとめる．トピックは，E1:アジアでの航空機墜落事故，E2:日本で起きた原子力発電所関連の事故，他計 4 件を用いた．

トレンド要約課題　時系列統計情報が関連する社会や経済の状況をトピックとして与えられ，関連する統計量の変化（動向）とその原因や影響を要約する．トピックは T1:ガソリンを巡る状況，T2:内閣の評価，他計 4 件を用いた．

課題においてこれらのトピックは，まとめるべき特徴や統計量を含めて利用者に提示される．例えば，E1「アジアでの航空機墜落事故について，発生日時，場所，事故状況，航空会社名を含む航空機の種類，死傷者数，原因を調べてください」，T1「ガソリンを巡る状況の調査として，ドバイ原油価格とレギュラーガソリン価格の変化を調べてください」のように提示される．

2.3　情報アクセス環境

2011 年に実施した VisEx 利用者実験には，オーガナイザが用意したベースラインシステムに加えて，4 システムが参加した．本稿ではそのうち，ベースラインシステム (BL) と UTLIS システム (UT) を対象にログデータの分析を行う．

BL システムは，一般的な Web 検索エンジンと同様に，基本的なキーワード検索機能と並べ替え機能を持ち，キーワード検索の結果として，10 件の記事の見出しとスニペットのリストからなる結果ページ[3]を返す．利用者はいずれかの見出しをクリックすることでその記事の内容を表示した文書ページに移動する．並べ替えは，日付と適合度の昇順と降順が選択できる．結果ページでは，次の 10 件，前の 10 件，指定した位置の 10 件の表示も指定できる．

UT システムは，BL システムに，記事の内容が関連する場所や記事の発行日を指定することで検索結果を絞り込む，絞り込み検索機能，指定した記事と類似した記事を検索する類似検索機能を追加したシステムである [5]．

追加されたふたつの機能は一般的なものであるだけでなく，次の点で興味深い．記事が関連する場所や時間での絞り込みは，ファセット検索の簡単な実現であると同時に地図やタイムスライダ [6] を用いた可視化インタフェースへの発展が考えられる．UT システムはそのような視覚的インタフェースを実現したものではないが，視覚的インタフェースと従来のキーワード検索がどのように組み合わされて使われるかの示唆が期待できる．類似検索は，インデクスを介さずに文書と文書を直接関係づけるという点で，文書集合にネットワーク構造を付与し，Web 文書のハイパーリンクによるのと同様のブラウジングを可能とするので，キーワード検索とブラウジングの関係についての示唆が期待できる．

Bate は情報に至るまでの人間の探索過程を調査し，いわゆる情報検索におけるキーワードを用いた文書検索とは異なる，様々な手法が用いられていることを明らかにした [1]．情報アクセス環境を利用した文書検索においても，ファセット検索的な絞り込みとブラウジングという手段の追加が情報アクセスの過程にどのよ

[3]Web 検索エンジンにおける SERP に相当する．

うな影響を与えるかが本稿でのログデータ分析の動機のひとつである．

2.4　利用者実験

1 環境 1 課題につき，それぞれ異なる 5 人の利用者が同じ順序で各 4 トピックについて実験を行った．ひとつのトピックに与えた時間は 50 分（3000 秒）である．練習トピックを用いて課題と環境に慣れる時間をとり，実験全体の前後と各トピックの実施後にアンケートを行っている．今回の分析では，2 環境× 5 人× 2 課題× 4 トピックの 80 件のログデータを分析した．

3　ログデータ分析の方針

情報アクセスは情報アクセス環境に対する一連の行為として実現される．ログデータに記録されるのは，それらの一部である．例えば，結果ページや文書ページの閲覧という行為そのものはログには記録されず，記録されるのは，そのページの表示開始という動作だけとなる．これらの記録されない行為を推測する必要がある．またログに記録されるような行為，例えば，文書ページの表示開始や絞り込み検索の実行も，実際に記録されるのは，特定の場所でのクリックやあるボタンの押下というブラウザへの動作であり，それらが情報アクセス環境としてどのような意味を持っているかの理解には，一定の解釈が必要になる．このような推測と解釈をコーディングと呼ぶ．

3.1　基本的なモデル

分析の対象となっている環境において基本的な情報アクセスは以下のように進められると考えられる．

まず，核部（検索条件入力・結果表示・文書表示が行われる）とエディタがそれぞれタブに割り当てられており，タブ選択によって核部の画面が表示されている．そこで，

1 例えばキーワードを入力し検索ボタンを押下することで，検索が実施され，結果ページが表示される．

2 得られた結果ページを閲覧する．

3 必要な情報が含まれていそうな文書を見つけると，その文書の見出しをクリックすることで文書ページを表示する（に移動する）．

4 文書ページを閲覧し，必要な情報を探す．

5 文書中に実際に必要な情報があれば，タブ選択によりエディタに移る．

6 タブ選択によって文書ページとエディタを行き来して，文字入力や削除，コピーやペーストによりレポートを作成する．

7 必要な情報をレポートにまとめた時点で，文書ページから後退（back ボタン押下）によって結果ページに戻る．

5’ 文書中に必要な情報がないと判断された場合は後退によって結果ページに戻る．

8 結果ページの閲覧を続ける．

9 結果ページ全体を閲覧し終わると，「次の 10 件」のクリックや新しいキーワードを用いた検索を行うことで，新しい結果ページを表示する．

充分な情報をまとめあげるまで，あるいは制限時間となるまでこれが繰り返される．

また，この間，どのページからでも必要に応じて，メニューによる指定等を用いて，外部ページ（核部画面とエディタ以外のページ）へ移動することができる．複数の外部ページをクリック等で移動し，その後，後退により戻ることになる．

この過程を，状態と，行為の実施による状態間の遷移としてモデル化する．状態とは，利用者が意味的にまとめられる一連の行為を行っている（と推測あるいは解釈される）期間をいう．ここで，行為はその解釈によって状態の一部であったり，状態遷移を引き起こすものであったりする．

状態として以下の 4 つをおく．

結果閲覧　結果ページを閲覧し，必要な情報を含んでいると思われる文書が存在するか，それがどれかを判断している．

文書閲覧（正の閲覧，負の閲覧）　文書ページを閲覧し，その文書が必要な情報を含んでいるかを判断している．このうち，含んでいるという判断に至った閲覧を正の閲覧，そうではない閲覧を負の閲覧とする．

情報編集　文書中から必要な情報を抜き出し，エディタを利用してそれをレポートにまとめあげている

外部閲覧　外部のページにアクセスし，そこで情報を収集している，

これに基づいて，上記の情報アクセスの過程は以下のように解釈される．

1 結果閲覧への状態遷移 → 2 結果閲覧 →
3 文書閲覧への状態遷移 → 4 文書閲覧 →
5 情報編集への状態遷移 → 6 情報編集 →
7 結果閲覧への状態遷移 →
（5’ 結果閲覧への状態遷移 → ）
8 結果閲覧 → 9 結果閲覧への状態遷移

定義された状態はほぼ表示されているページの種類に対応する．逆に言えば，表示されているページから状態を推定している．ただ，必ずしも一対一に対応している訳ではなく，情報編集は特定の文書ページとエディタとを行き来して文書内容を参照してレポートを作成している期間すべてで，その間に行われている文字入力やコピーやペースト等を含んでいる．一方で，文書閲覧は結果ページから文書ページに移動して，最初にエディタに移るもしくは結果ページに戻るまでとしており，文書ページの表示と閲覧はその状況によって２種類の状態に分類される．

状態遷移を引き起こす行為はその遷移によってのみ特徴づけられるが，結果閲覧から結果閲覧への遷移を引き起こす行為だけは，検索関連行為と呼び，それを更に分類する，主な検索関連行為は以下の通り．

- キーワード検索（KW 検索）
- 順序指定項目や昇順降順の変更（順序変更）
- 次 10 件の表示（次 10 件）
- 絞り込み検索
- 類似検索

なお，後退/前進による遷移はそれによって何が行われたに基づいて分類する．例えば，次 10 件の表示によって遷移した結果閲覧からの後退であれば前 10 件の表示に相当するとする．ただし，後述のタブ選択と同様，キーワード検索の後退は検索関連行為としない．

本分析では，これらの状態への滞在と遷移，そして検索関連行為に基づいてコーディングを行う．

このモデルでは，例えば以下が捨象されている．テキスト入力フィールドへの入力やラジオボタン等の選択に要する時間（これにはキーワード検索に用いるキーワードを工夫する時間も含まれる）や，情報編集に遷移する直前の文書内容のコピーに要する時間等は関連する状態に含まれてしまう．結果閲覧から情報編集への遷移はないものとされている．キーワードを工夫する行為やその時間は，情報アクセス環境の設計において重要なものであろうが，残念ながら，現在のログデータからはそれを得ることはできない．

3.2　モデルの緩和

前節で述べた基本的な情報アクセスでは，核部とエディタがそれぞれタブに割り当てられており，ひとつの核部の画面が結果ページ，文書ページの間を遷移する．実際には，環境のベースとなる Web ブラウザがタブブラウザであるため，３つ以上のタブを開き，複数の核部を行き来することが可能である．この場合，文書閲覧から文書閲覧への遷移，つまり複数の文書の見比べ，や結果閲覧から結果閲覧への検索関連行為によらない遷移，ある検索結果を残しておいて，別の検索を行うような検索行為の埋め込み等，が可能となる．

このような情報アクセスをモデル化として，上記の基本的な情報アクセスを表現する状態遷移のネットワークをタブにあわせて複数個用意しその間での状態遷移を考えることもできるが，本稿のモデル化では，状態遷移の制約を緩くしたひとつの状態遷移ネットワークに基づき，文書閲覧から文書閲覧への遷移等を許し，結果閲覧の間の遷移にタブ選択によるものを加えることとした．ここで，タブ選択における文書閲覧の遷移は，キーワード検索の後退と同じく，以前の状態への復帰と考え，検索関連行為とは別に扱う．この枠組みは複数の候補を比較し，それらの中から最もよい文書を選び出すような課題における利用者の行動を正しく表現しないことが危惧されるが，正の閲覧や負の閲覧等の文書を探し出すという情報アクセス行動の表現には充分と考える．

3.3　コーディングの実際

コーディングは以下の手順で行われる．コーディングはプログラムによって行った．

複数行為のまとめあげ　ログ取得機構が出力する粒度の細かい複数の動作をまとめる．例えば，一文字毎に記録される文字入力を文字列の入力にまとめる．文字の削除やスクロール等も同様にまとめあげる．ひとつの操作に対して得られる複数の記録をまとめて行為を分類する．例えば，後退動作を行うと，クリック動作，後退動作そのもの，URL 変更の 3 つの動作がログに記録される．これらを，戻り先 URL への後退動作としてまとめる．

