ベクトル表現可能な機械抽出トピックの定量的評価法

福井健一[†] 斉藤和巳[††]
木村昌弘[†††] 沼尾正行[†]

大規模文書群からのトピック自動抽出は文書群全体像の把握や文書分類などに有用であるとして，研究が行われている．しかし，ここで機械抽出トピックの評価は重要な課題であるが，人手によってトピック分類された文書群が得られたとしても，トピック表現の多様性によりトピック間の対応付けが困難となるため，単純に機械抽出トピックと比較評価できない．そこで本稿では，潜在的意味解析などによるベクトル表現可能なトピック抽出法を対象として，人手によってトピック分類された文書群を用いて抽出トピックの解釈可能度を定量的に評価することを試みた．日本語および英語の新聞記事などから *LSA* や *PCA* および Spherical *k*-means 法によって抽出したトピック群で，提案評価法の妥当性を検証した．

Evaluation of Vector Representable Topics That were Extracted Automatically

Ken-ichi Fukui,[†] Kazumi Saito,[††] Masahiro Kimura[†††]
and Masayuki Numao[†]

Automatic topic extraction from a large number of documents is useful to figure out an entire picture of the documents or to classify the documents. Here, it is an important issue to evaluate the automatically extracted topics, however, even if manually-labeled documents are obtained, it is impossible to compare automatically and manually derived topics due to complexity and uncertainty of the topics' structure. As the objective is vector representable topic extractions such as Latent Semantic Analysis, in this paper we tried to evaluate the interpretability of automatically extracted topics using the manulally-labeled documents. We validated the proposed evaluation method using topics extracted by *LSA*, *PCA* and Sherical *k*-means from Japanese and English news articles.

1. はじめに

近年，インターネットを通じてニュース記事，ブログ，電子メールなどから大規模な文書群を容易に得ることができるようになった．これら文書群の内容は，世界の最新の事件や出来事など，あるいは文書提供者間での議論などと様々で刻々と文書内容が変化する．このような文書群から，適切な主要トピックを自動抽出し，各トピックに関連する文書群を同定することにより，文書群を体系的に整理することができれば，文書群の全貌を容易に把握することができるとして研究が行われている．

† 大阪大学産業科学研究所
I.S.I.R, Osaka University
†† NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
††† 龍谷大学理工学部電子情報学科
Department of Electronics and Infomatics, Ryukoku University

現在，文書群からの機械的トピック抽出法は，文書内に出現する有意な単語群を基底とした特徴空間を考え，各文書を単語頻度ベクトルとして取り扱うベクトル空間モデルに基づく方法が一般的である．ここで，トピックとは同じ事柄について述べられている文書群が属するクラスを指し，同じトピックに属する文書は出現する単語頻度が似ていると考えられる．ベクトル空間モデルに基づいたトピック抽出には様々な方法により研究がなされている．たとえば，古典的クラスタリング手法によって得られた各クラスをトピックとする方法[1]や，同じトピックに属する文書はある特徴的な1つの軸の周りに分布していると考えると，特徴空間から特徴軸を発見する問題として定式化する方法がある．特徴軸を発見する方法としては，情報検索の分野で広く用いられている潜在的意味解析（*Latent Semantic Analysis: LSA*）[2]や，主成分分析（*Principal Component Analysis: PCA*）に基づく方法[3),4]，さらには近年信号処理の分野で発展した独立成分分析

（*Independent Component Analysis: ICA*）を用いる方法[5]がある．その他のトピック抽出法としては，あらかじめ選択された特徴語の時系列から χ^2 検定により傑出した単語群を抽出しボトムアップにトピックを構成する方法[6]や，混合ガウス分布に基づいて各トピックの分布を推定する方法[7]なども研究されている．

ここで，どのトピック抽出法を用いたとしても機械抽出されたトピックの評価は重要であるが難しい問題である．それは，トピックは一般的に多様なとらえ方が可能であるからと考えられる．たとえば，「米国大統領」に関するトピックと大きくとらえるか，「米朝実務会談」とより細かくとらえるかといった階層の問題もあれば，普通，人はこのようなとらえ方はしないと考えられるが「各国の議員選挙」という視点の違いの問題もある．このようなトピック表現の多様性により，機械的に抽出されたトピックと認知的に☆抽出されたトピック間を 1 対 1 対応で比較するのは難しいと考える．

本稿で対象とするような文書群は，一般的にはトピックを特徴付ける単語群をあらかじめ選択することは困難である．*LSA* を代表とするトピックを特徴軸として抽出する手法では，あらかじめ特徴語の選択を必要とせず，文書群の大まかな内容や概念に基づいた微妙なトピックを抽出できるとされ未知のトピックの発見に適していると考えられる．本稿では，これらベクトル表現可能なトピック抽出法を対象とした抽出トピックの評価方法を提案する．ここで，認知的に抽出されたトピックは信頼できるものと考えて，それらを説明変数として多対 1 対応で機械抽出トピックの解釈可能度を考える．認知トピックを代表する認知トピックベクトルを定義し，ある機械抽出トピックベクトルと認知トピックベクトルの線形和との近似度合いにより定式化した．

2 章は従来のトピック抽出研究における評価方法と関連研究について，3 章では評価の対象となる代表的なトピック抽出法，4 章では提案する評価方法ついて説明する．そして，5 章で実際のニュース記事および人工データを用いて提案評価方法の妥当性を検証した結果を示す．

2. 従来評価法および関連研究

トピック抽出に関連する代表的な研究に *Topic Detection and Tracking*（*TDT*）プロジェクトがある．これはあらかじめ人手によってトピックに分類された文書群を訓練データとして，トピックの検出と追跡を行っている[8),9)]．これは教師あり学習であるため定量的な評価が可能になっている．しかし，あらかじめ各文書を分類する作業は膨大となることや，日々，刻々とトピックが変化するニュース記事や電子メールなどを扱う場合，つねに訓練データを更新し続けなければならないという問題もある．

一方，本稿では与えられたトピックの検出と追跡ではなく，刻々と変化するトピックも扱えるような教師なしのトピック抽出法の評価を対象としている．従来，教師なしトピック抽出に関する研究では，次のような評価が行われてきた．たとえば，自動的にトピック分類された文書群に対して後付けでトピックの解釈を行っている[3),6)]．これは明らかに機械抽出結果主体の評価であり，真に機械抽出トピックを評価するには人主体の評価が必要不可欠であると考える．また，本稿が対象とするベクトル表現可能なトピック抽出法では，トピックを表す単語集合どうしの重複率や，認知トピックを代表するベクトルと機械抽出された特徴軸との *cosine* 類似度に基づいてトピックを検出したと見なしている[5]．しかし，これは 1 対 1 対応を前提にしており，トピック表現の多様性はまったく考慮されていない．

