大阪府立大学博士論文

独立成分分析を用いた
情報フィルタリングに関する研究

2007 年 1 月

横 井　健

目次

図目次

表目次

第 1 章

序論

近年、急速な情報化が進み、一般ユーザが数多くの電子化された情報を容易に入手できるようになっている。それらの情報を効率よく利用するために、ユーザが求める情報を探し出す情報検索システムが数多く開発され、とくに、キーワードを用いた検索システムが普及している。しかし、キーワード検索システムでは、多くのユーザにとって、必要な情報を得るための的確なキーワードを選別することは難しく、検索結果に不要な情報が含まれてしまうことがある。上述のような必要な情報が他の情報に埋もれてしまう現象を情報洪水 [1] と呼ぶ。このような状況を解決するために、検索結果の表示順序をソーティングするランキング手法や、ユーザに必要な情報のみを選び出す情報フィルタリング手法 [2][3][4][5] が提案、研究されている。

ユーザにとって必要な情報を選択する基準として、興味の有無が用いられ、その興味情報はユーザプロファイルと呼ばれる。ユーザプロファイルとしては、ユーザが興味のある単語で表現する方法 [6] や、過去にユーザが評価した情報からユーザの興味を抽出する手法 [7][8][9] が提案されている。ドキュメント検索システムにおいては、ドキュメントやユーザプロファイルは、ベクトル空間法 [10] により索引語と呼ばれるドキュメントの特徴を表わす単語に対する重みを要素とするベクトルで表現される。このとき、扱うドキュメントの増

大に伴い、索引語数が増え、ベクトルの次元数が増大するので、ユーザの興味抽出において、的確な興味を探索できなかったり、時間的な効率に問題点がある。

上述の問題を解決するため、ドキュメント中に含まれる潜在的意味を用いて、ドキュメントベクトルを変換する潜在的意味解析（LSA: Latent Semantic Analysis）と呼ばれる手法が提案されている [11]。潜在的意味とは、単語に対する重みのような表面的な特徴ではなく、統計的な解析などを施すことで浮かび上がってくる特徴である。LSA では、各索引語に付与された重みのドキュメント間の分散に着目して、ドキュメント間の相互関係を解析し、その潜在的意味を用いて、ドキュメントベクトルの次元削減を行なっている。一方、音声処理や画像処理の分野で注目を集めている独立成分分析（ICA: Independent Component Analysis）[12][13] を用いてドキュメント中の潜在的意味を抽出しようという試みも報告されている [14][16][17][18][19]。ICA では、独立性という性質に着目し、ドキュメント中の潜在的意味を解析している。ICA により求められる独立成分は、成分ごとにある特定のテーマに関連した索引語に対して大きな重みが付いており、それらをトピックと呼ぶ。一方で、LSA によって得られる潜在的意味に見られるような、ICA によって得られるトピックを用いてドキュメントを表現したり、情報フィルタリングにこれらのトピックを応用するといった研究は、現在までに報告されていない。

そこで、本論文では、ICA によって得られるトピックを情報フィルタリングに応用する手法を提案する。具体的には、トピックを用いてドキュメントベクトルを表現し、トピック空間上でユーザプロファイルを作成する手法と、トピックを用いて不必要な索引語を除去する 2 つの手法を提案し、日本語テストコレクションとして有名な NTCIR2[20] を用いて、提案手法の有効性を検証する。本論文の構成は以下のとおりである。

第 1 章では、本研究の背景ならびに目的を述べるとともに、研究内容の概要について述べる。

第 2 章では、ICA によって得られるトピックが構成する空間にドキュメントベクトルを

写像し、その空間上でユーザプロファイルの作成を行なう手法 [21] について述べる。従来手法の LSA では、潜在的意味空間上の基底ベクトルは直交性という性質に着目しているが、本手法では、独立性に着目することで、得られる潜在的意味にそのドキュメントに含まれるトピックという意味づけを与えることができる。また、ユーザプロファイル作成を作成する手法として、学習ドキュメントの重心を用いる方法と遺伝的アルゴリズムを用いる方法を採用する。以上の手順により、ユーザプロファイルの作成方法に関わらず、もとのドキュメントベクトルを直接用いる従来手法と比較して、ICA によって得られるトピックを用いてドキュメントを表現する提案手法が有効であることを確認する。

第 3 章では、ICA によって得られるトピックに対して、トピックの選択を行なうことで、ユーザプロファイルの精度を改善する手法 [22][23][24] について述べる。ICA によって得られたトピックを用いてユーザプロファイルの精度を改善しようとする際、不必要なトピック、ノイズ成分を除く必要がある。そこで、トピックの類似性に着目し、最大距離アルゴリズム（MDA: Maximum Distance Algorithm）を用いてトピックをクラスタリングし、トピックを類別する。一般的に MDA の距離関数としては、ユークリッド距離が利用されているが、ICA によって得られるトピックはスケールや順序に任意性があるので、ユークリッド距離ではその類似性を評価できない。そこで、トピックに含まれる索引語の重みの分布の類似性に着目し、それを測る尺度として、情報幾何の分野において注目されている統計的情報量を用いる。本手法では、Jensen-Shannon 情報量を MDA の距離関数に採用し、クラスタリングを行なう。さらに、トピックの選択を行なうことにより、推薦精度の改善を目指す。

第 4 章では、SVD と ICA を組み合わせた潜在的意味の解析手法と、その情報フィルタリングへの応用を検討する [25][26]。第 3 章の手法では、ICA によってトピックを求めた後、必要と思われるトピックを選別したが、ドキュメント数が増えるにしたがい、ICA の処理時間が増大する。そこで、ICA を適用する前に、累積寄与率に基づいて SVD の基底ベクトルを選択し、ノイズ削減と低次元化を行なう。また、それらの基底ベクトルが張る空間上で

ICA を適用することにより、ICA の処理時間の短縮やフィルタリング精度の改善を試みる。

第 5 章では、ICA を行列分解の一手法と考え、ドキュメントを構成する潜在的意味として ICA の混合行列を採用し、その潜在的意味を情報フィルタリングに応用する手法 [27] を提案する。従来、ICA の混合行列を潜在的意味とみなす研究 [15][18] が報告されており、その潜在的意味は、独立性や直交性などは仮定できないが、第 2 章の手法に比べ人間の考えるトピックの感覚に近い。そこで、それらの潜在的意味を用いて情報フィルタリングを行なうことで、先に述べた手法とは異なる結果が得られると考え、その潜在的意味空間上でユーザプロファイルを作成し、推薦精度の向上を目指す。