状態を構成する行為のまとめあげ　コピーや文字列入力等の編集関連の操作をまとめあげる．スクロール等，コーディングに不要な行為を削除する

行為の解釈　検索関連行為について分類を行う．ログに記録されるのは，ボタンの押下もしくはクリックと検索等実行時のパラメータの列であるので，前回の差分から新しいキーワード列による検索の実行なのか，次の 10 件の表示なのか等を解釈する．

表示ページの解釈　状態を判定する前段階として，表示されているページが結果ページ，文書ページ，エディタ，外部ページのいずれであるかを判断する．

後退/前進とタブの管理　実行された動作をタブ毎に記録し，後退/前進においてどの動作が取り消されたのか/再実行されたかを明らかにする．その

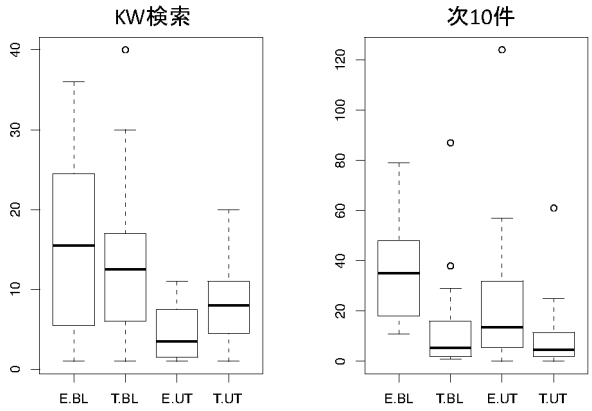


図 2: KW 検索と次 10 件の頻度

時点で表示されるページをタブ毎を管理し，タブ選択によってどのページが表示されたかを明らかにする．

状態の推定　連続した外部ページの表示を外部閲覧状態としてまとめる．文書ページ，エディタが表示されていてる期間を必要に応じてまとめあげ，文書閲覧と情報編集の状態に分類する．文書閲覧についてはその後に続く状態により正の閲覧と負の閲覧に分類する．

4　分析

4.1　検索関連行為と状態

コーディングされたログデータについて，環境と課題を要因として分析を行った．まず，ひとつのトピックに関する情報アクセス行動の中での，検索関連行為の頻度を調査した．検索関連行為のうちで中央値が 0 でなかったふたつ，KW 検索と次 10 件の分析を表 1（1,2 行）に示す．左側に中央値と IQR（四分位範囲）を，右側に分散分析の結果を示している．図 2 はその箱ひげ図である．これらからは，KW 検索の頻度は環境間の差（主効果）が有意，次 10 件は課題間の差が有意となる．箱ひげ図を眺めると幾つかはずれ値が存在することがわかる．この影響を排除するために，KW 検索について最大値ひとつ，次 10 件について最大値とその次のデータを除いて行った分析の結果が表 1（3,4 行）である．次 10 件でも環境間の差が有意となり，KW 検索，次 10 件の両方で交互作用が現れる．

表 1（5,6 行）には，UT システムにおける類似検索と絞り込み検索の実施頻度を示した．イベント収集課題において，類似検索は，トレンド要約課題と比較して有意に多く用いられている（$t(26.242) = 5.53, p < 0.001$）．

表 1（7,8 行）には，正閲覧文書数と負閲覧文書数を示した．環境，課題の差が共に有意で，負閲覧文書数には交互作用も見られる．

表 2 は，ひとつのトピックに関する情報アクセス行動の中での各状態への滞在時間（秒数）について，中央値と IQR，および分散分析の結果を示したものである．結果閲覧，負閲覧，情報編集のいずれでも，課題の差が有意である．

まず，課題の特徴として，トレンド要約課題は，イベント収集課題に比べて，検索関連行為の頻度が少なく，結果閲覧時間が短い．そして正閲覧文書数が少なく，情報編集時間が長い．このことからこの課題が，比較的見つけ出しやすい少数の適合文書から多くの情報をまとめあげるという傾向を持つと推測される．環境の差としては，類似検索と絞り込み検索を提供している UT システムで，KW 検索と次 10 件の実施頻度が減少していることがわかる．KW 検索と次 10 件に代わる情報アクセス手段として類似検索と絞り込み検索が利用されていると自然に考えることができる．この差は特に検索関連行為頻度の高いイベント収集課題で顕著である．

図 3 に 2 つの環境の 2 つの課題の各トピックから任意に 1 件を選んだ計 16 件の情報アクセス行動について，累積情報編集時間の変化を示す．いずれの場合でも情報編集時間は制限時間の後半でも増加しており，適合する文書を見つけ終わってしまった状況ではないことが見てとれる．このことと，環境間で結果閲覧と情報編集の時間に有意な差が見られなかったことから，必要な情報を見つけ出す効率という点ではふたつの環境に大きな差がないことが推測される．これに対して，正閲覧文書数が UT システムで有意に少ないことに明確な説明は与えられない．情報編集時間が得られた情報量に比例するとすれば，少ない文書から同じ量の情報を得ていたということで，UT システムがより適切な文書を見つけ出せているという可能性等も考えられるが，それらについては，利用者のレポートの分析等と組み併せて検証する必要がある．

4.2　検索関連行為の内訳

前節での分析は，KW 検索と次 10 件に代わる情報アクセス手段として類似検索と絞り込み検索が利用されていることを示唆していた．検索関連行為の内容についてふたつの分析することで，この点をより詳細に検討する．

第一の分析では，KW 検索をそこで用いられるキーワードによって分類する．与えられたトピック全部が適合するようなキーワードを全体 KW，与えられたトピックのうちの一部，特定のイベント等，が適合する

表 1: 検索関連行為頻度と閲覧文書数

	中央値 (IQR)				分散分析 F(1,76) (p 値)		
	BL		UT		主効果		交互作用
	イベント	トレンド	イベント	トレンド	環境	課題	環境:課題
KW 検索	15.5(18.5)	12.5(11)	3.5(5.5)	8(6.25)	**21.51(<.001)**	0.00(.96)	3.48(.07)
次 10 件	35.5(29.75)	6(13.5)	14(24.75)	4.5(9.25)	2.75(.10)	**13.94(<.001)**	0.50(.48)
KW 検索*	15.5(18.5)	12(9.5)	3.5(5.5)	8(6.25)	**20.92(<.001)**	0.10(.76)	**5.69(.02)**
次 10 件*	35.5(29.75)	6(12.5)	12(24)	4.5(9.25)	**6.41(<.01)**	**22.98(<.001)**	**4.94(.03)**
類似検索			8(4.25)	0(1.25)			
絞込検索			11(12.5)	12(10.5)			
正閲覧	15.5(6.25)	10.5(8.25)	13(2.75)	10.5(4.25)	**4.44(.04)**	**10.48(<.01)**	0.45(.51)
負閲覧	26.5(13.25)	9.5(9.0)	17(6.5)	8(9.5)	**10.01(<.01)**	**46.91(<.001)**	**5.76(.02)**

KW 検索*と次 10 件*ははずれ値を除いた分析

表 2: 状態滞在時間

	中央値 (IQR)				分散分析 F(1,76) (p 値)		
	BL		UT		主効果		交互作用
	イベント	トレンド	イベント	トレンド	環境	課題	環境:課題
結果閲覧	868.5(333)	708.5(418)	1004.5(464)	634.5(313)	0.00(.99)	**12.85(<.001)**	0.76(.39)
正閲覧	210(191)	255(174)	212(101)	165(110)	1.99(.16)	1.08(.30)	2.12(.15)
負閲覧	263.5(215)	170(169)	195 (155)	126.5(142)	0.44(.51)	**12.38(<.001)**	0.42(.52)
情報編集	1558.5(614)	1877(914)	1423(443)	1917(334)	0.05(.82)	**13.92(<.001)**	0.06(.81)

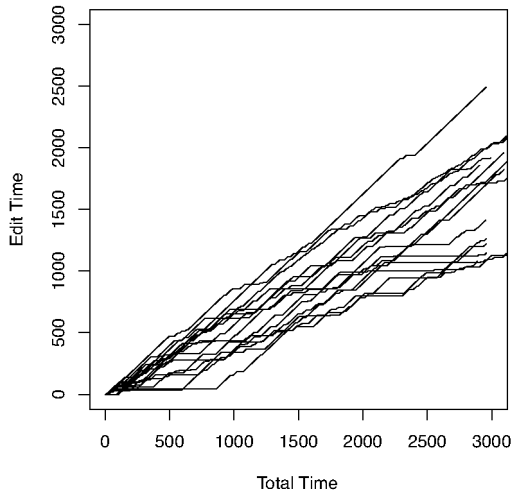


図 3: 累積情報編集時間

キーワードを特定 KW とする．E1 アジアでの航空機墜落事故であれば,「アジア」「航空機」「墜落」等が全体 KW,「中正国際空港」「ヘリコプター」「12 月 10 日」等が特定 KW となる．KW 検索を，すべてが全体 KW であるような検索と特定 KW を含む検索とに分類して，その実施頻度を調査した．結果を表 3 に示す．UT システムでは，課題によらず，特定 KW の検索が減少していることがわかる．全体 KW による検索はトレンド要約課題では UT システムの場合の方が多くなっており，環境と課題の交互作用が見られる．

第二の分析として，それぞれの検索手段の貢献の大きさを確認するために，正文書閲覧の直前の検索関連行為の分布を調査した．どのような検索関連行為が適合文書発見に繋がることが多かったかを求めている．中央値と IQR を表 4 に示す．BL システム，特にイベント収集課題では，適合文書が次 10 件に続いて得られることが多い．なお，この数字は，ある行為を行った後に適合文書が得られる割合ではなく，得られた適合文書がどの検索行為の結果なのかの割合を反映したものであり，BL システムにおけるイベント収集課題では，特に頻繁に次 10 件が実施されているので，この結果は次 10 件が効率的であることを示しているわけではない．UT システムにおいては，絞り込み検索と類似検索の結果として適合文書が得られることが多い．イベント収集課題では類似検索が，トレンド要約課題では絞り込み検索が貢献している．

絞り込み検索によって特定 KW による検索と似た結果を得ることができるので，UT システムにおいて，特定 KW による検索が，絞り込み検索に代替され，減少したことは至極当然であり，絞り込み検索が特定 KW による検索よりも好まれ，活用されたことを示唆している．全体 KW については，課題の違いが影響している．UT システムにおけるトレンド要約課題では，トピック全体への絞り込みは全体 KW による KW 検索，そこか