本稿での評価法のような，認知トピックを用いて抽出トピックの解釈可能度を定量的に測る試みはほとんどない．関連研究として，異なる文書体系間の構造マッチング[10]があげられる．そこでは *Naive Bayes* モデルに基づくパラメータ混合モデルによって両者の体系間の対応付けを得ているが，一般には解が定まらないため，与えられる体系の階層構造を利用している．しかし，本稿が対象とする文書群の体系は一般的には不明である．

3. ベクトル表現可能な代表的トピック抽出法

各文書は単語出現回数のみに着目した Bag-of-Words（BOW）モデルで表現し☆☆，本稿ではベクトル表現可能な代表的トピック抽出法として，潜在的意味解析（*LSA*），主成分分析（*PCA*），Sherical k-means 法（*SKM*）を用いた．各トピックは文書群の大まかな内容や概念を表すトピックベクトルとして抽出される．トピックベクトルは，*LSA* や *PCA* では固有ベクトルに対応し，また，*SKM* では各クラスタの中心ベクト

☆ 本稿では，“機械的” と対比させて単純に人手によることを指す．“機械的” は “自動的” と読み替えられる．

☆☆ 実際には，前処理として低頻度語や stop word と呼ばれる文書の内容に関与しない語は削除されており，有意な単語の集合となっている．

ルに対応する．

3.1 単語の重み付け

まず，各文書に含まれるそれぞれの単語の重要度を表す指標として，ある単語が文書中に出現する回数 *tf*（*term frequency*）と，その単語が文書群中に出現する文書数 *df*（*document frequency*）を用いる．*df* は単語の普遍性を表しており，*df* の値が小さいほど特徴的な単語であるといえる．そのため，*df* の逆数 *idf*（*inverse document frequency*）を *tf* に乗じた，$tf \cdot idf$ が用いられている．*tf* や *idf* の定義には様々提案されている[11]．総文書数を N，第 i 単語が現れる文書数を N_i とすると，第 n 文書の第 i 単語に対する $tf \cdot idf_{n,i}$ には，本稿では次の定義式を用いる：

$$tf \cdot idf_{n,i} = tf_{n,i} \cdot \log \frac{N}{N_i} \tag{1}$$

ここで，$tf_{n,i}$ は第 n 文書の第 i 単語の出現回数そのものを表す．また，$\sqrt{\sum_i {tf \cdot idf_{n,i}}^2} = 1$ となるように正規化を行うこともある．

3.2 *LSA* によるトピック抽出

文書特徴ベクトルを V 次元-列ベクトル $\mathbf{x}_n = (x_{n,1}, \cdots, x_{n,V})^t$ で表すことにする．ここで，$x_{n,i}$ は $tf \cdot idf_{n,i}$ などによる単語の重みを表す．この V 次元空間内において，文書群 $\{\mathbf{x}_n : n = 1, \cdots, N\}$ の分布を考えるとき，その分散が最大となる軸に第 1 トピックがあり，それに直交する軸で次にその分散が最大となる軸に第 2 トピックがある，というようにトピック軸が存在すると考える．したがって，$||\mathbf{u}|| = 1$ の下で，

$$E(\mathbf{u}) = \sum_{n=1}^{N} < \mathbf{x_n}, \mathbf{u} >^2 \tag{2}$$

なる目的関数を最大化することにより，トピック軸 $\mathbf{u}$ を抽出する．ここで，$<,>$ は内積を表す．これは，$V \times V$ 行列 $B = (b_{ij})$ の固有値問題として求められる．ここに，

$$b_{ij} = \sum_{n=1}^{N} x_{n,i} \cdot x_{n,j} \tag{3}$$

である．すなわち，B の第 k 固有値の固有ベクトル $\mathbf{u_k}$ が第 k トピック軸方向の単位ベクトルである．この単位ベクトルを第 k トピックを表すトピックベクトルと呼ぶことにする．

3.3 *PCA* によるトピック抽出

LSA に類似した手法に主成分分析（*PCA*）がある．*PCA* では，特徴軸の原点を文書群の中心に平行移動する点が *LSA* と異なる．すなわち，目的関数は次式となる：

$$E(\mathbf{u}) = \sum_{n=1}^{N} < (\mathbf{x_n} - \bar{\mathbf{x}}), \mathbf{u} >^2 \tag{4}$$

ここで，$\bar{\mathbf{x}}$ は全文書の特徴ベクトルの中心を表している．そして，行列 B は，

$$b_{ij} = \sum_{n=1}^{N} (x_{n,i} - \bar{x}_i) \cdot (x_{n,j} - \bar{x}_j) \tag{5}$$

で与えられる．*LSA* の場合と同様に，B の第 k 固有値の固有ベクトル $\mathbf{u_k}$ が第 k トピックを表すトピックベクトルである．

3.4 次元縮約

大規模文書群の分類・可視化[4),12)]や情報検索[13)]に応用する場合，その特徴空間は数千文書で数万次元と，超高次元になるためメモリ容量や計算量の問題がある．しかしながら，一般に文書特徴ベクトルは幸いスパースなベクトルになるため，その特性を活かし次元縮約が行われる．適切な成分への次元縮約は，縮約前後で任意のベクトル間距離は近似的に保存されることが知られている[14)]．本稿での *SKM* によるトピック抽出や提案するトピックの解釈可能度の計算においても，メモリや計算量の問題から次元縮約を行う．

具体的には，文書特徴空間を，*LSA* や *PCA* により抽出した特徴軸群を基底とする空間へ射影することにより次元縮約する．具体的には，v 個の特徴軸を抽出したとき，V 次元から v 次元への基底の変換行列を $v \times V$ 行列 $\mathbf{R} = (\mathbf{u}_1, \cdots, \mathbf{u}_v)$ として，射影後の文書特徴ベクトルを $\Phi_n = (\phi_{n,1}, \cdots, \phi_{n,v})^t$ とすると，次式により内積計算のみで簡単に得られる：

$$\Phi_n = \mathbf{R}^t \mathbf{x}_n \tag{6}$$

このとき，*LSA* や *PCA* による抽出トピックベクトルは，射影後の空間では基本ベクトル，すなわち $\{\mathbf{e}_k : k = 1, \cdots, v\}$（$k$ 番目の要素のみ 1 でその他はすべて 0 のベクトル群）となる．