第 6 章では、ICA によって得られるトピックに基づいて索引語の選別を行ない、情報フィルタリングの精度を改善する手法を提案する [28][29]。従来、ドキュメント中から重要語を抽出するキーワード抽出や索引付けに関する研究 [30] が行なわれている。それらの研究では、主にドキュメント中の高頻度語に焦点が当てられている [31]。一方、情報フィルタリングに用いるドキュメントベクトルの索引語を作成する際には、ドキュメント集合中に含まれるある特定のドキュメントに含まれる索引語に絞るのではなく、すべてのドキュメントから索引語を作成しなければならない。したがって、高頻度語が表わすドキュメントの大まかなトピックだけではなく、様々なトピックに関する単語も抽出しなければ、精度の良い情報フィルタリングを実現することが困難である。そこで、ICA によってドキュメントのトピックを解析し、そのトピックを表わす代表的な語と共起する語を索引語として用いることで、様々なトピックに関連する特徴的な語を索引語として選別する。さらに、それらの索引語を用いてドキュメントベクトルを再構築し、学習ドキュメントの重心を用いてユーザプロファイルの作成を行ない、推薦精度の向上を目指す。

最後に、第 7 章では以上の本研究について総括を行ない、今後の検討課題について述べる。

第 2 章

ドキュメントに対する独立成分分析を用いた情報フィルタリング

2.1　はじめに

昨今の情報化技術の発展に伴い、多くの情報がインターネットを通じて提供されるようになってきた。しかし、それらの量が多くなるにつれて、従来のデータベースに基づいた検索手法では、ユーザが的確な検索要求を作成できず、膨大な検索結果が得られてしまい、必要な情報を探し出すことが困難になってきた。ついには、必要な情報が他の情報に埋もれてしまう、情報洪水と呼ばれる現象が発生してきている。

この問題を解決するために、ユーザの興味に基づいて、情報を自動的に選択する情報フィルタリングが研究されている [2][3]。上述のユーザの興味は、情報フィルタリングの分野において、ユーザプロファイルと呼ばれている。本章では、独立成分分析（ICA:Independent Component Analysis）を用いて、ユーザプロファイルにもとづいた情報フィルタリングの精度改善を図る。ICA は信号処理や画像処理などの分野で、昨今注目を集めている手法である。ICA をドキュメントに適用すると、ドキュメントに含まれるトピックが得られることが

報告されている。ここで、トピックとはドキュメントを構成するカテゴリを意味する。

本章では、まず、ドキュメントに対して ICA を適用することでトピックを取得し、そのトピック空間上へドキュメントを写像する。つまり、ドキュメントをトピックで表現する。また、トピック空間上で、ユーザプロファイルを作成する。最後に、作成されたユーザプロファイルによってドキュメントの推薦を行ない、本手法の評価を行なう。さらに、得られたトピックについての検討も行なう。

2.2 従来研究

ドキュメントを対象とした情報フィルタリングシステムにおいては、ベクトル空間法と呼ばれるドキュメントの表現方法がしばしば用いられている。ベクトル空間法では、ドキュメントは、索引語に対する重みを要素とするベクトルで表現する。それゆえ、扱うドキュメント数が多くなるにしたがって、索引語数も増大し、ベクトルの次元数が大きくなる。

上述の問題を解決するための手法として潜在的意味解析（LSA:Latent Semantic Analysis）が有名である。LSA は特異値分解を用いた解析手法で、特異値の小さな特異ベクトルを除去し、その特異ベクトルの張る空間にドキュメントベクトルを写像することで次元削減を実現する手法である。

ドキュメントを対象とした情報フィルタリングの従来手法としては以下の手法が挙げられる。まず、ユーザの興味（ユーザプロファイル）によってフィルタリングを行なう手法がある。よく知られたユーザプロファイルを作成する手法として学習ドキュメントの重心を用いた手法 [32] がある。この手法では、ユーザプロファイルを興味のあるドキュメントのベクトルとは近く、興味のないドキュメントのベクトルとは離れるように作成する。この手法は、他の情報フィルタリング手法の比較手法としても良く用いられる。サポートベクタマシーン（SVM:Support Vector Machine）を利用したフィルタリング手法 [33] も提案されて

いる。SVM は、クラス分類の問題に良く利用されている。この手法では、学習ドキュメントをサポートベクタとし、サンプル間と分離超平面のマージンが最大となるように、超平面を決定する。また、協調フィルタリングなどの手法 [34][35] も良く研究されている。この手法では、趣向のよく似たユーザの興味情報（ユーザプロファイル）を用いて、情報をフィルタリングする手法である。すでに、Amazon.com[36] などのイーコマースで実用化もされている。

2.3 ユーザプロファイル

本節では、ドキュメントのベクトル空間法による表現、ユーザプロファイルの作成方法、および、評価方法について述べる。

2.3.1 ドキュメントの表現方法

ベクトル空間法においてドキュメントは、索引語に対する重みを要素とする列ベクトルで表現される。以下では、これをドキュメントベクトルという。索引語とは、ドキュメントの特徴を表わす単語である。本論文では、ドキュメント中に含まれる名詞から、「こと」や「もの」といった単語を除いたものを索引語として用いる。ある i 番目のドキュメントに対して、索引語数を V とし、j 番目の索引語に対する重みを w_{ij} とすると、ドキュメントベクトル $\mathbf{d}_i$ は、

$$\mathbf{d}_i = [w_{i1} \quad w_{i2} \quad \cdots \quad w_{iV}]^t \tag{2.1}$$

で表わされる。$[\cdot]^t$ は転置を示し、w_{ij} は、tf-idf 法 [37] により求める。tf-idf 法による w_{ij} の計算式の一例を式 (2.2) に示す。

$$w_{ij} = tf_{ij} \log(\frac{n}{df_j}) \tag{2.2}$$

式（2.2）中の tf_{ij} は、単語頻度と呼ばれ、i 番目のドキュメント中の j 番目の索引語の出現頻度を表わしている。一方、df_j は、文書頻度と呼ばれ、j 番目の索引語を含むドキュメント数を表わしている。また、n は全ドキュメント数を表わす。

2.3.2 ユーザプロファイルの作成方法

ユーザの興味を抽出したユーザプロファイル $\mathbf{u}$ は、式 (2.3) に示すように、ドキュメントベクトルと同じく索引語に対する重みを要素とした列ベクトルとする。

$$\mathbf{u} = [u_1 \quad u_2 \quad \cdots \quad u_V]^t \tag{2.3}$$

ユーザプロファイルに含まれる各索引語に対する重みは、興味のある索引語に対しては大きな重み、興味のない索引語に対しては小さな重みが付く。また、その重みの決定には、ユーザが評価したドキュメントを用いる。本章では、ユーザプロファイルの作成手法として、従来提案されている学習ドキュメントの重心を用いて作成する手法と、遺伝的アルゴリズム（GA: Genetic Algorithm）[38] による手法を用いる。

まず、重心によって求められるユーザプロファイルは、式（2.4）となる。

$$\mathbf{u} = \alpha \sum_{\mathbf{d}_k \in D_I} \mathbf{d}_k + \beta \sum_{\mathbf{d}_l \in D_U} \mathbf{d}_l \tag{2.4}$$

ここで、$\mathbf{u}$ はユーザプロファイル、$\mathbf{d}_k$、$\mathbf{d}_l$ はそれぞれ興味ありのドキュメントベクトルと興味なしのドキュメントベクトル、α、β はそれぞれ興味のあるドキュメントと興味のないドキュメントに対する係数である。D_I と D_U はそれぞれ、興味のあるドキュメントの集合、と興味のないドキュメントの集合である。