表 3: 全体 KW と特定 KW

	中央値 (IQR)				分散分析 F(1,76) (p 値)		
	BL		UT		主効果		交互作用
	イベント	トレンド	イベント	トレンド	環境	課題	環境:課題
全体 KW	3.5(7.5)	3(4)	2(3)	4.5(5.5)	1.62(.21)	0.03(.87)	**8.8(<.01)**
特定 KW	9(10.5)	6(11.25)	0.5(2.25)	2(4.75)	**25.70(<.001)**	0.18(.67)	0.48(.49)

表 4: 正文書閲覧に貢献した検索関連行為

	中央値 (IQR)			
	BL		UT	
	イベント	トレンド	イベント	トレンド
KW 検索	3 (7.25)	7.5 (5.25)	0 (1)	2.5 (3.25)
次 10 件	8.5 (8.5)	3.5 (5.25)	2 (6.25)	1 (2)
順序変更	1.5 (2.25)	0 (1)	0 (0.25)	0 (1)
類似検索			5.5 (7)	0 (2)
絞込検索			1.5 (4.5)	5.5 (8.5)

らの絞り込みは時間を指定した絞り込み検索という役割分担が行われたことが原因で，その両方を KW 検索で行う BL システムよりも全体 KW による検索が増えたと考えられる．イベント要約課題では，全体 KW による検索も UT システムの方が少ない．この理由は明らかでないが，検索結果文書数が多くても絞り込み検索があるので，KW 検索での工夫を行わなかった，UT システムの絞り込み検索はその直前の KW 検索結果に復帰することが容易であるのに対し，BL システムで後退等を用いて以前の検索結果に戻るのはやや面倒なので，同じキーワード列を再度入力してしまうこと等が回数の増加に繋がった等が理由として考えられる．なお，KW 検索は，以前の結果表示ページに後退やタブ選択で戻った場合は検索が行われたとは数えないが，以前と同じキーワードを改めて入力して検索した場合は新しい 1 回と数えている．

特定 KW の位置づけは課題によって異なる．トレンド要約課題で使われる特定 KW の一部は時期の指定である．一方，イベント収集課題では，地名や施設名が中心となる．大きな違いとして，前者はトピックを得た時点で利用者に明らかであるのに対し，後者は関連する文書を見つけるまで利用者はこれが適合キーワードだと明らかでないことがあげられる．つまり,「中正国際空港」や「東海村」がトピックとなっているイベントに関連することは適合文書を少なくともひとつ見つけた時点ではじめて利用者に理解される．このことは，BL システムにおける特定 KW の検索が絞り込み検索と似ているだけでなく，類似検索の役割も果たしていることを示している．逆に言えば，類似検索はブラウジングを提供するだけではなく，KW 検索の代替とも捉えることができる．

第二の分析の結果からは，BL システムによる情報アクセス行為が KW 検索を用いて適合情報だけを絞り込むというより，KW 検索は準備的な限定に用い，そこから先はその結果を順に眺めることで利用者自身が判断して選び出す形で行われていることが推察される．このパタンは，特に特定 KW が事前に明らかでないイベント収集課題において顕著である．UT システムにおいては，全体 KW による KW 検索の後の結果を順次眺めていく行為や，特定 KW による絞り込みや類似検索相当の検索が，絞り込み検索や類似検索に置き換えられているのが見てとれる．ふたつの課題における絞り込み検索と類似検索の貢献の違いは，上で述べた特定 KW の位置づけの違いに基づくと推察される．

5　考察

ログデータ分析を通じて，課題の特徴とそれを反映した情報アクセス環境の利用のされ方がある程度まで明らかにできた．レポート作成という課題は，イベントの列挙であれ，トレンドの要約であれ，あるトピックについての情報を網羅的に必要とするので，情報アクセスは再現率を重視したものとなり，そのため，まず，トピックに関する文書を含んでいるであろう集合を大きく選び出し，その後に更に適合文書の選択が行われている．後者の選択は利用者自身が閲覧を通じて順次判断することで行われる場合も多い．絞り込み検索や類似検索も後者の選択を支援し，その効果は選択のためのキーワードが事前に明らかでないような課題で有益である．その意味では同じレポート作成でも課題のこの特徴が必要な情報アクセス手段に影響を与える．

つまり，このような課題では，大きく選び出された

文書の集合を整理して俯瞰的に表示するような結果表示ページや，その集合を更に絞り込む手段を利用者に提示するような仕組みが，検索そのものの高度化，例えばキーワード検索の精度向上，よりも重要となると考えられる．今回の絞り込み検索や類似検索はその簡単な例でしかないが，対話的な情報アクセスに関する技術がいわゆる伝統的な情報検索の技術とは異なる方向性を持つことを示唆している．

この知見はレポート作成という課題の特徴を反映したものである．利用者が既に充分に理解している難解な質問の回答を含んだ文書をひとつ見つけ出すというような課題では状況は全く異なる．また，情報編集に要する時間が情報アクセス行為の半分以上を占めるという状況もレポート作成という課題の特徴であろう．この点でも課題の特徴は情報アクセス環境の設計に大きく影響する．

ログデータ分析を通じて以上の洞察が得られたと述べたが，得られてしまえば，それは内省的な熟考によっても明らかにできたようにも思われる．その意味では当たり前のことしか示されなかったわけで，今回の分析がログデータ分析の有効性を充分示したとは言い難い．更にモデルの設計やコーディングはかなり手間のかかる作業である．今回のモデルは結局単純なものとなってしまっているにもかかわらず，それに基づくコーディングの設計にはかなりの時間を要した．特定のコーディイングを前提としたログ取得となっていなかったこと，利用者の行為のバラエティが多いことがその理由である．今回の経験を活かして，より安価で有効なログデータ分析が行えるのかを検討する必要がある．

6　むすび

探索的情報アクセス環境の評価実験である VisEx を通じて得られたログデータの分析を行い，課題の特徴とそれを反映した情報アクセス環境の利用のされ方を明らかにし，情報アクセス環境設計の示唆を得た．一方で，ログデータ分析の難しさも明らかになった．まだ行われていない分析も多い．VisEx で得られたデータの中でも，利用者のレポートの内容の分析と突き合わせることでログデータを別の視点から眺める必要がある．また，キーワード検索におけるキーワードの変化の追跡，キーワード検索の結果からのページ推移の分析で興味深い結果が得られている [7] ので，そのような研究とも比較検討ができればと思うが，今回の課題とデータ数では実現は難しかった．更に，考察で述べたように対話的・探索的情報アクセスにも様々なタイプの課題があるはずで，その整理も行い，それぞれの特徴と情報アクセス環境との関係を明らかにしていきたい．

謝辞

VisEx は筆者と松下光範氏（関西大学），上保秀夫氏（筑波大学）によって運営された．VisEx をタスクに加えていただいた NTCIR は神門典子氏（国立情報学研究所）を中心に運営されている．また，高間康史氏（首都大学東京）は VisEx に参加いただいた．これらを通じての貴重な議論や頂いた数々のアドバイスに心より感謝する．本研究は基盤研究 (B)「視覚情報を活用した対話的情報アクセスのための情報編纂研究基盤の構築」の一環として行われてる．

参考文献

[1] M.J. Bates: The Design of Browsing and Berrypicking Techniques for the Online Search Interface. in *Online Information Review*, Vol. 13, No. 5, pp. 407–424, 1989.

[2] S.T. Dumais and N.J. Belkin: The TREC Interactive Tracks: Putting the User into Search. in E.M. Voorhees and D.K. Harman ed. *TREC Experiment and Evaluation in Information Retrieval*, pp. 123-152, The MIT Press, 2005.

[3] 加藤恒昭，松下光範，上保秀夫: VisEx 予備実験報告. 第 5 回情報編纂研究会, 2011.

[4] 加藤恒昭，松下光範，上保秀夫: NTCIR-9 VisEx の概要. 第 7 回情報編纂研究会, 2011.

[5] 加藤恒昭: . 情報アクセス過程に対する検索手段増加の影響 － VisEx での実験－ 第 7 回情報編纂研究会, 2011.

[6] B. Shneiderman: Dynamic Queries for Visual Information Seeking. in *IEEE Software* Vol.11, No.6., pp. 70–77, 1994.

[7] R.W. White and J. Huang: Assessing the Scenic Route: Measuring the Value of Search Trails in Web Logs. in *The Procs. of SIGIR'10*, pp. 587–594, 2010.

テキスト分類のための潜在トピックを考慮したグラフ構成

Latent Topic-based Graph Construction for Text Classification

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

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

Abstract: This paper aims to raise the accuracy of multi-class text classification by means of graph-based semi-supervised learning (GBSSL). It is essential to construct a proper graph expressing the relation among nodes in GBSSL. We propose a method to construct a similarity graph by employing both surface information and latent information to express similarity between nodes. Experimenting on Reuters-21578 corpus, we have confirmed that our proposed method works well for raising the accuracy of GBSSL in multi-class text classification task.

1　序論

機械学習手法は，教師あり学習，教師なし学習，半教師あり学習などがある．半教師あり学習 (Semi-Supervised Learning: SSL) 法は，少量のラベルありデータを用いて，多量のラベルなしデータに付与するラベルを予測する手法である．その中でも，グラフ構造に基づく半教師あり学習 (Graph-Based Semi-Supervised Learning: GBSSL) 法は，文書分類タスクにおいて，Support Vector Machine (SVM)[2] などの学習法と比べてより有効な手法であることが知られている [4].

GBSSL 法の精度は，一方でどのような教師データ (ラベルありデータ) を与えるかによって左右され，他方で，どのようなグラフを構成するかによって左右されることが分かっている [7, 9]．前者に関連して重要となるのが，どのようにして情報量の大きいデータを選出するかである．その良い事例が能動学習法であり，質の高い教師データを選出するための方法である．GBSSL 法の精度を改善するため，いくつかの能動学習法が提案されている [7, 10]．また，後者に関連して重要となるのは，グラフのノード間の関係性をどのように表現するかである [9]．一般に，GBSSL 法のグラフスパース化手法には，k-近傍グラフが用いられることが多い．しかしながら，k-近傍グラフではその構成上，ハブ点と呼ばれる高次数のノードができやすく，このハブ点は GBSSL 法の精度を悪化させるということが報告されている [11]．ノードに次数制約を設けた，グラフスパース化手法もまたいくつか提案されている [11, 12].