3.5 Spherical *k*-means 法によるトピック抽出

k-means 法は非階層型クラスタリングの代表的手法である．文書群のクラスタリングにおいては，文書間類似度として *cosine* 類似度が広く用いられている．このため，超球面上のクラスタリングになるため，Spherical *k*-means と呼ばれる．アルゴリズムを次に示す．

(1) K 個の中心ベクトル Ψ_k（$k = 1, \cdots, K$）の初期値をランダムに与える．
(2) 中心ベクトルと文書特徴ベクトル間の *cosine* 類似度に基づいてサンプルをクラスタに分類する．
(3) 分類されたクラスタの文書集合により中心ベク

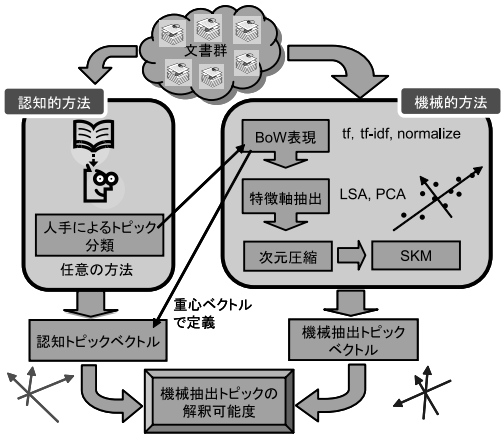

図 1 提案評価方法のフローチャート
Fig. 1 The flow chart of proposed evaluation methodology.

トルを更新する．

(4) 中心ベクトルの変化がなくなるまで (2)，(3) を繰り返す．

ここで，$||\Psi_k|| = ||\Phi_n|| = 1$ に正規化されているとき，*cosine* 類似度は次式により与えられる：

$$\cos\theta_{k,n} =< \Psi_k, \Phi_n > \tag{7}$$

そして，最終的に得られる K 個のクラスタの中心ベクトルをトピックベクトルとする．

4. トピックの解釈可能度

4.1 概　　略

本稿では，特徴軸として抽出されるトピックの解釈可能度を定量化することを試みた．提案評価方法の流れ図を**図 1** に示す．一方では，*LSA* や *PCA*，もしくは *SKM* などによってトピック抽出を行う．他方では同じ文書群に対して任意の認知的な方法でトピック分類がなされているものとする．それらを認知トピックと呼ぶことにし，認知トピックごとの重心ベクトルにより認知トピックを代表する認知トピックベクトルを構成する．両者のトピックベクトルを用いて機械抽出トピックの解釈可能度を定義した．

基本的な考え方は，両極端の場合，ある機械抽出トピックが認知トピック群の線形結合で一致できれば，その機械抽出トピックは認知トピック群を説明変数として解釈可能であり，逆にどの認知トピックとも直交したならば，解釈不可能であると考える（**図 2**）．このことから，認知トピック群の線形和での近似度合いにより機械抽出トピックの解釈可能度を定式化する．

しかし，たとえ線形結合による近似度合いが高かったとしても，必ずしも少数の認知トピックで説明されるとは限らない．そのため，各認知トピックとの類似度と併用して解釈する必要がある．しかしながら，逆に線形結合での近似度合いが低い場合，客観性のある認知トピック群が得られているとしたら，それらには含まれないようなトピックで構成されている機械抽出トピックは解釈困難と考えられる．すなわち，提案するトピックの解釈可能度は，ほとんど意味を持たない抽出トピックを除外するフィルタの役割は果たすと考えられる．

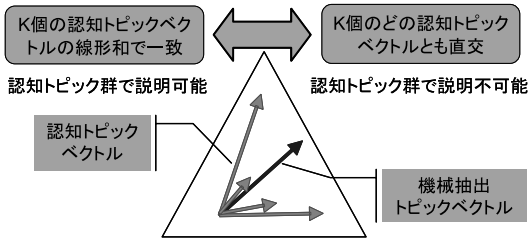

図 2 トピックの解釈可能度の幾何学的解釈
Fig. 2 Topological interpretation of interpretability of a topic.

4.2 定　式　化

ある機械抽出トピックベクトル Ψ（*LSA* や *PCA* では固有ベクトル，*SKM* では中心ベクトルに対応）に近くなるように，K 個の認知トピック群の重み付き和ベクトル $\sum_{k=1}^{K} w_k\tilde{\Psi}_k$ の K 次元–重みベクトル $\mathbf{w}$ を求める．ただし，$||\Psi|| = ||\tilde{\Psi}_1||\cdots||\tilde{\Psi}_k|| = 1$ に正規化されているとする．各ベクトルを列ベクトルとすると，$v \times K$ 行列 $\mathbf{A} = [\tilde{\Psi}_1, \cdots, \tilde{\Psi}_K]$ を定義できる．すると，重み付き和ベクトルは $\sum_{k=1}^{K} w_k\tilde{\Psi}_k = \mathbf{Aw}$ と表される．

いま，ベクトル Ψ と $\mathbf{Aw}$ の *cosine* 類似度の自乗を最大化するように，ベクトル $\mathbf{w}$ を求めるとする．すると，目的関数は以下で定義できる．

$$P = \frac{(\Psi^t\mathbf{Aw})^2}{\mathbf{w}^t\mathbf{A}^t\mathbf{Aw}} \tag{8}$$

上記目的関数は数理物理学における一般化レイリー商（*generalized Rayleigh quotient*）と同形である[15]．その解 $\hat{\mathbf{w}}$ は次式により与えられる（導出は付録を参照）：

$$\hat{\mathbf{w}} \propto (\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\Psi \tag{9}$$

ここで，目的関数は *cosine* 自乗で評価しているため，$\mathbf{w}$ の定数倍には不変なことに注意する．また，認知トピック群である行列 $\mathbf{A}$ の列ベクトルは，人が分類したトピックであるため一次独立であると考えられる．これはたとえば，$|\mathbf{A}^t\mathbf{A}|$ は *Gram* の行列式であるから，$|\mathbf{A}^t\mathbf{A}| \neq 0$ で行列 $\mathbf{A}$ の列ベクトルの一次独立性は確かめられる．ただし，一次独立であるために

表 1 データセットの基本統計量
Table 1 Basic statistics of the data sets.

	総文書数	総単語種類数	トピック数
毎日新聞	2,694	18,070	39
TDT3	34,413	78,452	115

表 2 認知抽出トピック一覧（毎日新聞データ）
Table 2 The list of cognitively extracted topics (Newspaper *Mainichi* data set).