つぎに、GA によるユーザプロファイルの作成方法について述べる。GA を用いたユーザプロファイルの作成手法は文献 [39] や文献 [40] で提案されている。ユーザプロファイルの作成に GA を用いる利点は、評価ドキュメントが多くなるに従いユーザの興味空間が大きくなった場合に、GA の持つ広範囲の探索能力が効果的に働く点にある。ユーザプロファイル

u は、各索引語に対する重み u_i を、5 ビットのバイナリで表現し、単純 GA を用いて、最適なユーザプロファイルを探索する。なお、本論文では興味のあるドキュメントとの内積値が大きく、興味のないドキュメントとの内積値が小さくなるように、次のような評価関数を用いる。

$$f(\mathbf{u}) = \alpha \sum_{\mathbf{d}_k \in D_I} \mathbf{u}^t \mathbf{d}_k + \beta \sum_{\mathbf{d}_l \in D_U} \mathbf{u}^t \mathbf{d}_l \tag{2.5}$$

2.3.3 ユーザプロファイルの評価方法

作成したユーザプロファイルと評価用ドキュメントの類似度により推薦するドキュメントを決定し、その推薦精度によって評価を行なう。ユーザプロファイル **u** と各ドキュメント $\mathbf{d}_i$ の類似度を、

$$Sim = \mathbf{u}^t \mathbf{d}_i \tag{2.6}$$

で定義する。また、評価方法としては、0 から 1 まで 0.1 刻みで再現率を変化させたときの各再現率における補間適合率を用いた再現率・適合率曲線と、それら 11 点の補間適合率の平均（11 点平均適合率 [41]）を用いる。

適合率は式 (2.7)、再現率は式 (2.8) で計算される。

$$\text{適合率} = \frac{\text{検索された文書中の該当文書数}}{\text{検索された文書数}} \tag{2.7}$$

$$\text{再現率} = \frac{\text{検索された文書中の該当文書数}}{\text{全文書中の該当文書数}} \tag{2.8}$$

補間適合率とは、任意の再現率における適合率のことであり、再現率 r における補間適合率 $P(r)$ は、ユーザプロファイルとの類似度の上位 i 番目までの文書を用いて算出した再現率 R_i と適合率 P_i を用いて、次式のように表わす。

$$P(r) = \max_{r \le R_i} P_i \tag{2.9}$$

2.4　独立成分分析

この節では、ドキュメントに独立成分分析を適用する手法について説明する。

2.4.1　文書-単語モデルに対する独立成分分析

独立成分分析（ICA:Independent Component Analysis）は複数の観測された混合信号を統計的に独立な信号に分離する手法である。ICA をドキュメントに適用する場合、混合信号はドキュメントベクトルとし、独立な信号はそのドキュメント中に含まれるトピックとする。ここで、トピックとはあるカテゴリにおける索引語の重みを表現したものである。ドキュメントに対して ICA を適用する場合、それらのトピックが組み合わされて各ドキュメントが構成されるとする。このとき、各トピックはそれぞれ異なった話題を対象としているため、それぞれのトピックは独立であると考えることができる。

いま、n 個のドキュメントベクトルを $\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_n$ とし、このドキュメントベクトルが未知の m 個のトピックのベクトル $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_m$ の線形結合で与えられるとする。ドキュメントベクトル数よりトピック数が多いと、解が一意に定まらないので、$n \geq m$ とする。このとき、ドキュメントベクトルとトピックベクトルを並べた行列は、式（2.10）のようになる。

$$\begin{aligned} D &= [\mathbf{d}_1 \quad \mathbf{d}_2 \quad \cdots \quad \mathbf{d}_n]^t = [\mathbf{t}_{d1} \quad \mathbf{t}_{d2} \quad \cdots \quad \mathbf{t}_{dV}] \\ S &= [\mathbf{s}_1 \quad \mathbf{s}_2 \quad \cdots \quad \mathbf{s}_m]^t = [\mathbf{t}_{s1} \quad \mathbf{t}_{s2} \quad \cdots \quad \mathbf{t}_{sV}] \end{aligned} \tag{2.10}$$

また、D を文書-単語行列、S をトピック行列と呼ぶ。ここで、$\mathbf{t}_{dj}$ は次式のように表わされる。

$$\mathbf{t}_{dj} = [w_{1j} \quad w_{2j} \quad \cdots \quad w_{nj}]^t \tag{2.11}$$

$\mathbf{t}_{dj}$ は各文書における索引語の重みをベクトルとして表現したものである。$\mathbf{t}_{sj}$ も同様に定義する。

ここで、$\mathbf{t}_{dj}$ の各単語の重みは、各トピックに含まれる単語の重み $\mathbf{t}_{sj}$ の混合であると考える。つまり、$\mathbf{t}_{dj}$ を式 (2.12) のモデルで表現する。

$$\mathbf{t}_{dj} = A\mathbf{t}_{sj} \tag{2.12}$$

A は未知の $n \times m$ 混合行列である。

式 (2.12) は、各単語の重みは独立であるという仮定から、すべての $\mathbf{t}_{dj}$ で成立するので、行列を用いて式（2.13）のように記述できる。

$$D = AS \tag{2.13}$$

ICA は、トピック行列 S や混合行列 A が未知の場合に、文書-単語行列 D のみを用いて、トピック行列 S を推定する手法である。実際は、文書-単語行列 D のみから、復元行列 W を用いて、互いに統計的独立となる復元信号 Y を求める。

$$Y = [\mathbf{y}_1 \quad \mathbf{y}_2 \quad \cdots \quad \mathbf{y}_n]^t = WD \tag{2.14}$$

ただし、ICA の性質から Y の成分の大きさと順序には任意性が残る [13]。

復元行列 W を推定する手法として、Fast ICA[42] が提案されている。本論文では Fast ICA を用いて、独立成分を求める。次節でその Fast ICA について述べる。

ICA によって求められた Y の成分は、互いに統計的独立である。したがって、Y が入力ドキュメント中に含まれるトピックの推定値と考えられる。

文書-単語行列 D を、トピックで構成される空間へ射影する。ここでは、式（2.15）のようにトピックで表現されたドキュメントベクトル行列 $\hat{D}$ を用いて、ユーザプロファイルを作成する。

$$\hat{D} = Y^t D \tag{2.15}$$

2.4.2 Fast ICA

本研究では、Fast ICA アルゴリズムを用いて、独立成分を求める。独立成分を求める際に Fast ICA で利用される更新式として、式（2.16）を用いる。なお、$g(\nu)$ は 非ガウス性を測るための関数であり、$\tanh(\nu)$ や、$\nu e^{-\nu^2/2}$ などが用いられる。

$$\mathbf{w}' = E[Yg(\mathbf{w}^t Y)] - E[g'(\mathbf{w}^t Y)]\mathbf{w} \tag{2.16}$$

$$g'(\nu) \triangleq \frac{\partial g(\nu)}{\partial \nu} = \begin{bmatrix} \frac{\partial g(\nu_1)}{\partial \nu_1} & \cdots & \frac{\partial g(\nu_1)}{\partial \nu_V} \\ \vdots & \ddots & \vdots \\ \frac{\partial g(\nu_V)}{\partial \nu_1} & \cdots & \frac{\partial g(\nu_V)}{\partial \nu_V} \end{bmatrix}$$