*連絡先： お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース
〒 112-8610 東京都文京区大塚 2-1-1
E-mail: g0920506@is.ocha.ac.jp

本研究では，GBSSL 法を用いた多クラス文書分類におけるグラフ構成手法の提案を行う．グラフ構成において，必須の要件であるノード間の類似度に，文書間の潜在的な類似度を新たに取り入れる．一般にこれまで，テキストデータから構成されるグラフにおいては，単語の頻度情報に基づく文書間の表層的な類似度が多く採用されてきたが，我々はこれに加えて新たに，確率的言語モデルに基づく文書間の潜在的な類似度を加えたものをノード間の類似度として採用する．また，これら表層的な類似度と潜在的な類似度を $(1-\alpha):\alpha$ $(0 \leq \alpha \leq 1)$ の割合で混合させ，α をパラメータとして動かし，両情報を同時に採用する．

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

2　文書分類のための GBSSL 手法

本研究で提示する，多クラス文書分類のタスクにおける GBSSL 法の詳細は，以下に述べる通りである．

2.1　グラフ構成

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

で V と E は，それぞれグラフのノード集合と辺集合を表す．

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

2.2 グラフにおける類似度

テキストデータにおける文書間の類似度を測る指標として，表層情報に基づく類似度と潜在情報に基づく類似度の二種類の類似度を採用する．文書の表層情報としては，文書に含まれる単語の出現頻度に着目した *tfidf* ベクトル [3] が多く用いられる．ここでは，表層情報に基づく類似度 ($sim_{surface}$) を，*tfidf* ベクトルのコサイン類似度の値とする．また，文書の潜在情報として，複数文書内に隠れトピックが存在することを仮定し，その隠れトピックに関して生起する単語の確率分布 (トピック分布) を用いる．ここでは，潜在情報に基づく類似度 (sim_{latent}) を，シグモイド関数 (式 (4))) を用いて，トピック分布間の距離を類似度に変換したものとする．トピック分布間の距離は $L2$ ノルム距離 (式 (5)) を用いて求める．トピック分布の推定には，Latent Dirichlet Allocation (LDA) 法 [1] を用いる．

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

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

$$sim_{surface}(S,T) = \cos(tfidf(S), tfidf(T)) \quad (2)$$

$$sim_{latent}(P,Q) = \frac{2}{1+\exp^{L^2(P,Q)}} \quad (3)$$

$$\sigma_1(x) = \frac{1}{1+\exp^{-x}} \quad (4)$$

$$L^2(P,Q) = \int (P(\boldsymbol{x}) - Q(\boldsymbol{x}))^2 d\boldsymbol{x} \quad (5)$$

2.3 ラベル伝搬法

本研究における GBSSL 法として，ラベル伝搬法 [5, 8] を採用する．ラベル伝搬法は，「グラフ上において，辺で繋がるノード同士は同じカテゴリに属す」という仮定に基づき，カテゴリラベル未知のノード (すなわち，テストデータ) について予測を行う手法である．

類似度行列を $\boldsymbol{W}$，ノード数を n 個 (このうち教師データ数は l 個) とする．n 個のノードに対する予測値 $\boldsymbol{f}$ は，以下の最適化問題の目的関数 (式 (6)) の解 (式 (8)) として求まる．式 (6) の第 1 項は，各ノードの予測値と教師データの正解値の差を表し，第 2 項は，類似度グラフ上で隣接するノード同士の予測値の差を表す．$\lambda (> 0)$ は両項のバランスをとる定数である．

式 (6) は $\boldsymbol{L}$ を用いて，式 (7) と変形できる．$\boldsymbol{L} (\equiv \boldsymbol{D} - \boldsymbol{W})$ はラプラシアン行列と呼ばれ，対角行列 $\boldsymbol{D}$ は $\boldsymbol{W}$ の各行 (又は列) の和を対角成分に持つ行列である．

$$\begin{aligned} J(\boldsymbol{f}) &= \sum_{i=1}^{l} (y^{(i)} - f^{(i)})^2 \\ &\quad +\lambda \sum_{i<j} w^{(i,j)} (f^{(i)} - f^{(j)})^2 \end{aligned} \quad (6)$$

$$= ||\boldsymbol{y} - \boldsymbol{f}||_2^2 + \lambda \boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} \quad (7)$$

$$\boldsymbol{f} = (\boldsymbol{I} + \lambda \boldsymbol{L})^{-1} \boldsymbol{y} \quad (8)$$

3 実験

3.1 実験仕様

テキスト分類問題の対象データには，Reuters-21578 (Reuters)[1] を用いる．Reuters は 135 のトピックカテゴリからなる Reuters newswire の英文記事を集めたデータセットである．本実験では “ModApte”分割に従って，本文とタイトルのみからなる記事データを抽出し，全データに対してストップワードの除去とステミング処理を行う．その後，同じデータセットを用いて GBSSL 手法でマルチラベル文書分類を行っている Subramanya ら [4] の実験仕様に合わせ，10 種のカテゴリ **earn**，**acq**，**money-fx**，**grain**，**crude**，**trade**，**interest**，**ship**，**wheat**，**corn** に対する分類精度を求める．Reuters の記事データはマルチラベルを有するため，ここでは各カテゴリ毎に one-versus-rest 法を適用した二値分類を行い，一定の閾値以上のカテゴリラベルを文書に付与するラベルとして採用する．

データセットは，テストデータ (ラベルなしデータ)$u = 3299$ 個を共通とし，これに教師データ $l = 20$ 個を加えたものを 16 セット用意する．データセットに含まれるデータ総数は $n = 3319$ 個である．教師データとして加えるカテゴリは，上記 10 種のカテゴリにそれら以外のカテゴリ (**others**) を加えた全 11 種とする．データ

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

セットに加える教師データ l 個のカテゴリは 11 種のカテゴリからランダムに選択するが，全 11 種のカテゴリの教師データが少なくとも 1 個ずつ含まれるように選択する．

グラフ構成の際に求める，潜在トピックの推定方法には，崩壊型ギブスサンプリングを用い，その反復回数は 200 回とする．最適トピック数はパープレキシティの値を算出し，その 5 回平均の値で決定する．$\alpha = 0$ のときは文書の表層情報のみを扱うため，推定を行う必要がない．このため，類似度が一意的に決まる．他方，$\alpha \neq 0$ のときは文書の潜在トピックの推定を行うため，類似度が一意的に決まらない．このため，5 回平均した値を用いることとする．ノード間の類似度におけるパラメータ α は $[0, 1]$ の範囲を 0.1 刻みで動かす．

ラベル伝搬法で用いた類似度グラフのノード数は $|V| = n(= 3319)$ である．k-近傍グラフの大きさのパラメータ k は $\{2, 10, 50, 100, 250, 500, 1000, 2000, n\}$，ラベル伝搬法のパラメータ λ は $\{1, 0.1, 0.01, 1e{-}4, 1e{-}8\}$ の範囲を動かす．15 セット中 5 つのデータセットによって，各カテゴリに対する最適パラメータ (k, λ) の組を決定した後，それらのパラメータの値を用いて，残り 10 セットに対して文書分類を行い，各カテゴリ毎に PRBEP を求め，各試行毎の各カテゴリに対する PRBEP の平均値を算出する．指標 PRBEP は，$Precision$(適合率) と $Recall$(再現率) が一致するときの値である．

3.2　実験結果

$[0, 1]$ における 0.1 刻み毎の各 α の値に対して，カテゴリ毎に決定した最適パラメータ (k, λ) を表 1 に示す．各カテゴリに対し，これらの最適パラメータを用いて行った実験結果を図 1〜10 に示す．横軸は α の値を表し，縦軸は PRBEP の値を表す．図 1〜10 は，各 α の値に対して行った 10 回の試行の各カテゴリ PRBEP の平均値を示している．各 α 毎に全カテゴリの PRBEP を合算して求め，その平均値の変移を図 11 に示す．図 12，13 は $\alpha = 0, 0.2, 1$ のときの，カテゴリ毎のテストデータ数とその PRBEP との相関関係を表している．横軸は各カテゴリに含まれるテストデータの数を表し，縦軸は PRBEP の値を表す．青の点線，黒の実線そして赤の一点鎖線は，それぞれ $\alpha = 0, 0.2, 1$ のときの結果を表す．

図 1〜11 において，$\alpha = 0$ の場合は，表層情報のみを用いた場合の結果であり，本研究におけるベースラインである．また，$\alpha = 1$ の場合は，潜在情報のみを用いた場合の結果である．それ以外 ($\alpha \neq 0$ または 1) は，潜在情報と表層情報を一定の割合 $(\alpha : (1 - \alpha))$ で混合した場合であり，両情報を用いた結果を示している．

まず，図 1〜10 に関連しては次の通りである．$\alpha = 0$ の時よりも，$\alpha \neq 0$ の時の PRBEP が必ず大きい値をとるのは，図 1, 2, 3, 6, 7, 8 である．他方，逆に $\alpha = 0$ の時よりも，$\alpha \neq 0$ の時の PRBEP が α の値によって小さい値をとるのは，図 4, 5, 9, 10 である．

次に，図 11 からは以下のことが分かる．マクロ平均値の最大値は 51.0 ($\alpha = 0.2$) であり，最小値は 44.5 ($\alpha = 1$) である．ただし，$\alpha = 0$ の時の値は 45.2 である．したがって，最大マクロ平均値 51.0 ($\alpha = 0.2$) は $\alpha = 1$ の時より 6.5%高く，更にベースラインである $\alpha = 0$ の時より 5.9%高いことが分かる．また，$\alpha = 0$ 〜0.2 の時，マクロ平均値は単調増加しており (45.2 → 51.0)，$\alpha = 0.2$ 以上では，マクロ平均値は単調減少している (51.0 → 44.5)．

図 12, 13 は，$\alpha = 0, 1$，並びにマクロ平均値で最大値をとる $\alpha = 0.2$ における，各カテゴリのテストデータ数とその精度の相関関係を表している．テストデータ数の多いカテゴリほど，潜在トピックを考慮した $\alpha \neq 0$ における精度は改善されていることが分かる．しかしながら，データ数が 200 個以下であるカテゴリにおいては必ずしも同様の改善傾向は見られない．

4　考察

図 1〜10 の各図において，PRBEP が最大値をとる時の α の値は各カテゴリ毎に異なっており，一律ではない．故に，精度が最大となる時の，α の値 (すなわち表層情報と潜在情報の混合割合) を一意的に決めることは難しい．しかしながら，半数以上のカテゴリにおいては，$\alpha = 0$ に対して $\alpha \neq 0$ のときの PRBEP は増加傾向を示しており，残りのカテゴリにおいても，適切な α が求まりさえすれば全てのカテゴリにおいてベースラインを超えることが分かる．

図 11 は，全カテゴリのマクロ平均 PRBEP を示している．$\alpha = 1$ を除いた全ての $\alpha \neq 0$ において，ベースラインである $\alpha = 0$ の時のマクロ平均値よりも高くなっている．特に，各 α をベースラインと比較すると，$\alpha = 0.1, 0.2, 0.3, 0.4, 0.5$ のとき，t 検定によって 5%有意でベースラインに対して精度向上があることが分かった．