No.	トピックタイトル
1	北朝鮮核査察問題
2	南北対話問題
3	南北特使交換問題
4	米朝実務会談
5	北朝鮮制裁問題
6	IAEA 脱退宣言
7	金日成・カーター会談
8	南北首脳会談
9	米朝対話再開
10	米朝高官協議
11	アフガニスタン内戦
12	サパティスタ民族解放軍武装蜂起
13	カンボジア問題
14	PLO とイスラエル問題
15	南アフリカ総選挙情勢
16	ボプタツワナ問題
17	シスカイ問題
18	ホワイトウォーター問題
19	ロシア PFP 調印
20	ボスニア紛争情勢
21	和平協議
22	四者会議
23	七カ国外相会談
24	分割地図協議
25	ソマリア問題
26	ベトナム国交正常化問題
27	カザフスタン総選挙
28	エルサルバドル大統領総選挙
29	クリミア大統領選挙
30	クリミア独立住民投票
31	ウクライナ議会選挙
32	バルチン黒海艦隊問題
33	クリミア独立問題
34	ルワンダ内戦
35	イエメン内戦
36	マラウイ自由選挙
37	モルドバ国会選挙
38	東ティモール国際会議問題
39	タイ民主憲法制定問題

はベクトルの数よりも次元数の方が多い必要がある．すなわち，$K \leq v$ を満たす必要がある．このとき，行列 $\mathbf{A}^t\mathbf{A}$ は非特異となるため解は存在する．したがって，機械抽出トピックの解釈可能度 p は次で与えられる：

$$p = \frac{(\Psi^t \mathbf{A}\hat{\mathbf{w}})^2}{\hat{\mathbf{w}}^t \mathbf{A}^t \mathbf{A}\hat{\mathbf{w}}} \tag{10}$$

このとき，p は [0,1] の値をとり，$p = 1$ のとき，その機械抽出トピックは，認知トピック群の線形結合で *cosine* 類似度の意味で一致させることが可能であることを意味する．一方，機械抽出トピックと認知トピック群は互いに直交するとき，$p = 0$ となる．

また，上記目的関数では *cosine* 類似度の自乗で評価しているため，*cosine* 類似度が負になる場合でも両者は近いと評価される．しかし，その解は $\mathbf{w}$ の定数倍に対して不変であるので，$-\mathbf{w}$ ととることで *cosine* 類似度として近くなる $\mathbf{w}$ を得ることができる．

表 3 認知抽出トピック一覧（TDT3 データから抜粋）
Table 3 The list of cognitively extracted topics (TDT3 data set).

No.	トピックタイトル
1	Cambodian Government Coalition
2	Hurricane Mitch
3	Pinochet Trial
4	Chukwu Octuplets Born in Houston
5	Osama bin Laden Indictment
6	NBA Labor Disputes
7	Congolese Rebels vs. Pres. Kabila
8	November APEC Summit Meeting
9	Anti-Doping Proposals
10	Car Bomb in Jerusalem

5. 検証実験

5.1 データセット

本実験では，1994 年 1 月〜6 月までの毎日新聞国際面記事（日本語），および TDT3☆で公開されているデータセット（英語）の 2 種類を用いた．各データセットの基本統計量を**表 1** に示す．毎日新聞記事については，斉藤らの認知科学的実験[16] によってトピックが付与されたデータセットを用いた．そこで抽出されたトピック一覧を**表 2** に示す．一方，TDT3 のデータセットには，1998 年 10 月〜12 月までの New York Times などの新聞記事や CNN，ABC ニュース番組などの音声から抽出された文字情報が含まれており，各文書はガイドラインに従って人手によってトピックが付与されている．そのトピックの一部を**表 3** に示す．また，BOW に用いる単語集合を選択する前処理として，不要な単語を設定し stop word の除去を行い，TDT3 のデータに関しては英語の語幹を取り出す Porter stemming[17] も施した．

5.2 人工データ

次に，比較対象として 2 種類の人工トピックベクトルを用意した．1 つは，ランダムに生成した正規直交基底をトピック軸とするランダムトピックである．こ

☆ http://projects.ldc.upenn.edu/TDT3/

れはたとえば，ランダムにベクトル群を生成し，*Gram* の行列式によりそれらベクトル群の一次独立性を確認した後，*Gram-Schmidt* の直交化法により得られる．もう 1 つは，認知トピック群の線形結合で構成したトピックベクトルに乱数ノイズを加えたノイズ付き人工トピック Ψ^* である．これは次式により生成した：

$$\Psi^* = r\sum_{k=1}^{K} s_k\tilde{\Psi}_k + (1-r)\Theta \qquad (11)$$

ここで，s_k は $\sum_k s_k = 1, s_k \in [0,1]$ を満たす乱数による第 k 認知トピックの結合加重である．Θ は $||\Theta|| = 1$,（ベクトルの各要素）$\in [0,1]$ を満たす乱数による v 次元–ノイズベクトルである．$r \in [0,1]$ は人工トピックベクトルとノイズベクトルとのバランスを表す重みである．また，$||\Psi^*|| = 1$ に正規化した．

5.3 検証の考え方

情報検索の分野では正解文書集合を取り出すことが目標であるため，システムが出力した文書集合と正解集合との合致率を定量的に評価できる．このため多くの事例研究から明かになっている事柄がある．たとえば，多くの場合，文書特徴ベクトルに用いる単語の重み付けとして，単純な頻度である *tf* よりも，単語の普遍性を考慮した *tf-idf* の方が良い結果が得られ，またノルム 1 に正規化した方が良い結果が得られている[11]．

一方，*LSA* や *PCA* によるトピック抽出はその性質上，分散の大きい方が順番に軸を抽出していくため，下位の抽出軸ほど特徴的ではなくなっていくと考えられる．本稿では，特に上位数トピックに関しては人にとって何らかの有意なトピック軸を抽出していると仮定する．ここで，有意なトピック軸とは，その軸近傍の文書群は同じトピックに属すると人は判断するような軸を指す．ゆえに，認知トピック群の線形結合で近似できる可能性が高く，トピックの解釈可能度は高い評価が得られると考えられる．一方，ランダムトピックは何ら文書に関する情報を用いていないので，有意ではないトピック軸であり，低い評価になると考えられる．

本検証実験では，これらの仮定や情報検索分野での事実に基づいて，本提案評価法によってトピック抽出においても同様のことが示せることを確かめる．すなわち，有意であると仮定もしくは考えられるトピックに関する評価が高く，そうでなければ低い結果が得られたならば，提案評価方法が正しく機能している事実の 1 つになると考える．

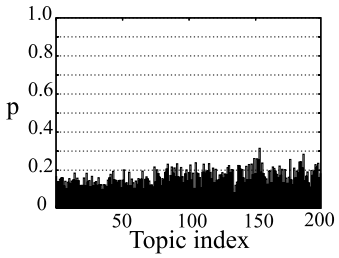

図 7 ランダムトピックの解釈可能度（毎日新聞データ）
Fig. 7 Interpretability of random topics (Newspaper *Mainichi* data set).