ここで、$E[\cdot]$ は期待値を表わす。

以下に、Fast ICA アルゴリズムの手順について示す。

1. 入力信号 D の平均値を 0 にする。

$$D' = D - E[D] \tag{2.17}$$

2. PCA によってホワイトニングを行なう。なお、Rot は回転行列を示す。

$$E[D'^t D'] = R^t \Sigma R$$

$$Rot = R^t \Sigma^{\frac{1}{2}} R \tag{2.18}$$

$$D'' = RotD' \tag{2.19}$$

   ここで、$\Sigma^{1/2} = diag[\lambda_1^{1/2}, \lambda_2^{1/2}, \cdots, \lambda_V^{1/2}]$ とする。

3. $\mathbf{w}$ の初期値をランダムに選び、正規化をする。
4. 更新式（2.16）によって、$\mathbf{w}$ を更新する なお、$g(\nu) = \nu e^{-\frac{\nu^2}{2}}$ を用いる。
5. $\mathbf{w}'$ を正規化する。
6. $\mathbf{w}'$ と $\mathbf{w}$ の内積が 1 に近ければ、計算を終了し、$\mathbf{w}_k = \mathbf{w}'$ とする。そうでなければ、4 に戻って $\mathbf{w}'$ を再度更新する。

7. $k = 1$ であれば 3 に戻り次の重みベクトルを求める。$k \geq 2$ であれば、$\mathbf{w}_k$ と $\mathbf{w}_1$ から $\mathbf{w}_{k-1}$ とを、Schmidt の直交化法により直交化する。
8. $k = m$ となるまで 3 以下を繰り返す。
9. m 個の重みベクトル $\mathbf{w}_k$ から、$W = [\mathbf{w}_1 ... \mathbf{w}_m]$ を作成し次の式により Y を求める。

$$Y = WD'' \tag{2.20}$$

2.5　実験と結果

この節では実験とその結果について述べる。

2.5.1　実験環境

実験では、テストコレクション NTCIR2 から、タスク No.110（情報検索の可視化）に含まれる 625 件のドキュメントを用いた。なお、本実験では、キーワード検索等で必要と考えられるドキュメントを絞り込んだ後に、さらにその絞り込まれた中から必要なドキュメントを抽出するという状況を想起している。したがって、本実験で用いるサンプル数は少なくても良い。これらのドキュメントにはあらかじめタスクとの関連度合いとして、S、A、B、C の 4 段階にランク付けされている。S は一番タスクとの関連度合いが強く、A、B と順に関連度合いが下がり、C は関連がないドキュメントを示す。本節では、C が付いているドキュメントを興味のないドキュメントとし、残りのドキュメントを興味のあるドキュメントとした。その結果、625 件のうち、34 件のドキュメントを興味ありドキュメントと判定した。また、それら 625 件のドキュメントを 125 件ずつの 5 つのデータセットに分割し、5 つのうち 3 つを学習データ、残り 2 つを評価データとして交叉検定を行なった。すべての組み合わせについて実験を行なうと、10 通りの実験が行なえる。表 2.1 に 5 つに分割したデータの詳細を示す。

表 2.1　実験データ

	all	set1	set2	set3	set4	set5
interesting	34	7	13	5	3	6
uninteresting	591	118	112	120	122	119

これらのドキュメントに対して、まず、茶筌 [43] を用いて形態素解析を行ない、名詞を抽出した。その後、“こと” や “もの” といったストップワードを削除した。以上の処理により得られた索引語数は 1,145 個となり、ドキュメントベクトルは 1,145 次元となった。そして、それらを用いてユーザプロファイルを作成した。

各実験はもとのドキュメントベクトルから遺伝的アルゴリズムを用いてユーザプロファイルを作成する方法（GA)、もとのドキュメントベクトルから、それらのドキュメントベクトルの重心を用いてユーザプロファイルを作成する方法（FB)、本章で提案した ICA によって得られるトピックを用いて表現したドキュメントから遺伝的アルゴリズムによってユーザプロファイルを作成する方法（ICA+GA)、トピックによって表現されたドキュメントから重心を用いてユーザプロファイルを作成する方法（ICA+FB）の 4 種類の手法で、ユーザプロファイルを作成した。

結果の評価には再現率・適合率曲線と 11 点平均適合率を用いた。なお、各再現率における適合率は補間適合率を用いた。以下に実験の流れを示す。

1. ドキュメントベクトルを作成する。
2. ICA を実行する。
3. ユーザプロファイルを作成する。
4. ユーザプロファイルを用いて評価用データを推薦し、再現率・適合率曲線、11 点平均適合率により評価を行なう。

2.5.2　GA によるユーザプロファイルの作成

ここでは、遺伝的アルゴリズムによるユーザプロファイルの作成について述べる。GA で用いる適合度関数は式（2.5）とする。また、係数 α、β は、興味のあるドキュメントと興味のないドキュメントの比率が 1:1 になるように、以下のように定める。