図 12, 13 からは，ノード間の類似度に文書間の潜在トピックを考慮することが GBSSL の精度改善に繋がることが期待され、それは特に各カテゴリのデータ数が十分多量にあるときであるということが期待される．カテゴリ **wheat**, **corn** において，$\alpha = 0$ に対して $\alpha = 1$ のときの PRBEP が著しく悪化したのは，これらのカテゴリにおけるテストデータが少量であるため，LDA による十分なトピック推定が行えなかったためだと考えられる．

以上のことから，GBSSL 法のグラフ構成としては，表層情報のみを用いるよりも潜在情報も加えた両情報を

表 1: カテゴリ毎の最適パラメータ (k, λ)

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

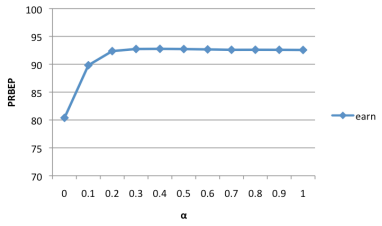


図 1: **earn** の平均 PRBEP

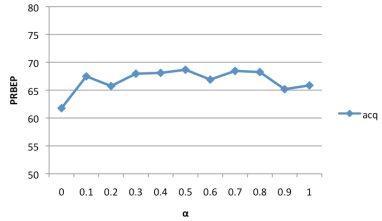


図 2: **acq** の平均 PRBEP

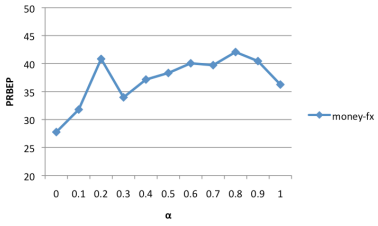


図 3: **money-fx** の平均 PRBEP

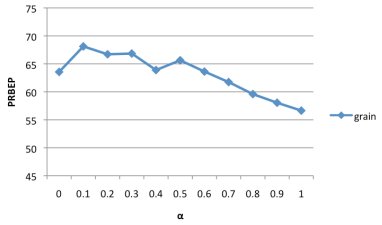


図 4: **grain** の平均 PRBEP

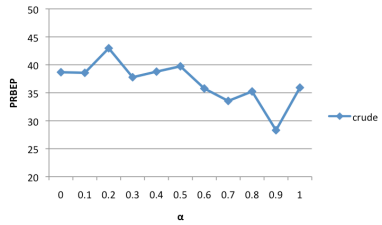


図 5: **crude** の平均 PRBEP

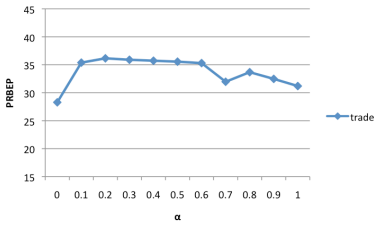


図 6: **trade** の平均 PRBEP

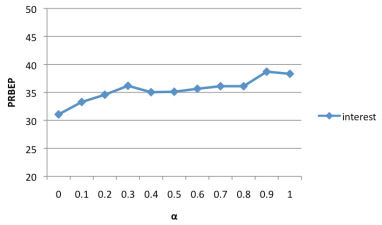


図 7: **interest** の平均 PRBEP

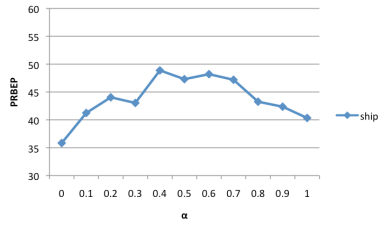


図 8: **ship** の平均 PRBEP

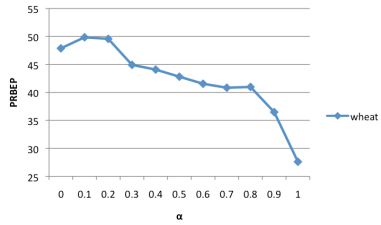


図 9: **wheat** の平均 PRBEP

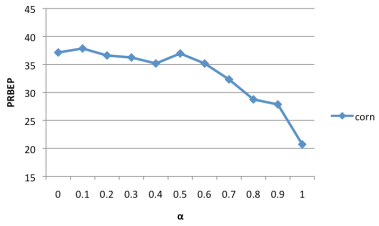


図 10: **corn** の平均 PRBEP

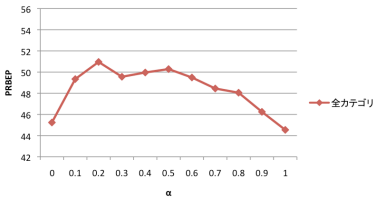


図 11: 全カテゴリのマクロ平均 PRBEP

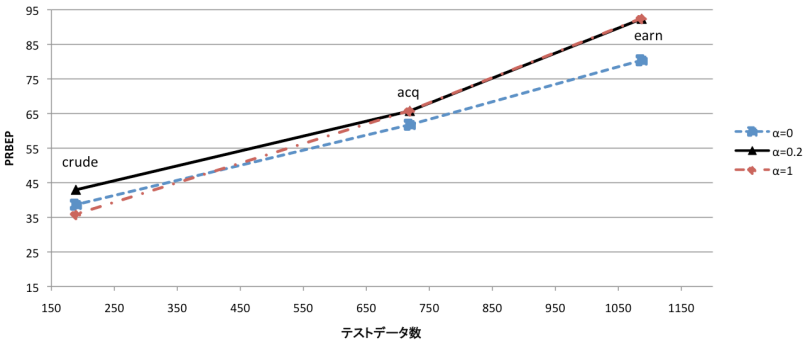


図 12: 各カテゴリにおけるテストデータ数と PRBEP との相関関係．$\alpha = 0, 0.2, 1$ における，カテゴリ **earn**，**acq**，**money-fx** のテストデータ数 (横軸) と PRBEP(縦軸)

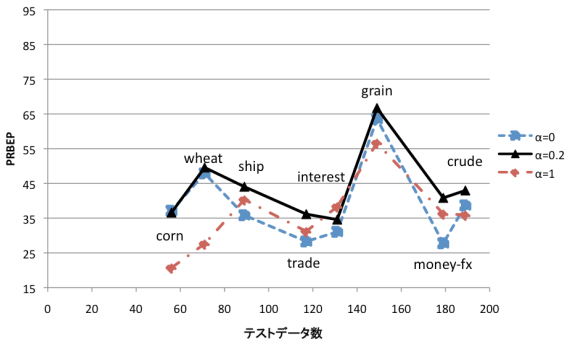


図 13: 各カテゴリにおけるテストデータ数と PRBEP との相関関係．$\alpha = 0, 0.2, 1$ における，カテゴリ **money-fx**，**grain**，**crude**，**trade**，**interest**，**ship**，**wheat**，**corn** のテストデータ数 (横軸) と PRBEP(縦軸)

用いる方が GBSSL 法の精度は向上することが分かる．また，十分なデータ数があるときのみ，潜在情報による精度向上への寄与率が上がることも期待される．したがって，両情報の混合割合 α の最適値が求まり，各カテゴリそれぞれにおいて十分な量のテストデータがありさえすれば，単に表層情報や潜在情報のみを用いる場合よりも，高い精度が得られるだろう．

5　結論

我々は，表層情報と潜在情報に基づく類似度グラフの構成法を提案した．マルチラベルを有する Reuters-21578 コーパスを用いた実験の結果から，GBSSL 法におけるグラフ構成では表層情報と潜在情報のどちらかだけを用いるよりも，両情報を混合させて同時に用いた方が GBSSL 法における文書分類の精度を向上させることが分かった．

今後の課題としては，我々が今回得た結論 (表層情報と潜在情報の両情報を用いる方がそれらを単体で用いるよりも精度が高い) を他のデータセットを用いて検証することであり，グラフスパース化手法などの工夫を行うことなどを通して更なる精度の向上を図ることである．

参考文献

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

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

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

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

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

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

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

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

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

[10] Gu, Q. and Han, J.: Towards Active Learning on Graphs: An Error Bound Minimization Approach, Data Mining, IEEE International Conference (2012)

[11] Ozaki, K., Shimbo, M., Komachi, M. and Matsumoto, Y.: Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language Data, Proceedings of the Fifteenth Conference on Computational Natural Language Learning (2011)

[12] Jebara, T., Wang, J. and Chang, S.: Graph construction and b-matching for semi-supervised learning, Proceedings of the 26th Annual International Conference on Machine Learning (2009)

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4回)
SIG-AM-04-05

単語の共起グラフを用いた潜在的意味に基づく効果的な文書分類の検証

Validation on Efficient Text Classification Based on Latent Semantic with a Graph of Co-occurring Terms

小倉由佳里 *　　小林一郎
Yukari Ogura　　Ichiro Kobayashi

お茶の水女子大学大学院人間文化創成科学研究科理学専攻
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

Abstract: We have proposed a method to raise the accuracy of text classification based on latent topic information, introducing several techniques such as extracting important words with PageRank algorithm and reducing the size of target documents by replacing them with important sentences in themselves. We have experimented on text classification with Reuters-21578 data set and confirmed that our proposed method worked to raise the accuracy of text classification. In this paper, we aim to verify our method with additional experiments using 20 Newsgroups data set and report the experimental result.

1　はじめに

近年，インターネットの発達に伴い，爆発的に増大した莫大な量のテキストデータを扱う問題がある．そのため大量のテキストを，自動でカテゴリごとに分類できるような文書分類手法が必要とされている．本研究の先行研究となる [13] では，文書の潜在的意味を考慮した分類手法が提案された．そこでは，文書分類の方針として，まず語彙の重要度に基づき重要文抽出を行い，元の文書を重要文のみで構成し，分類対象となる文書の精錬化を図る．語彙の重要度を決める指標としては，一般に $tf \cdot idf$ や語彙の頻度などが用いられるが，語の共起関係からグラフを構成し，PageRank アルゴリズムを用いて重要語の決定が行われた．次に，潜在的意味解析手法を用いて，文書の潜在トピックごとの確率分布をもとに，k-means 法でクラスタリングが行われている．また，実験では Reuters-21578 のデータセットを使用し，提案する手法の有効性を検証した．本稿では，提案する手法の汎用性を検証するために，20 Newsgroups のデータセットを用いた実験結果について報告し，考察を行う．

*連絡先：お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース小林研究室
〒 112-8610 東京都文京区大塚 2-1-1
E-mail: g0920509@is.ocha.ac.jp

2　関連研究

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

本研究の先行研究 [13] では，文書を潜在情報に基づいて分類することを目的とし，Newman ら [8] による潜在的情報の首尾一貫性は単語の共起関係により形成されるという報告を参考に，共起語からなるグラフを構

築し，それに PageRank アルゴリズムを適用することにより，抽出された重要語から重要文を決定する．その重要文を用いて，潜在情報に敏感な文書群を再構成し，文書分類を行う手法を提案した．以下，先行研究 [13] での説明を重複するが，提案手法の内容を再掲しつつ，追加実験の結果と報告と考察を行う．