5.4 *LSA* や *PCA* を用いた検証

5.4.1 トピック軸抽出順による比較

毎日新聞データおよび TDT3 データに対して，*LSA* および *PCA* による抽出トピックの解釈可能度を評価した結果を**図 3**，**図 4**，**図 5**，**図 6** に示す．図の縦軸は解釈可能度 p，横軸は抽出トピックインデックス，すなわち式 (3) の第 k 固有ベクトルに対応するトピックを表している．どちらのデータセットにおいても，また *LSA* と *PCA* おいても，さらに (a)〜(d) いずれの重み付けにおいても，評価値は上位数トピックは高く，その後急速に減少していっている．このことは，*LSA* や *PCA* による抽出法の特性と一致している．一方，ランダムトピックの場合（**図 7**）にその傾向は見られず，トピックごとの際だった差はない．これは明らかに解釈可能度は，認知トピックとは無関係なランダムトピックには反応せず，有意であると考えられる抽出トピックに反応していることを示している．

5.4.2 重み付けによる違い

一方，異なる重み付けによる抽出トピックの解釈可能度を比較する．**表 4**，**表 5**，**表 6**，**表 7** は，図 3〜図 6 に対応している．それぞれ (a) *tf*, (b) *tf-idf*, (c) *normalized tf*, (d) *normalized tf-idf* を表しており，表中の数値は，*LSA* や *PCA* によって抽出されたトピックのうち，一定以上の解釈可能度が得られたトピック数を表している．また，最下段は抽出トピック中の最大値を表している．

まず，*LSA* による抽出トピックについては，毎日新聞データ（表 4）では，ベクトルを正規化していない (a), (b) よりも正規化した (c), (d) の方が明らかに多くのトピックで高い評価値を得ている（(a) と (c), (b) と (d) を比較）．しかし，*tf* のみと *tf-idf* を比較すると（(a) と (b), (c) と (d) を比較），(a), (c) の方が $p > 0.9$ のトピックを若干多く抽出しているが，$p > 0.8$ 以下は大きな差はなく，また最大値は (d) が最も高いため優劣はつけ難い．一方，TDT3 データ

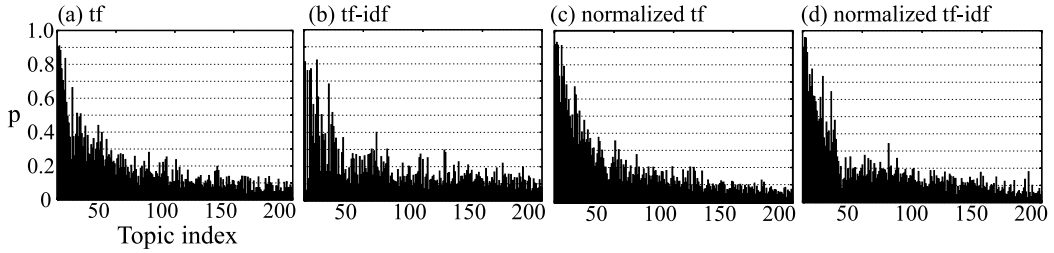

図 3 *LSA* による抽出トピックの解釈可能度（毎日新聞データ）
Fig. 3 Interpretability of the topics extracted by *LSA* (Newspaper *Mainichi* data set).

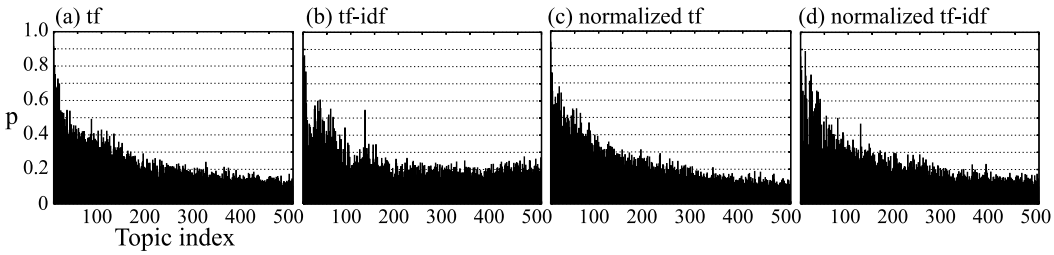

図 4 *LSA* による抽出トピックの解釈可能度（TDT3 データ）
Fig. 4 Interpretability of the topics extracted by *LSA* (TDT3 data set).

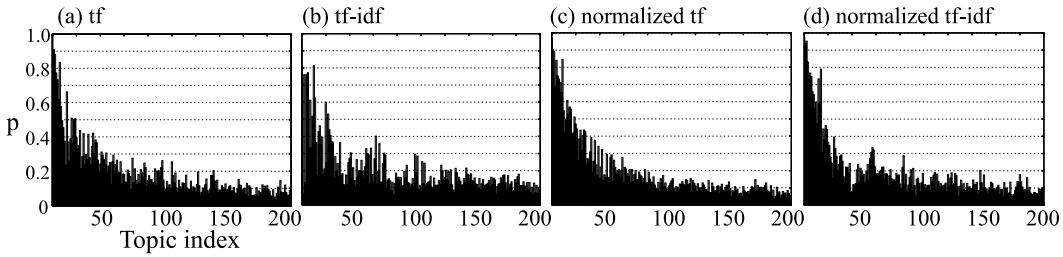

図 5 *PCA* による抽出トピックの解釈可能度（毎日新聞データ）
Fig. 5 Interpretability of the topics extracted by *PCA* (Newspaper *Mainichi* data set).