$$\begin{aligned} \alpha &= +1 \\ \beta &= -N_I/N_U \end{aligned} \tag{2.21}$$

交叉は 2 点交叉を用いた。他のパラメータは表 2.2 に示す。

表 2.2　GA のパラメータ

	世代数	交叉率	突然変異率
GA	10000	1	0.005
ICA+GA	5000	1	0.05

再現率・適合率曲線を、図.2.1 に示す。つぎに、11 点平均適合率を表 2.3 に示す。

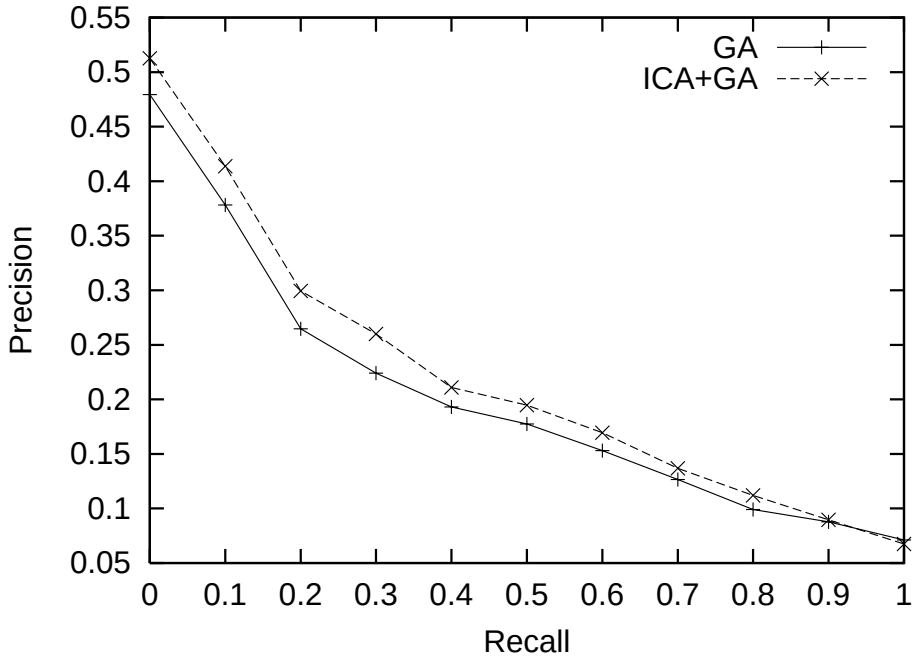

図 2.1　再現率・適合率曲線

表 2.3　11 点平均適合率の比較

GA	0.20
ICA+GA	0.22

2.5.3　重心によるユーザプロファイルの作成

ここでは、重心によるユーザプロファイルについて述べる。式（2.4）中の係数 α と β は、式（2.21）と同様に定める。

適合率・再現率曲線の結果を図 2.2 に示す。11 点平均適合率は表 2.4 に示す。

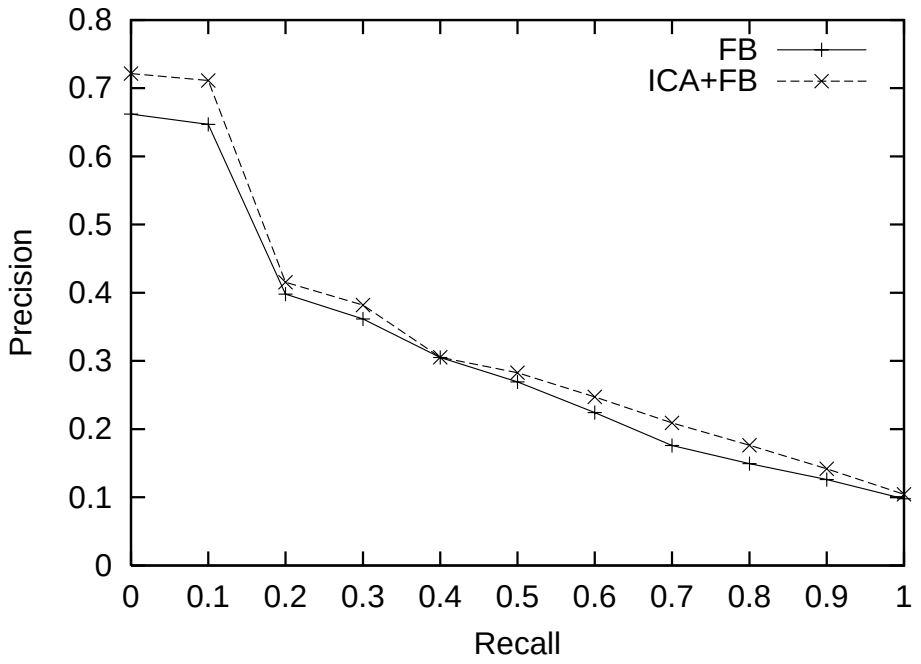

図 2.2　再現率・適合率曲線

表 2.4　11 点平均適合率の比較

FB	0.31
ICA+FB	0.34

2.5.4　独立成分の例

ここでは、表 2.5 と表 2.6 に、得られた独立成分の例について示す。それぞれ、各索引語に対する重みで並べ替え、その重みの大きな索引語を示している。表 2.5 は、トピックを表わした独立成分の一例である。一方、表 2.6 は、トピックとして意味を抽出できない独立成分の例である。

表 2.5　独立成分の例

形態素解析	16.90
文字列検索	0.44
形態素	0.26
メディア	0.09
関連フィードバック	0.09
マルチ	0.08
文章構成	0.08
文書検索システム	0.07

表 2.6　ノイズ成分

総合得点	5.94
ベクトル	5.86
動的イメージ	5.31
斜め方向	5.03
ビット	4.66
地方都市	4.57
ファイル	4.28
テレビ	4.25

2.6　検討と考察

図 2.1 と図 2.2 を見ると、ICA は重心を用いた手法でも GA を用いた手法でもともに、すべての再現率において適合率がわずかながら改善していることが分かる。また、表 2.3、表 2.4 に示した 11 点平均適合率を見ても ICA がわずかながら適合率を改善させていることが分かる。ICA は前処理として適用ができるので、実運用を考えた際には、ユーザプロファイルの作成のみが処理速度として求められる。ユーザプロファイルの作成を ICA のトピック空間上で行なうことにより、GA、学習ドキュメントの重心ともにドキュメントベクトルの次元を削減することができ、その結果、精度がわずかながらも改善するとともに、処理速度も向上するので、ICA によって得られるトピック空間上でユーザプロファイルを作成する本手法は有効であると考えられる。

つぎに、得られた独立成分に関する検討を行なう。本実験のデータは、あるテーマに対して検索を行なった結果得られたドキュメントを集めたものである。したがって、そのテーマに関連するトピックが ICA を用いて得られると考えられる。しかし、表 2.6 に示した例で

は、お互いにあまり関連性の低い単語が同じような重みを持っているので、そのような特定のトピックを汲み取ることが困難である。したがって、この成分はノイズであると考えられる。本章では、すべての独立成分を使ったために、ノイズの除去は行なっていない。それゆえ、それらのノイズを除去できれば、精度は改善すると考えられる。さらに、独立成分がドキュメント中に含まれるトピックを表わすことを確認した。表 2.5 より、この成分は、“文字列検索”や “関連フィードバック”など、情報検索に関するいくつかの単語に大きな重みが付いている。したがって、この成分は、情報検索に関するトピックであると考えることができる。

2.7 まとめ

本章では、ICA をユーザプロファイル作成の前処理として用いた。その結果、ICA を適用することによって、ユーザプロファイル作成において、その作成手法によらず本手法が有効であることを確認した。さらに、不要なトピックを除去することにより、今後さらなる精度改善が期待できると考えられる。最後に、独立成分がトピックを示していることも確認した。

第 3 章

独立成分の選択による情報フィルタリング

3.1　はじめに

2 章で述べたように、ICA によって得られるトピックを用いてドキュメントを再表現し、トピック空間上で情報フィルタリングを行なうことは有効である。得られたトピックにはノイズと考えられる成分も含まれており、情報フィルタリングの精度を改善しようとする際には、それらを除去することが必要となる。しかしながら、LSA にみられるような小さな特異値に対応する特異ベクトルを削除するというようなノイズを除去するために利用可能な基準がいまだ提案されていない。

ところで、ドキュメントベクトルに対しては、クラスタリングして整理を行なう手法 [44] が提案されている。トピックもドキュメントベクトルと同様の形状をしているため、クラスタリングを行なうことで、トピックを整理することが可能であると考えられる。さらに、トピックの整理を行なった後、必要なトピックを選択することで、ノイズを除去することができると思われる。