3　提案方法

3.1　PageRank アルゴリズムによる重要語の決定

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

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

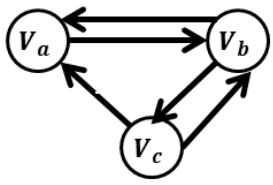


図 1: リンク関係の例

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

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

3.2　潜在情報による分類

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

3.3　提案手法における処理の流れ

本手法における，文書分類の流れを説明する．

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

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

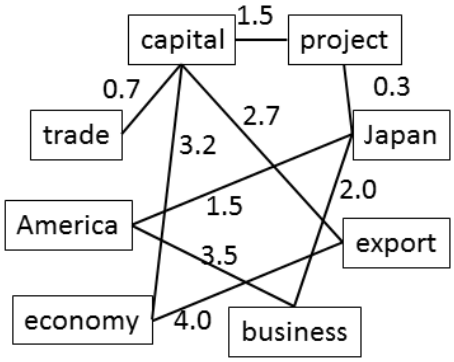


図 2: 類似度グラフ

step3 重要文の抽出

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

step4 分類

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

4　実験

4.1　実験仕様

実験対象データには，Reuters-21578[1] のテストデータと 20 Newsgroups[2]を使用した．提案する手法は，対象文書から重要文を抽出し，文書を精錬してから文書分類を行うため，文数の少ない文書では提案手法の効果が判別できないため，1 文書中の文章数が 5 文以上である文書を利用した．

Reuters-21578 のカテゴリは，文書分類の他研究 [9], [11] においても用いられている上位 10 件のカテゴリ，acq，corn，crude，earn，grain，interest，money-fx，ship，trade，wheat を利用した．その結果，文書数 792 件，語彙数 15,835 語，カテゴリ数 10 の文書群を対象に，タグの除去，ステミング処理，ストップワード除去を施し実験を行った．

20 Newsgroups は，20 のニュースカテゴリからなるデータセットである．文書数 11,269 件，語彙数 53,975 語で構成されている．本研究では，文献 [12] を参考にし，comp.graphics，rec.sport.baseball，sci.space，talk.politics.mideast の 4 カテゴリからそれぞれ 200 件ずつをランダムに選び，文書数 800 件，語彙数 14,198 語のデータを使用した．以後，このデータセットを 4-News と記述する．

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

4.2　評価手法

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

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

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

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

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

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

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

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

[1] http://www.daviddlewis.com/resources/testcollections/reuter21578
[2] http://qwone.com/ jason/20Newsgroups/

4.3　実験結果

k-means 法を 10 回行い，その平均値を測った．ただし，LDA を用いて，文書のトピックごとの確率分布から分類を行う場合には，出力される確率分布 θ が毎回変化する．そのため 1 つの θ に対して k-means 法を 10 回行い，マクロ平均を測った．文書間の類似度指標には，Jensen-Shanon 距離を用いた．重要度の高い上位 3 単語を含む文を抽出して分類を行った場合の正解率と F 値の結果をそれぞれ表 1，表 2 に示す．また，重要文抽出を行った後の文書群の単語数の変化を表 3，表 4 に示す．重要文抽出した後の文書群の単語数が，元の文書群の 8 割程度になるよう設定し，分類を行った場合の正解率と F 値の結果をそれぞれ表 5，表 6 に示す．

表 1: 正解率

単語の重要度	Reuters-21578	4-News
PageRank	**0.5671**	0.6415
$tf \cdot idf$	0.5500	0.5915
重要文抽出なし	0.5177	**0.8563**

表 2: F 値

単語の重要度	Reuters-21578	4-News
PageRank	**0.4852**	0.6321
$tf \cdot idf$	0.4347	0.5091
重要文抽出なし	0.4262	**0.8494**

表 3: Reuters-21578 の単語数の変化

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

表 4: 4-News の単語数の変化

手法	10 語	15 語	20 語	25 語	30 語
PageRank	10,731	10,958	11,078	11,171	11,241
$tf \cdot idf$	11,048	11,441	11,731	11,849	11,937

表 5: 正解率

単語の重要度	Reuters-21578	4-News
PageRank	**0.5529**	0.8175
$tf \cdot idf$	0.5499	0.7948
重要文抽出なし	0.5177	**0.8494**

5　考察

実験結果の表 1，表 2 より，4-News を用いた実験では，Reuters-21578 を用いた実験と同じ結果は得られなかった．Reuters-21578 では，重要文抽出により文書が精錬されたことから，文書の特徴を表現するのに必要な文のみが残り，文書のトピックごとの確率分布の差が測りやすくなったのではないかと考えられた．しかし，4-News を用いた実験では，重要文抽出を行わない場合の方が，行う場合よりも精度が高い結果となった．これは，データセットの性質の違いであると考えられる．この原因としては，4-News は元の文書群の単語数が Reuters-21578 より少ないことから，重要文抽出を行ったことにより，単語数がさらに減り，本来大量の文書の下で行う学習の効果が下がり，LDA の精度が下がったのではないかと考えられる．また，重要度の高い単語上位 3 単語を含む文を抽出した後の文書群の単語数の変化から考察すると，Reuters-21578 では元の文書群の 8 割程度抽出できているのに対し，4-News では 6 割程度しか抽出できておらず，このため精度が大きく下がったと考えられる．重要文抽出に関しては，Reuters-21578，4-News 共に，$tf \cdot idf$ を用いた場合に比べ，PageRank を用いて重要文の抽出を行った場合に文書分類の精度の向上が見られた．このことから，文書の 3 文中での単語の共起関係からグラフを構成し，単語の重要度を PageRank アルゴリズムを用いて決定することにより，分類に適した単語の重要度が得られることが検証された．

また表 3，表 4 から，重要文抽出したあとの単語数の比較では，$tf \cdot idf$ と比較して，PageRank を用いた場合に，より語彙数，文数が減っていることが分かる．$tf \cdot idf$ の場合，特定の文書に多く出現している単語の値が高くなるため，$tf \cdot idf$ が高い単語は，その文書中の多くの文に出現している可能性が高い．そのため，$tf \cdot idf$ の高い単語を含む文を抽出すると，自然と多くの文を抽出することになるのではないかと考えられる．Reuters-21578 と 4-News での結果を比較してみると，Reuters-21578 において，重要度の高い上位 1 単語を含む文を抽出した後の単語数と，4-News において，重要度の高い上位 10 単語を含む文を抽出した後の単語数の，元の文書群に占める割合がほぼ等しくなっている．これは，Reuters-21578 は 10 カテゴリであるのに対し，4-News は 4 カテゴリであることから，4-News では，同じカテゴリで似た単語の重要度が高くなっているから抽出単語数が少ないのではないかと考えられる．

表 5，表 6 では，抽出後の単語数を元の文書群の 8 割にして実験を行った．結果は，表 1，表 2 と同じ傾向

表 6: F 値

単語の重要度	Reuters-21578	4-News
PageRank	**0.4582**	0.8116
$tf \cdot idf$	0.4347	0.7948
重要文抽出なし	0.4262	**0.8494**

が見られた．これらと比較すると，4-News では，抽出後の単語数の割合を増やしたため，重要文抽出する場合において精度の向上が見られたが，重要文抽出をせず分類を行う場合に一番精度が高くなる結果となった．

6　おわりに

本研究では，先行研究 [13] で提案された PageRank を用いた重要語の抽出を行い, それに基づいて重要文を抽出し, 潜在的意味によるクラスタリングを行う手法の汎用性を検証するために，20 Newsgroups のデータセットを用いて，追加実験を行った．実験から，Reuters-21578 では提案手法の有効性が確認されたが，20 Newsgroups を用いた実験では，重要文抽出を行わない場合に最も精度が高くなる結果となり，データセットにより結果に違いが見られた．

今後の課題としては, さらに他のデータセットを用いた実験を行うつもりである．また，文書量が LDA の精度に影響することが考えられることから，文書数をさらに増やした実験を行いたいと考えている．さらに，トピック数を変化させた場合の実験結果の比較を行う．また, 現在は k-means 法での分類しか行っていないため, 他の多クラス分類手法との比較を行うつもりである．

参考文献

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

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

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

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

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

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

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

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

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

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

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

[12] Liping Jing, Michael K.Ng, Jun Xu, Joshua Zhexue Huamg.: Subspace Clustering of Text Documents with Feature Weighing K-Means Algorithm *PAKDD 2005, LNAI 3518*, pp. 802–812 (2005)

[13] 小倉　由佳里，小林　一郎．: 単語の共起グラフを用いた潜在的意味に基づく効果的な文書分類への取り組み，インタラクティブ情報アクセスと可視化マイニング第 *3* 回研究会,(2013)

TETDMを用いた汎用性を考慮したシステムの設計指針に関する基礎的検討

Consideration of Design Guide for Constructing General Purpose System using TETDM

梶並 知記[1*]　田代 航一[2]　利根川 拓馬[2]　北村 侑也[2]　高間 康史[2]
Tomoki Kajinami[1]　Koichi Tashiro[2]　Takuma Tonegawa[2]
Yuuya Kitamura[2]　Yasufumi Takama[2]

[1] 神奈川工科大学
[1] Kanagwa Institute of Technology
[2] 首都大学東京
[2] Tokyo Metropolitan University

Abstract: This paper considers a collaborative policy for combining tools, in development of system using TETDM. TETDM is an total environment for text data mining, can prepare for various mining tasks by combination of small mining tools. However, an useful guide in the design of system constructed with several small tools developed by different tool developers has not been considered. This paper describes a design guide adjusting user's purpose and system's specifications for constructing general purpose system, and shows an example of practice.