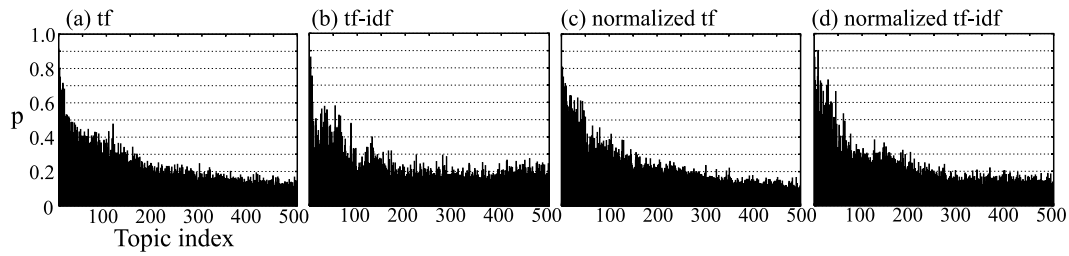

図 6 *PCA* による抽出トピックの解釈可能度（TDT3 データ）
Fig. 6 Interpretability of the topics extracted by *PCA* (TDT3 data set).

（表 5）では，(a) と (c) はほとんど差はないが，(b) と (d) を比較すると，正規化した (d) の方が $p > 0.5$〜> 0.7 でより多くのトピックが得られている．また，*tf* のみと *tf-idf* を比較すると，(a) と (b) では，最大値は (a) の方が高いが，(b) の方が $p > 0.8$ のトピックを 2 つ多く抽出しており優劣はつけ難い．しかし，(c) と (d) を比較すると，最大値は (c) の方が高いが (d) では $p > 0.5$〜> 0.8 のトピック数が多いため，全体としては (d) の方が若干良い結果であると考えられる．TDT3 データは毎日新聞データに比べて文書数・トピック数ともに大規模であるため，全体的に毎日新聞データほど顕著な差が出なかったと考えられる．

表 4 解釈可能度の各階級のトピック数および最大値（毎日新聞データ，*LSA*）
Table 4 The number of topics for each rank and the maximum value of interpretability (Newspaper *Mainichi* data set, *LSA*).

p	(a)	(b)	(c)	(d)
> 0.9	2	0	**4**	3
> 0.8	2	2	**4**	**4**
> 0.7	4	5	**7**	**7**
> 0.6	8	6	10	**13**
> 0.5	9	11	**17**	15
> 0.4	16	13	22	**23**
> 0.3	31	23	**35**	27
MAX	0.9094	0.8229	0.9134	**0.9597**

表 5 解釈可能度の各階級のトピック数および最大値（TDT3 データ，*LSA*）
Table 5 The number of topics for each rank and the maximum value of interpretability (TDT3 data set, *LSA*).

p	(a)	(b)	(c)	(d)
> 0.9	**1**	0	**1**	0
> 0.8	1	**3**	1	2
> 0.7	4	4	2	**6**
> 0.6	9	5	9	**16**
> 0.5	14	17	23	**26**
> 0.4	49	40	**54**	45
> 0.3	**116**	70	107	97
MAX	0.9140	0.8632	**0.9364**	0.8883

一方，*PCA* による抽出トピックについても，*LSA* の場合と同様の傾向が見られる．毎日新聞データ（表 6）では，(a) と (c)，(b) と (d) を比較すると，正規化した方が明らかに解釈可能度のより高いトピックが多数得られている．しかし，*tf* のみ (a) 方が *tf-idf* の (b) より最大値も高く，若干多くのトピックを抽出している．(c) と (d) の比較では (d) の方がより多く評価値の高いトピックが得られている．一方，TDT3 データ（表 7）では，(a) と (c) を比較すると，最大値は (a) の方が高いが (c) は $p > 0.6$ のトピックが多いため優劣はつけ難い．(b) と (d) を比較すると，明らかに正規化した (d) 方が良い結果が得られている．また，*tf* のみと *tf-idf* の比較では，(a) と (b) では最大値は (a) の方が高いが $p > 0.8$ のトピックを 2 つ多く抽出しており優劣はつけ難い．(c) と (d) では最大値，トピック数ともに (d) の方が良い結果となった．

総合すると，*LSA*・*PCA* のどちらの抽出トピックにおいても，毎日新聞データでは正規化した方が評価値の高いトピックが明らかに多く得られ，TDT3 データでは同様に正規化の効果は若干見られ，また正規化した *tf* よりも正規化した *tf-idf* の方が良い結果となった．以上より，検証の仮定とおおむね一致していると考えられる．

表 6 解釈可能度の各階級のトピック数および最大値（毎日新聞データ，*PCA*）
Table 6 The number of topics for each rank and the maximum value of interpretability (Newspaper *Mainichi* data set, *PCA*).

p	(a)	(b)	(c)	(d)
> 0.9	2	0	1	**3**
> 0.8	4	2	**5**	4
> 0.7	6	5	8	**9**
> 0.6	8	8	**11**	**11**
> 0.5	12	10	**16**	14
> 0.4	18	14	**26**	18
> 0.3	30	28	**33**	27
MAX	0.9089	0.8129	0.9700	**0.9725**

表 7 解釈可能度の各階級のトピック数および最大値（TDT3 データ，*PCA*）
Table 7 The number of topics for each rank and the maximum value of interpretability (TDT3 data set, *PCA*).

p	(a)	(b)	(c)	(d)
> 0.9	**1**	0	0	**1**
> 0.8	1	**3**	1	2
> 0.7	4	4	4	**7**
> 0.6	9	5	**15**	14
> 0.5	17	17	29	**32**
> 0.4	49	39	**52**	50
> 0.3	**115**	78	98	88
MAX	**0.9138**	0.8624	0.8093	0.9019

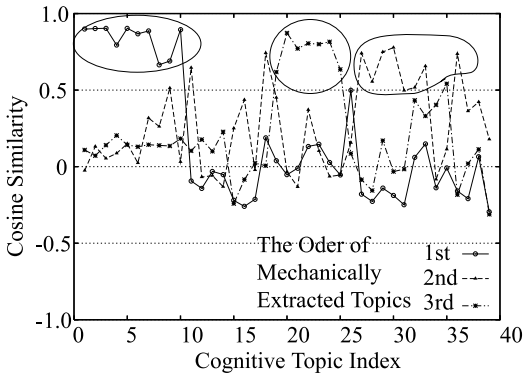

図 8 *LSA* による Top3 抽出トピックの認知トピック群に対する *cosine* 類似度（毎日新聞データ）
Fig. 8 Cosine similarity of Top 3 topics extracted by *LSA* to the cognitively extracted topics (Newspaper *Mainichi* data set).