そこで、本章では、トピックの意味的な類似性に着目してトピックの選択を行ない、情報フィルタリング精度を改善する手法を提案する。トピックを選択するために、まず、トピックのクラスタリングを行ない、トピックの選択を行なう。従来のクラスタリング手法ではユークリッド距離が各データ間の距離として用いられる。しかし、Fast ICA によって得られたトピックは、スケールと順序に任意性が存在するため、ユークリッド距離ではトピック間の類似性を正確に評価できず、クラスタリングを行なうことができないという問題点がある。本章では、情報幾何の分野で注目されている分布間距離のひとつである、Jensen-Shannon 情報量（JS 情報量）[45][46] を導入した最大距離アルゴリズム（MDA:Maximum Distance Algorithm）[47][48] を提案し、クラスタリングを行なう。さらに、選択されたトピックを用いてドキュメントベクトルを再表現し、ユーザプロファイルを作成する。なお、ユーザプロファイルの作成には学習ドキュメントの重心を用いる。これらの手法の有用性を確認するために、テストコレクション NTCIR2 を用いた評価実験を行なう。また、MDA によってクラスタリングされた独立成分について検討する。

3.2　提案手法

本手法は、ICA をドキュメントに適用して得られたトピックに対して、MDA を用いてクラスタリングを行なった後、必要なトピックを選択する。さらに、選択したトピックを利用してドキュメントベクトルを再表現することにより、ユーザプロファイルの精度向上を図る。なお、トピックは 2.4 節で説明した手法によって求める。以下ではトピックを得た後の MDA の処理について説明する。

ICA で得られたトピック中からトピックを選択することは、情報フィルタリングの精度改善に効果があると考えられる。本章では、クラスタ数をあらかじめ指定する必要のないクラスタリング手法である MDA を用いて、トピックをクラスタリングし、選択を行なう。

このアルゴリズムは既存のクラスタ中心から一番遠い点を探し出し、その距離が十分遠い場合に、その点を中心とした新しいクラスタを作成する手法である。したがって、ICA で求められたトピック Y の中で類似なトピックをクラスタとしてまとめ、異なった性質のトピックを別のクラスタに分類することができる。

MDA で用いられる距離関数としては、ユークリッド距離がしばしば用いられる。しかし、Fast ICA によって得られたトピックはスケールと順序に任意性が存在するので、ユークリッド距離ではトピック間の類似性を正確に評価できない。したがって、他の距離関数を用いる。情報幾何学においては、Kullback Leibler 情報量（KL 情報量）[49] などが距離関数として良く知られているが、ここでは、KL 情報量に対称性を持たせるよう拡張した JS 情報量を用いて、トピックをクラスタリングする。JS 情報量を用いるためには、トピックを確率として表現する必要がある。コルモゴロフの公理 [50] によれば、各事象の確率は 0〜1 の範囲にあり、その総和は 1 である。この条件を満たすために、以下の手順によりトピックを変換する。

1. $m = \min_{i,k} y_i(k)$ を求める。なお、$y_i(k)$ は $\mathbf{y}_i$ の k 成分を示す。
2. $y'_i(k) = y_i(k) + m$ によって、トピックのすべて要素を正の値へシフトする。
3. $\sum_k^V y'_i(k) = 1$ になるように変換する。

本手法では、ベクトル空間においてクラスタリングを行なう。ただし、JS 情報量を用いてトピック間の距離を計算するときに限り、先に述べた処理により、トピックを確率に変換し距離を求める。以下に MDA の処理を示す。

1. 最も離れたクラスタ間の距離に対する比率 r を設定する。
2. トピック $\mathbf{y'_1}$ をクラスタ中心 $\bar{\mathbf{Z}}_\mathbf{1}$ とする。
3. $\mathbf{y}'_2, \cdots, \mathbf{y}'_n$ に対して、$DIST_i = \min_j dist(\mathbf{y'_i}, \bar{\mathbf{Z}}_\mathbf{j})$ を求める。$dist(\mathbf{y'_i}, \bar{\mathbf{Z}}_\mathbf{j})$ は式（3.1）で与

えられる JS 情報量を示す。

$$dist(\mathbf{y'_i}, \mathbf{\bar{Z}_j}) = \sum_{k}^{V} H[\frac{1}{2}\{y'_i(k) + \bar{Z}_j(k)\}] - \sum_{k}^{V} \frac{1}{2}\{H[y'_i(k)] + H[\bar{Z}_j(k)]\} \quad (3.1)$$

ただし、$H[x]$ は式 (3.2) で定義する。

$$H[x] = -x \log x \quad (3.2)$$

4. $l = \max_i DIST_i$ を求める。l となる成分を $\mathbf{y}_k$ とする。
5. $l/MAX > r$ ならば、$\mathbf{y_k}$ をクラスタ中心 $\mathbf{\bar{Z}_{j+1}}$ とする新しいクラスタを作成し、(3）へ戻る。ここで、$MAX = \max_{i,j} dist(\mathbf{\bar{Z}_i}, \mathbf{\bar{Z}_j})$ は、最も離れたクラスタ中心間の距離である。
6. 各成分の属するクラスを決定する。

不必要なトピックを除去するために、クラスタに含まれる成分数が多いクラスタから順に、その全要素を必要成分として選択する。　同一クラスタ内に含まれる要素は、ある特定のトピックに関して細分化された成分であると考えられる。それゆえ、この選択手法は、主要なトピックを詳細に表現する成分を選択するものである。本章で扱うドキュメントは単一のタスクテーマに関連するものであるので、上述の選択方法により、小さなクラスタに分類されるトピックが、メインとなるタスクテーマと関連が薄いと考えられる。

3.3　実験と結果

MDA を用いたトピック選択によるフィルタリング精度改善への有効性を確認するために、テストコレクションを用いた評価実験を行なう。以下では、その実験について述べる。

3.3.1 実験環境とその方法

本章では、前章の実験と同様、テストコレクション NTCIR2 からタスク No.110（情報検索の可視化）に含まれる 625 件のドキュメントを用いた。

本章で用いる索引語は、前章で用いた索引語に加えて、漢字のみで構成される単語が隣接する場合、それらを結合して新たに一つの索引語とした。その結果、得られた索引語数は 5,548 個となり、ドキュメントベクトルの次元数は 5,548 次元となった。これらの選択された単語を用いて、tf-idf 法により重み付けを行ない、625 件のドキュメントに対して、ドキュメントベクトルの作成を行なった。

作成された 625 本のドキュメントベクトルに ICA を適用し、623 本の独立成分を得た。なお、ICA で独立成分を得る際に用いる式（2.16）における Fast ICA アルゴリズム内の非ガウス性を測るための関数には、$\tanh(v)$ を用いた。つぎに、その 623 本の独立成分を MDA を用いてクラスタリングし、フィルタリングに不必要と考えられる成分を除去した。この際、選択成分数が 300 となるように、MDA のパラメータ r は 0.47 とした。また、MDA によって選択した独立成分を用いて、入力ドキュメントを変換した。変換後のドキュメントベクトルの次元数は、300 次元になる。選択成分数の 300 という値は、選択成分数を 200 と 400 として同様の実験を行ない、その結果を踏襲した上で決定した。

ユーザプロファイルの作成には、式 (2.4) の学習ドキュメントの重心を利用した。なお、式 (2.4) 中の係数 α と β は、興味のある記事と興味のない記事の比率が 1 : 1 になるように式（2.21）と同じ値を用いた。