1　はじめに

本稿では，TETDMを用いたシステム開発における，ツール同士の連携方針について検討する．TETDMは，テキストデータマイニングのための統合環境であり，小規模なツール同士を連携させることで多様なタスクへ対応可能としている [6]．ツールの種類は,「マイニング処理ツール」と「可視化ツール」の２つに分類され，ユーザは任意のマイニング処理ツールと可視化ツールを１つずつ選択し，それらを１対１で組み合わせて使うことで，テキスト分析を行う．ここで，１対１の組み合わせは複数種類同時に使用することが可能で，TETDM上の複数枚のパネルそれぞれに，マイニング処理ツールと可視化ツールの組が１つずつ配置される．これにより，同時にさまざまな観点からテキスト分析を行うことが可能になっている．また，TETDMでは，ツール間の連動として，他のツールから出力されるデータを別のツールで利用する仕組みが用意されている．これにより，テキスト分析にとどまらず，複数のツールからなるテキストデータマイニングシステムを開発するプラットフォームとしてTETDMを活用することも，可能となっている．

*連絡先：神奈川工科大学情報学部情報工学科
〒 243-0292 神奈川県厚木市下荻野 1030
E-mail: kajinami@ic.kanagawa-it.ac.jp

しかしながら，複数の開発者が個別に開発した小規模ツール同士を連携してシステムを設計する指針については未検討である．本稿では，対話的なクラスタリング環境の構築を目的としたシステム開発において，目的優先と手段優先志向を摺り合せるシステム設計指針について述べる．TETDMの仕様を変更せず，仕様の制限がある中で文章を対象にした対話的なクラスタリング環境を構築する本研究の意義は，以下の３点である．

1. TETDMに備わっているツール連動の仕様から逸脱せず対話的なクラスタリング環境をTETDM上に実装する例を示すことで，TETDM上に新たなプラットフォームを構築するシステム開発に応用できる．
2. 可視化ツールとの組み合わせを想定しない複数のマイニング処理ツールを統合的に扱う手法を提案することで，マイニング処理ツールと可視化ツールを１対１対応させるTETDMの特徴を活かしつつ，TETDMを拡張する方向性を示す．
3. 対話的なクラスタリングのための，統一的なデータのやり取りを可能とすることで，クラスタリングに関連するツールを，同一環境上で比較し易くなる．

本稿では，複数のツールを連携し TETDM 上に対話的なクラスタリング環境を構築することを目指すが，特定のクラスタリング手法に特化したり，技術文書の分類や商品レビューの分類といった特定のタスクに特化するものではなく，汎用的なものである．そのため，汎用性を意識した，ツール連携の方略を検討する．

本稿の構成は以下のとおりである．2 節で，TETDM の応用に関する研究について述べ，本稿の位置づけを明確にする．3 節で，対話的なクラスタリング環境におけるツールの役割の同定や，ツール同士でやりとりするデータの内容を抽象化，データの型を定義する．4 節で，ツールを統合的に扱う管理パネル方式の提案を行い，5 節で，試験的なシステム実装例を示す．

2　関連研究

2.1　TETDM の活用や拡張

TETDM を用いることで，ユーザはさまざまなツールから，システム上可能な範囲で任意の組み合わせを選択して，テキスト分析処理の結果を得ることができる．実践的な活用例として，医療現場でのカルテ分析がある [7]．また，R といった既存の分析ソフトウェアと連携し，TETDM を拡張する研究もおこなわれている [8]．TETDM の拡張に関する研究として，マイニング処理ツールと可視化ツールの組み合わせをユーザが能動的に選択する必要がある TETDM の特徴に着目しているものがある．TETDM のコアとなるプログラム部分もオープンソースであることを活かしてツールの組み合わせ作業の支援が行われている [2][4]．

本稿では，テキスト分析を行う特定の現場を想定したものではなく，また，既存のソフトウェアとの密な連携を目指すものではない．本稿は TETDM の拡張に関する研究であるが，TETDM のコアとなるプログラム部分には触れず，マイニング処理ツールと可視化ツールの 2 種類のツールを実装する枠組み，TETDM の仕様に従いツール同士を連携させる枠組みの中で，新たなプラットフォームを構築するものである．

2.2　対話的なクラスタリング

対話的なクラスタリングは，ユーザの要求に応じたクラスタリング結果を出力するための方法で，ユーザによるクラスタリングに必要なパラメータ，制約の入力を支援する [3]．ユーザは，自身の意図や背景知識を考慮したクラスタリングへの制約付与を行い，クラスタリングした結果とのインタラクションを繰り返し，望みのクラスタリング結果を得る．対話的なクラスタリングは，文書の分類に応用されている [5]．

本稿では，対話的なクラスタリング環境の構築を目指すが，ツールの組み合わせによってさまざまな視点からテキストデータを眺め，インタラクティブに分析する TETDM と，異なるクラスタリング結果を並列に眺め，そこからユーザの意図や背景知識に応じて，反復的にクラスタリングを行う対話的なクラスタリング環境には親和性があると考える．

3　クラスタリングのためのツール連動

本稿では，ユーザが複数のクラスタリング結果を見比べることができ，また使用するクラスタリング手法，各種パラメータの設定が動的に行える環境の構築を想定したシステム設計の方略を考える．また，クラスタリング結果（可視化）としてユーザが見たい情報は，クラスタ集合，それに含まれるクラスタ，クラスタに分類されている文書，文書内の単語の 4 種類であると想定する．3.1 節で，クラスタリングの流れを 3 段階にわけ，マイニング処理ツールや可視化ツールとの対応について述べる．3.2 節で，対話的なクラスタリングのための，ツール間連動で用いるデータ型について述べる．3.3 節で，複数人からなるシステム開発の中で実際に行った，ツールの分類作業，ツール間連動の整合性確認作業について述べる．

3.1　クラスタリングの流れ

クラスタリングの実行手順を大きく 3 段階に分けると，以下のようになる．可視化処理段階と TETDM の可視化ツールは自然と適合するが，前処理とクラスタリング処理は，ともにマイニング処理ツールとして実装する．

前処理 クラスタリングする文書のベクトル化・特徴量の算出する段階

クラスタリング処理 任意のクラスタリング手法によりクラスタリングする段階

可視化処理 選択したクラスタリング手法に応じた/ユーザの意図に応じた可視化手法によって，クラスタリング結果を出力/フィルタリングする段階

複数のツールを組み合わせてクラスタリングシステム全体を構成するため，最低でも各段階 1 つずつのツールを連結することで，クラスタリングが一通り完了できることになる．

3.2　クラスタリングに必要なデータ型

本節では，ツールの役割分担を考える際，システム設計の際に採用されるデータの流れに着目する考え方 [1] を参考に，ツール間でやりとりするデータに具体性を持たせて検討を行う．

ここでは，前処理段階，クラスタリング処理段階，可視化段階の間にどのようなデータが必要であるか検討する．できるだけ複雑にならず，なおかつユーザが必要とする要素（クラスタ集合，それに含まれるクラスタ，クラスタに分類されている文書，文書内の単語）を表現するのに十分なデータ型である必要がある．なお，TETDM の仕様に従い，ユーザからシステムに入力するものはテキスト形式の文書ファイルとする．入力文書は単一ファイルとは限らず，複数の文書ファイルにも対応できる．また，TETDM の標準的な機能により，文書を段落や文章，単語に分割する操作は完了しており，文書内の文章数や単語数などは特定の変数に格納され，また特定の単語などを，配列の要素数（ID）を指定することで一意に定めることができることを前提としている．したがって，本稿では，ツール間で具体的にやりとりするデータの内容を文書ベクトルリスト，クラスタ文書リスト，クラスタ単語リストの 3 つとし，TETDM で用意されている，ツール連動用のデータ型に対応させる．文書ベクトルリストは，文書と単語の 2 次行列で定義する．中身は，任意の特徴量（TF-IDF など）によって計算された各単語の重みとなる．クラスタ文書リストは，クラスタを行，クラスタに含まれる文書を列とする 2 次行列で定義する．クラスタ単語リストは，クラスタを行，クラスタに含まれる単語を列とする 2 次行列で定義する．

表 1 は，文書ベクトルリスト，クラスタ文書リスト，クラスタ単語リストについて，TETDM で用意されているデータ型との対応を示している．クラスタ文書リストの部分に，boolean と double の 2 つの型があるが，列数が全文書数ありクラスタに含まれている文書を 1，含まれていない文書を 0 とする 2 値表現を行う場合と，あるクラスタに含まれている文書 ID を配列として格納する場合の両方に対応するためである．

表 1: 具体的なデータとデータの型.

データ	型
文書ベクトルリスト	double[][]
クラスタ文書リスト	boolean[][]，int[][]
クラスタ単語リスト	double[][]

3.3　実際の設計方略

3.1 節と 3.2 節で述べた，段階分類とデータの定義に基づき，本稿で実際に行ったシステム設計方略は以下のとおりである．

1. ツール名と入出力データの内容と処理内容を記載するカードを用意
2. 複数の開発者（プロジェクトメンバ）による，カードへの記載
3. ツール同士の入出力データのマッチングを精査
4. ツールの入出力データ再検討や，ツールの分割や統合

1 つのツールを 1 枚のカードで表現し，前処理，クラスタリング，可視化の 3 段階に分類されたツールをつなぐために，データ入出力の整合性をとる流れである．データ入出力の整合性がとれない場合は，処理内容と入出力データの関係が適切かどうか，またツールの処理内容を分割または統合可能かどうか検討する．なお，前処理，クラスタリング，可視化のいずれかに当てはめるのが難しいツール，特定のクラスタリング手法に依存するツールに関しては，別途オプションカテゴリとする．

上記方略の（1）と（2）が，開発プロジェクトの目的を考慮した目的優先の志向に対応し，(3）と（4）が，TETDM の仕様から実現可能な手段を考慮した手段優先の志向に対応する．すなわち，開発者やユーザの考える,「実現したいこと」の「入出力データが何か」検討し，TETDM のツール連動の仕組みに適合するようなデータの流れになるよう，調整していく．

表 2 に，具体的に出されたツール案の一部を示す．前述したクラスタリングの段階ごとに，ツールを分類している．括弧内のものは，オプションカテゴリのものである．また，本研究は教育機関で実施しており，著者らの一部（工学系学生，大学院生）のクラスタリング手法に関する学習も兼ねている．したがって，ここで既存のクラスタリング手法のすべてを列挙することは目指していない．

表 2: クラスタリングの段階とツール群.

段階	ツール
前処理	TF-IDF 計算, BM25 計算
クラスタリング	K-means，階層的クラスタリング，制約付き階層的クラスタリング，(重心計算，距離計算)
可視化	ネットワーク型図，階層構造図

4　管理パネル方式によるツール管理

本節では，複数のマイニング処理ツールの管理を行う管理パネルを TETDM 上に構築し，1 つのパネルを利用してツールの組み合わせを変更する，管理パネルモデルを提案する．便宜上，ここでは TETDM で採用されている基本的なツールの管理を「基本方式」，管理パネルモデルによるツールの管理を「管理パネル方式」を呼ぶ．

4.1　基本方式の問題点

図 1 に，基本方式に基づく，クラスタリング環境を示す．基本方式では，マイニング処理ツール同士の連携（データのやり取り）が許されているものの，1 枚のパネルにマイニング処理ツールと可視化ツールを 1 対 1 で組み合わせて配置する．ツール開発者の視点では，マイニング処理ツールを開発する際に，必ずなんらかの可視化ツールとセットで使われることを想定しておかなくてはならない．TETDM で用意されている，マイニング処理ツール同士の連動機能を用いて，他のマイニング処理ツールでのみ使われるデータを出力するツールの作成も可能であるが，原則的に，TETDM ではマイニング処理ツールと可視化ツールを 1 対 1 対応させる設計方針となっている．このことは，ツール利用者（ユーザ）側においても，問題となる．ツールの組み合わせを指定する，どのツールとどのツールが組み合わせ可能なのか事前に知っておく，またはツールに付属する説明文を熟読して調べる必要がある．そのため，ツール選択の改善を試みた研究もなされている [2][4].