5.4.3 認知トピックとの対応関係

ここで，実際に *LSA* による抽出トピックがどのような認知トピック群で説明されるのかを例示する．**図 8** は上位 3 抽出トピックの各認知トピックに対する類似度を示している．縦軸は *cosine* 類似度を，そして横軸は表 2 に対応する認知トピックのインデックスを表

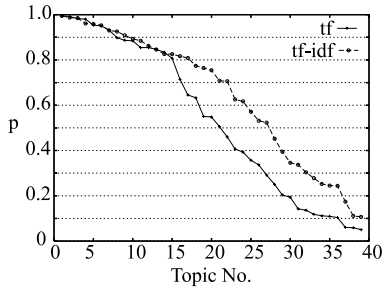

図 9 *SKM* による抽出トピックの解釈可能度（毎日新聞データ）
Fig. 9 Interpretability of the topics extracted by *SKM* (Newspaper *Mainichi* data set).

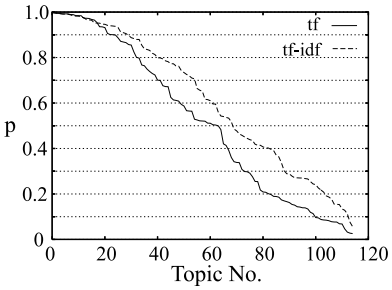

図 10 *SKM* による抽出トピックの解釈可能度（TDT3 データ）
Fig. 10 Interpretability of the topics extracted by *SKM* (TDT3 data set).

している．第 1 抽出トピックは，認知トピック No.1 から No.10 に対する類似度がきわめて高い．表 2 から解釈すると，No.1 から No.10 から第 1 抽出トピックは「北朝鮮の外交問題」と大きくとらえられる．同様に，第 2 抽出トピックは「各国の選挙」，第 3 抽出トピックは「中東諸国の紛争問題」と大まかに解釈できる．

5.5 *SKM* を用いた検証

次に，*SKM* による抽出トピックの解釈可能度を**図 9** および**図 10** に示す．縦軸はトピックの解釈可能度を表し，横軸は抽出トピックを解釈可能度の降順に並べてある．*SKM* における中心ベクトルの初期値依存性を排除するため，初期値を変えた 100 回のうち，K 個の抽出トピックの解釈可能度の平均値が最も高かったときの結果を用いた．抽出するトピック数（クラスタ数）は認知トピック数と同数，すなわち毎日新聞データでは $K = 39$，TDT3 データでは $K = 115$ とした．また，文書間類似度として *cosine* 類似度を用いているため，ベクトルのノルムを 1 に正規化することによる違いはない．そのため，*tf* および *tf-idf* についてのみ記載してある．

図 9，図 10 のどちらにおいても，解釈可能度の高いトピックもあれば低いトピックもある．また毎日新聞データではおよそ $p < 0.8$，TDT3 データではおよそ $p < 0.95$ のトピックについては *tf-idf* の方が *tf* より明らかに高くなっており，有意であると仮定している *tf-idf* の方が良い結果となった．

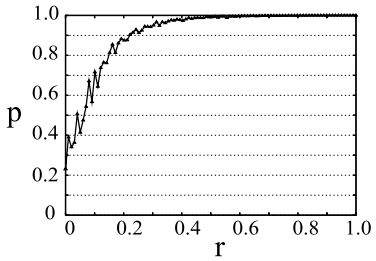

図 11 トピックの解釈可能度とノイズの関係（毎日新聞データ）
Fig. 11 Relationship between interpretability of the topic and noise (Newspaper *Mainichi* data set).

5.6 解釈可能度とノイズとの関係

最後に，**図 11** にノイズ付き人工トピックを用いて，ノイズ項とのバランス r を変化させて評価した結果を示す．r の値が小さい，すなわちノイズの大きいとき，評価値はランダムトピック時に近い値を示し，そしてノイズ項の減少（r 増加）にともなって急速に上昇した．$r = 1$ のとき，Ψ^* は認知トピックの線形結合になるため，目的関数式 (8) からも明らかに $p = 1$ となる．

5.7 課　　題

評価尺度が正しく働いていることを検証するのは難しい問題である．その尺度がある手法の何らかの性能を正しく測定できることは，それは異なる手法間の差分を定量的に明らかにできることを意味する．その意味において，評価尺度は次の性質を持つ必要がある．

(1) ある性能を客観的に計測していること
(2) 性能の優劣と高い相関があるとともに，十分な分解能や線形性を有すること
(3) 多様なデータセットや（適用可能な）異なる手法において評価が安定していること

本稿では，評価尺度を検証するにあたって，従来研究の知見に基づいたいくつかの仮定をもとに異なる手法間の差異を議論することで (1)，(2) の検討を行ったが，その他の観点における検討も必要である．(1)，(2) について，抽出トピックの解釈可能度として，線形結合での近似度合いによって定義していたが，人が解釈したときの解釈可能度合いとの相関についての検討も必要である．また，より微妙な差異を検出するためには，図 11 に示すような特性曲線に線形性があり，高い分解能を持つ必要がある．(3) については日本語

および英語の規模の異なるデータセットを用いて，異なる手法については *LSA*，*PCA*，*SKM* によるトピック抽出法を適用し，4 種類の単語の重み付けを比較した．さらに信頼性を得るためには別のデータセットや異なるトピック抽出法の適用も必要である．

また一方で，トピックの解釈可能度における認知トピック群の重みベクトル成分の偏り具合は重要であると考えられる．なぜなら，極端な場合，成分が一様に分布している場合は，その抽出トピックは全認知トピックに関するトピックと考えられる．本稿で定式化した解釈可能度としては高い値となっても，結局は特定のトピックを表していない可能性がある．ゆえに，重みベクトル成分の偏り具合と実際の認知トピックとの対応関係についても検討する必要がある．

6. ま と め

本稿では，人手によってトピック分類された文書群を用いて，*LSA* を代表とするベクトル表現可能な機械抽出トピックの解釈可能度の定量的評価方法を提案した．基本的な考え方は，ある機械抽出トピックに対して認知トピック群を説明変数とし多対 1 で対応すると考え，ある機械抽出トピックベクトルに対して認知トピックベクトル群の線形結合での近似度合いにより定式化した．そして，実際のニュース記事などの日本語および英語の 2 種類のデータセットから *LSA*，*PCA* および Spherical k-means 法によって抽出したトピック群，および人工データのトピック解釈可能度を評価した結果，従来研究の知見に基づく仮定の範囲においてはおおむね妥当な結果が得られた．ただし，解釈可能度において求まる重みベクトルと実際の認知トピックとの対応関係については今後検討が必要である．