また、ユーザプロファイル作成の際には交叉検定を行なった。625 件のドキュメントを 125 件ずつ 5 つのサブセットに分け、そのうち 3 つのサブセットを組み合わせて、ユーザプロファイルを作成するための学習データ、残りの 2 つのサブセットを評価データとし、10 パターンのデータに対して実験を行なった。各サブセットに含まれる興味のあるドキュメン

トと興味のないドキュメントの数を表 3.1 に示す。

最後に、作成したユーザプロファイルを用いて、評価用ドキュメントを推薦し、再現率・適合率曲線と 11 点平均適合率を用いてその推薦精度を評価した。

表 3.1　実験データ

	All	Set1	Set2	Set3	Set4	Set5
Interesting(No.)	34	7	13	5	3	6
Uninteresting(No.)	591	118	112	120	122	119

以下に、実験の流れをまとめる。

1. ベクトル空間法によりドキュメントをベクトル表現する。
2. ICA を適用する。
3. 求められた独立成分を MDA によりクラスタリングし、不要な成分を除去する。
4. 求められた成分を用いて、入力ドキュメントベクトルを変換する。
5. 学習ドキュメントの重心を用いて、ユーザプロファイルを作成する。
6. ドキュメントの推薦を行ない、11 点平均適合率を用いて評価を行なう。

実験は、ICA を適用せず元のデータを用いてユーザプロファイルを作成した場合 (Original)、ICA を適用した場合 (ICA)、および、ICA 適用後に MDA で独立成分の選択を行なった場合 (MDA) の 3 つに対して行ない、トピック選択のフィルタリング精度への影響を検討した。また、本章で提案している MDA を利用したトピック選択手法の有効性を確認するために、Kurtosis の高いトピック 300 本を必要トピックとして選択する実験 (Kurtosis) を行なった。Kurtosis は、非ガウス性を測る尺度で、ICA において独立性の基準として用いられている。なお、各トピック $\mathbf{y}_i$ の Kurtosis は、次式で計算される。

$$kurt(\mathbf{y}_i) = E[\mathbf{y}_i^4] - 3(E[\mathbf{y}_i^2])^2 \tag{3.3}$$

3.3.2 実験結果

図 3.1 は、各再現率における適合率を表わしている。なお、交叉検定によるトレーニングセット間の標準偏差をエラーバーとして採用している。Kurtosis によるトピック選択を適用した場合と、MDA によるトピック選択の場合を比較するために、2 手法の各再現率における適合率を図 3.2 に示す。11 点平均適合率を表 3.2 に示す。MDA の選択成分数を変更した場合の 11 点平均適合率の比較を表 3.3 に示す。また、MDA によりクラスタリングした成分の例を表 3.4、表 3.5 に示す。各成分の重みによって並べ替え、重みの大きい一部を例として示している。表 3.4 は、クラスタ中心になっているトピックの一例である。表 3.5 は、表 3.4 中の Cluster1 に含まれる成分の一例である。これらの成分はクラスタ中心から距離の近い順に選択している。

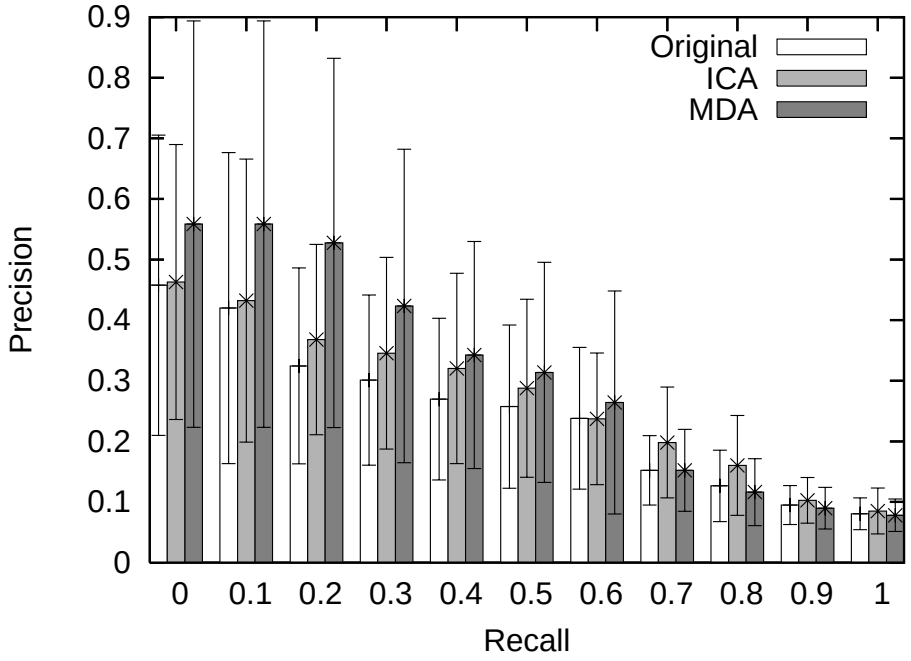

図 3.1 各再現率における適合率

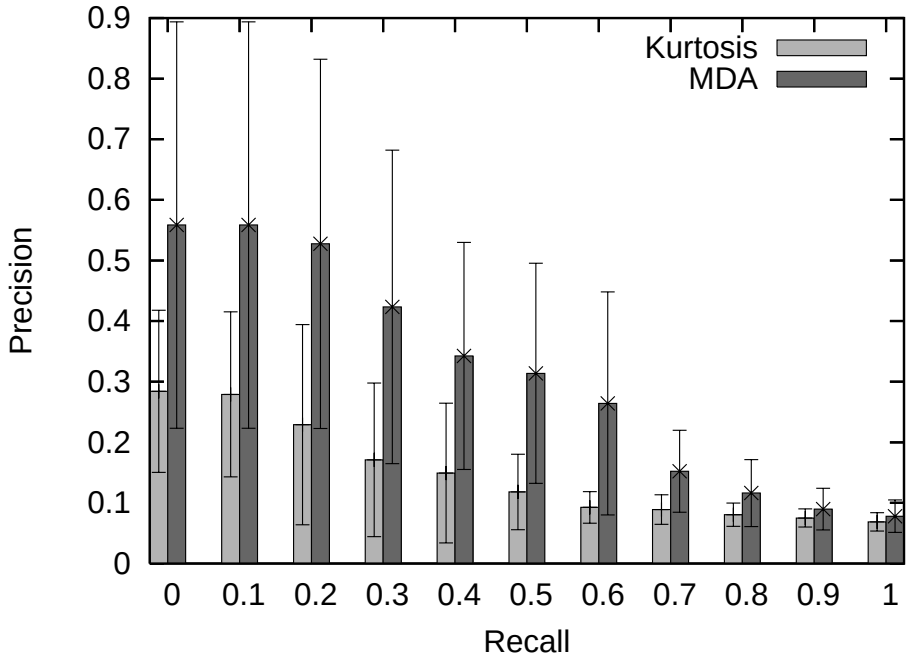

図 3.2　Kurotsis による各再現率における適合率

表 3.2　11 点平均適合率の比較

	Original	ICA	MDA	Kurtosis
Average	0.248	0.273	**0.311**	0.149

表 3.3　選択成分数による比較

Compnent #	200	300	400
Average	0.183	**0.311**	0.207

3.4　検討と考察

まず、図 3.1 を見ると、交叉検定の学習セットによって結果のばらつきがあるものの、第 2 章で述べたように、もとのデータから作成したユーザプロファイルより、ICA を適用した