対話的なクラスタリング環境を構築する際，ユーザにとって重要なことは，どのパネルにどのような (名前・機能の) ツールを組み合わせるかより，クラスタリングに必要なパラメータ（前処理段階）をいかに与えるか，また実行したいクラスタリング手法が選択できるかといった点である．したがって，TETDM 上に実装されているクラスタリング関連ツールを統合的に扱う，インタフェースの必要性が生じる．

4.2　管理パネルモデル

図 2 に，管理パネルモデルによるクラスタリング環境の概要を示す．本稿で提案する管理パネルモデルは，クラスタリングのために TETDM 上に仮想的な統合環境を構築するものである．パネルへマイニング処理ツールや可視化ツールを配置する仕様や，定義されているツール間のデータ連動に用いるメソッドやデータ型など，そのまま使用している．TETDM のコアとなるプログラム部分に手を加えることはない．

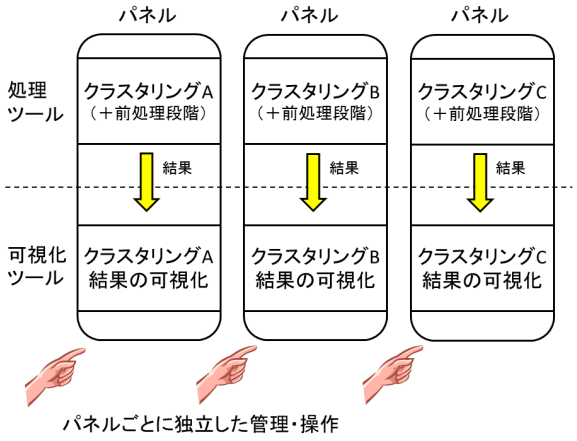


図 1: 基本方式によるクラスタリング．

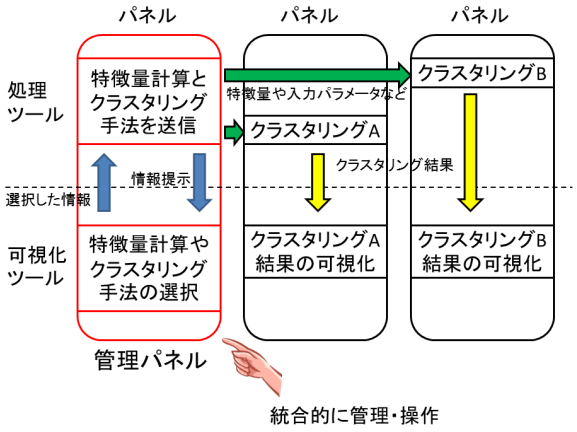


図 2: 管理パネル方式によるクラスタリング．

管理パネルは，TETDM の仕様に従い，マイニング処理ツールと可視化ツールが 1 対 1 対応して，TETDM のパネル上に配置される．管理パネルが行うことは，文字通り，クラスタリングに関連するツール群の管理であり，ユーザに提示する情報は，どのような前処理が可能か，どのようなクラスタリング手法が利用可能かの 2 点である．ユーザは，管理パネル上で，自身が利用したい前処理方法，クラスタリング手法を選択する．管理パネルのマイニング処理ツールは，ユーザが選択した前処理とクラスタリング処理の組み合わせに応じて，関連するマイニング処理ツールを動作させる．この際，前処理とクラスタリング処理の組み合わせを変えた，異なる処理を並列に実行可能である．

管理パネル以外のパネルには，管理パネルで選択した情報を引き継ぐマイニング処理モジュールを配置する．これにより，ユーザからは特定の前処理やクラス

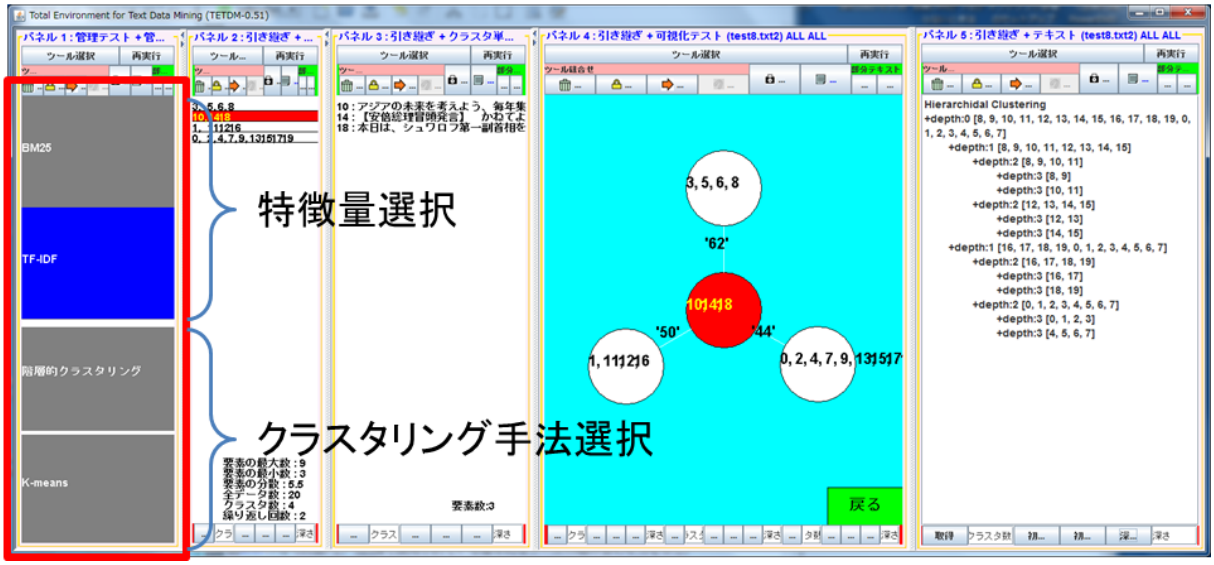


図 3: 管理パネルモデルに基づくクラスタリング環境.

タリング手法を個別に選択する指定する処理が隠されることになる．ユーザが管理パネル以外のパネルで自らの要求に応じて選択するのは，どのような情報が見たいか，すなわち可視化手法の選択だけである．

管理パネル方式では，ユーザは，パネルごとにマイニング処理ツールと可視化ツールの組み合わせに悩む必要がなくなる．ユーザは，管理パネル 1 つで複数のツールをまとめて取り扱うことができるようになった上で，クラスタリング結果を出力する可視化ツールを配置したパネルに現れる，ボタンや入力フォームなどを利用し，出力のフィルタリング，結果に対するフィードバックを行うこともできる．3.1 節で述べたオプションカテゴリのツールは，特定のクラスタリング手法との結びつきが強いマイニング処理ツールとなるため，そのクラスタリング手法の結果を出力するパネルからパラメータ入力を受け付ける形で実装することが望ましい．管理パネル方式を採用することで，TETDM の特徴でもあるパネルごとのインタラクション機能を維持したまま，管理パネルで複数のツールを統合する，汎用的な対話的クラスタリング環境が構築できることになる．基本方式ではパネルの独立性が高いのに対して，管理パネル方式では，複数のパネルにまたがって共通する処理を行うマイニング処理ツールを統合管理する．

5　クラスタリング環境の実装

3 節で述べた連動ルールと，4 節で述べた管理パネルモデルに従った，試験的なクラスタリング環境を TETDM 上に構築する．

図 3 に，クラスタリング環境の実行例を示す．図中の，左端のパネル（図中赤枠で囲った）が，管理パネルである．特徴量選択とクラスタリング手法の選択が可能である．現在のクラスタリング環境では，特徴量を TFIDF と BM25 から選択でき，クラスタリング手法を K-means と階層的クラスタリング（最近隣法）から選択できる．クラスタリング結果は，管理パネルの右側に並んでいるパネルに表示されており，テキストや図形を用いて結果を提示している．

6　おわりに

本稿では，TETDM を用いたシステム開発における，ツール同士の連携方針について検討した．対話的なクラスタリング環境の構築を目的としたシステム開発を想定し，ユーザの要求と TETDM の仕様に関する摺り合せを行った．データの流れに着目し，クラスタリングの段階ごとに必要なデータの型と，TETDM で定義されているツールの連動に関する仕様を対応させた．また，対話的なクラスタリング環境に適する，パラメータの指定やクラスタリング手法の選択を支援する，管理パネルモデルを提案した．本稿では，2 種類の特徴量計算手法の指定と，クラスタリング手法の指定が可能な試験的なシステムの実装を行った．

今後，ユーザによる制約の指定，ユーザからのフィードバックに応じた処理を可能にするほか，前処理のツール，クラスタリング手法のツールを増やし，クラスタリング環境を充実させる．

本稿で提案する枠組みにより構築する環境は，クラスタリングによる文書分析を行いたいユーザだけでなく，新規のクラスタリング手法や可視化手法などを他

の手法と比較して評価を行いたい研究者にも有益であると考える.

参考文献

[1] トム・デマルコ（著）, 高梨智弘, 黒田純一郎 (監訳) : 構造化分析とシステム仕様—目指すシステムを明確にするモデル化技法—, 日経 BP 出版センター (1994)

[2] 中垣内李菜, 川本佳代, 砂山渡 : 統合環境 TETDM を用いたテキストマイニングにおける初心者のためのツール選択支援, 第 27 回人工知能学会全国大会, 3B3-NFC-01a-1 (2013)

[3] 中村朋健, 上土井陽子, 若林真一, 吉田典可 : クラスタリング結果の特徴抽出を用いる高次元データの対話的クラスタリング, 情報処理学会論文誌 : データベース, Vol.47, No.SIG 19（TOD 32）, pp.28–41 (2006)

[4] 大塚直也, 松下光範 : テキスト分析における試行錯誤の支援に向けて—TETDM のインタフェースに関する一考察—, 第 2 回インタラクティブ情報アクセスと可視化マイニング研究会, SIG-AM-02-10, pp. 56–61 (2012)

[5] 佐藤祐介, 岩山真 : 半教師有りクラスタリングを適用した対話型文書分類技術の提案, 情報処理学会研究報告, Vol. 2009-DBS-148, No. 7, pp.1–6 (2009)

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

[7] 谷恵里香, 砂山渡 : 電子カルテにおける新人とベテランの特徴比較支援システム, 第 3 回インタラクティブ情報アクセスと可視化マイニング研究会, SIG-AM-03-07, pp. 37–43 (2013)

[8] 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)