参 考 文 献

1) Schultz, J.M. and Liberman, M.: Topic Detection and Tracking using idf-Weighted Cosine Coefficient, *Proc. DARPA Broadcast News Workshop*, pp.189–192 (1999).
2) Landauer, T.K. and Dumais, S.T.: A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge, *Psychological Review*, Vol.104, pp.211–240 (1997).
3) Kimura, M., Saito, K. and Ueda, N.: Multinominal PCA for extracting major latent topics from document streams, *Proc. 2005 International Joint Conference on Neural Networks*, pp.238–243 (2005).
4) Fukui, K., Saito, K., Kimura, M. and Numao, M.: Visualizing Dynamics of the Hot Topics Uing Sequence Based Self-Organizing Maps, *Lecture Note in Computer Science*, Vol.3684, pp.745–751 (2005).
5) 濱本雅史，北川博之，Pan, J.Y.，Faloutsos, C.：独立成分分析を用いたテキストデータからのトピック検出，電子情報通信学会第 15 回データ工学ワークショップ（DEWS2004）(2004).
6) Swan, R. and Jensen, D.: TimeMines: Constructing Timelines with Statistical Models of Word Usage, *KDD-2000* (2000).
7) Morinaga, S. and Yamanishi, K.: Tracking Dynamics of Topic Trends Using a Finite Mixture Model, *KDD2004* (2004).
8) Allan, J., Carbonell, J., Doddington, G., Yamron, J. and Yang, Y.: Topic Detection and Tracking Pilot Study: Final Report, *Proc. DARPA Broadcast News Transcription and Understanding Workshop* (1998).
9) Wayne, C.: Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation, *Language Resources and Evaluation Conference*, pp.1487–1494 (2000).
10) 斉藤和巳，上田修功，金田有二：確率モデルを用いた文書分類体系間の構造マッチング，情報処理学会研究報告自然言語処理，Vol.2004, No.47, pp.33–38 (2004).
11) Salton, G. and Buckley, C.: Term-weighting Approaches in Automatic Text Retrieval, *Information Processing & Management*, Vol.24, No.5, pp.513–523 (1988).
12) Kohonen, T., Kaski, S., Lagus, K., Salojrvi, J., Honkela, J., Paatero, V. and Saarela, A.: Self organization of a massive document collection, *IEEE Trans. Neural Networks*, Vol.11, No.3, pp.574–585 (2000).
13) 佐々木稔，大谷貴志，北　研二：情報検索のための概念ベクトル生成手法，電子情報通信学会技術報告，Vol.102, No.181, pp.29–34 (2002).
14) Kaski, S.: Dimensionality Reductions by Random Mapping: Fast Similarity Computation for Clustering, *Proc. Int. Joint Conf. on Neural Networks* (*IJCNN'98*), pp.413–418 (1998).
15) Duda, R.O., Hart, P.E. and Stork, D.G.: *Pattern Classification*, 2nd Edition, Wiley-Interscience (2000).
16) 斉藤和巳，木村昌弘，上田修功：文書トピックに関する認知科学的実験，*SIG-KBS-A405-10*, pp.57–62 (2005).
17) Jones, K.S. and Willet, P.: *Readings in Information Retrieval*, Morgan Kaufmann, San Francisco (1997).

付　　録

$$P=\frac{(\Psi^t\mathbf{Aw})^2}{\mathbf{w}^t\mathbf{A}^t\mathbf{Aw}}=\frac{(\mathbf{w}^t\mathbf{A}^t\Psi)(\Psi^t\mathbf{Aw})}{\mathbf{w}^t\mathbf{A}^t\mathbf{Aw}} \quad (12)$$

これより，定数 λ として，

$$\mathbf{A}^t\Psi\Psi^t\mathbf{Aw}=\lambda\mathbf{A}^t\mathbf{Aw} \quad (13)$$

が成り立つことが分かる．行列 $\mathbf{A}^t\mathbf{A}$ が正則ならば，

$$(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\Psi\Psi^t\mathbf{Aw}=\lambda\mathbf{w} \quad (14)$$

となり，これは，行列 $(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\Psi\Psi^t\mathbf{A}$ に対する固有値問題である．ここで，行列 $\Psi^t\mathbf{Aw}$ が 1 次元に縮退していることに着目すると，式 (14) の左辺は，

$$(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\Psi(\Psi^t\mathbf{Aw})\propto(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\Psi \quad (15)$$

となり，これより固有ベクトル $\hat{\mathbf{w}}$ は唯一，

$$\hat{\mathbf{w}}=(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\Psi \quad (16)$$

となる．

(平成 18 年 4 月 27 日受付)
(平成 18 年 6 月 16 日再受付)
(平成 18 年 7 月 6 日採録)

福井　健一

2001 年名古屋大学情報文化学部自然情報学科中退（飛び級進学）．2003 年名古屋大学大学院人間情報学研究科物質・生命情報学専攻修士課程修了．2005 年より大阪大学産業科学研究所新産業創造物質基盤技術研究センター特任助手．機械学習，データマイニング等に興味を持つ．人工知能学会会員．

斉藤　和巳

1963 年生．1985 年慶應義塾大学理工学部数理科学科卒業．工学博士．同年 NTT 入社．1991 年より 1 年間オタワ大学客員研究員．神経回路網，機械学習，複雑ネットワーク等の研究に従事．現在，NTT コミュニケーション科学基礎研究所主任研究員（特別研究員），奈良先端科学技術大学院大学客員助教授．情報処理学会論文賞（1997 年），人工知能学会論文賞（1999 年）等受賞．電子情報通信学会，人工知能学会，日本神経回路学会，IEEE 各会員．

木村　昌弘

1989 年大阪大学大学院理学研究科数学専攻修士課程修了．同年日本電信電話（株）入社．現在，龍谷大学理工学部電子情報学科助教授．博士（理学）．ニューラルコンピュテーション，複雑系の数理モデリングおよび数理解析，Web マイニングの研究に興味を持つ．電子情報通信学会，人工知能学会，日本神経回路学会，日本応用数理学会，日本数学会各会員．

沼尾　正行（正会員）

1982 年東京工業大学工学部電気電子工学科卒業．1987 年同大学院情報工学専攻博士課程修了．工学博士．東京工業大学大学院情報理工学研究科計算工学専攻助教授を経て，2003 年より大阪大学産業科学研究所教授．1989〜1990 年スタンフォード大学 CSLI 客員研究員．人工知能，機械学習，関数型言語等の研究に従事．人工知能学会，日本認知科学会，日本ソフトウェア科学会，電子情報通信学会，AAAI，ACM 各会員．