表 3.4　クラスタ中心の例

Center1	Center2	Center3
行動選択確率	閲覧支援機能	流体解析
統合アルゴリズム	検索対象言語	特徴マップ
不完全知覚問題	対訳生成言語	対象画像
次元の呪い問題	選択例	顔部分空間
性能比較	主要キーワード	並列計算サーバー
ベンチマーク	閲覧支援情報	画像データ生成
セマンティック	コーパス	画像データ転送
セマンティックボックス	閲覧支援	画像圧縮技術
多面体	言語情報検索手法	実時間可視化
オブジェクト情報	文書検索技術	並列計算機
表現方法	帰納的学習	領域分割法
ユーザインタフェース	適応型	画像データ圧縮
位置関係	テキストデータ	数値流体力学
仮想空間	検索要求記述	Java アプレット
多義情報	相対位置関係	臨界投機
情報視覚化	オブジェクト多次元ベクトル行列	インストール
音程情報	検索手法	並列処理
ルール	キーワード対訳表示	写真画像
旋律情報	英日交換	焼き付け
デュアル	多言語情報検索	顔画像の抽出

表 3.5　クラスタ 1 に含まれる要素の例

Component1	Component2	Component3
チャット	マニュアル	フィルター
コミュニケーション	関連項目	定型フィルター
複数ユーザ	マニュアルシステム提案	集合フィルター
チャット機能	マニュアル本文	ドロップ
情報探索タスク	次元マップ	ドラッグ
検索コマンド	クリック	カレンダー
チャット参加	マウス	表示位置
チャット自体	タイトル	情報管理
コマンド	次元空間	情報視覚化
タスク	電子マニュアル	情報視覚
繊維デザイン画像データベース	ワープロ	ファイル
遠隔実験	目次構造	ユーザテスト
まとめ文章	検索機能	ブックマーク
遠隔地	キーワード検索	位相学習
データベースシステム	リファレンスマニュアル	キーワード検索
ターゲット	目次検索	テスト
サイバースペース	ワープロユーザ	設計情報
文字数	分類提示	移動ロボット
キーワード検索	目次	問題構造
インド	索引検索	発見支援

場合のユーザプロファイルの方が適合率はよい結果となっている。これは、ICA を適用することによりトピックが明確になり、適合率の改善が図られたためと思われる。また、MDA を用いた場合は、変換前のデータを利用してユーザプロファイルを作成した場合よりも適合率は改善されている。さらに、ICA を適用した場合と比べると、再現率の低い場所において、とくに適合率が改善されている。

また、表 3.2 を見ると、提案手法である MDA による独立成分を選択する手法の 11 点平均適合率が他の手法に比べて向上していることが分かる。これは、MDA を利用した場合に、ドキュメント推薦に不必要な成分を取り除くことができ、ノイズ低減が図られたためと思われる。

さらに、図 3.2 と表 3.2 から、Kurtosis の高い成分を選択する手法と、本章で提案している、MDA によりトピックを選択する手法を比較すると、本提案手法の方が良い結果を残している。これにより、意味的な特徴から成分選択をする手法が有効であると考えられる。

本章では表 3.3 に示すように、選択する成分数を変えた実験をいくつか行なった上で、独立成分数を 300 に決定した。選択する独立成分数が多すぎるとノイズが混じり、少なすぎるとドキュメントの表現力が低下する。したがって、推薦精度改善のためには最適な成分数を決定する必要がある。この最適な成分数はドキュメント集合によっても変化するものと考えられる。

つぎに、MDA によりクラスタリングされたトピックについて考察をする。クラスタ中心に着目すると、表 3.4 より、各クラスタ中心は、異なったトピックを代表しており、クラスタ分類が行なていることが分かる。表 3.4 に示した例からでも、Cluster1 は、自然言語処理に関連するトピック、Cluster2 は、キーワード検索に関するトピック、Cluster3 は、画像認識に関するトピックを表現していると推定できる。

表 3.5 は表 3.4 に示されている Cluster1 に含まれる成分の一例を示している。まず、Component1 と Component2 は、それぞれ “チャット” や “マニュアル” に関連した単語を多

く含んでいることから、自然言語処理の応用事例を表わしていると考えられる。実際、これらの単語が含まれているもとのドキュメントは、チャットから言葉の意味を表わす概念を自動獲得したり、電子マニュアルから重要語を自動抽出するといったような内容であった。一方、Component3 は、Component1 や Component2 とは違い、“フィルタ” などの単語から、フィルタリングに関する成分であると考えられる。とくに、“情報視覚化” などの単語がいくつか含まれており、情報の可視化に大きなウェイトのある成分である。これは、代表ベクトルに含まれる “情報視覚化”などの単語により分類されたものと考えられる。このように、クラスタに含まれる成分は代表ベクトルが表現するトピックを細分化したものであると考えることができる。

以上のことから、MDA によって類似トピックをクラスタリングでき、それらのトピックを選択することにより、効果的な成分選択ができると考えられる。

3.5　まとめ

本章では、ICA によって得られるトピックが構成する空間上におけるユーザプロファイル用いた情報フィルタリングを実現するにあたり、ICA によって得られた成分には検索に不必要な成分が含まれていることに着目し、空間を構成するトピックを選択することで、そのフィルタリング精度を改善する手法を提案した。とくに、そのトピックの選択にあたっては、トピックが持つ類似性に着目し、JS 情報量による MDA を用いてトピックをクラスタリングし、不必要な成分を除去した。また、MDA により、それらの ICA で得られる独立成分（トピック）がクラスタリングされることも確認した。

ここでは対象ドキュメントを情報検索に限定して実験を行なったので、今後は対象ドキュメントを変えて実験を行なうなど、本手法の更なる検討が必要である。

第 4 章

特異値分解と独立成分分析による潜在的意味を用いた情報フィルタリング

4.1　はじめに

ドキュメントは計算機上ではドキュメント中に含まれる索引語に対する重みを要素とするドキュメントベクトルで表現され、ドキュメント中に含まれる索引語数と同じ次元数を有している。このようなドキュメントベクトルに対して ICA を適用し、独立成分を探索すると計算時間がかかる。この原因としては、1) 上記手法では計算時間がドキュメントベクトルの次元数に比例し、索引語数の増加がその次元数を大きくすること、および、2) 探索すべき独立成分数が増加することが考えられる。

本章では、LSA で用いられている特異値分解（SVD: Singular Value Decomposition）[41] によりドキュメントベクトルを低次元化し、ICA の処理時間を短縮することを目的とする。また、得られた潜在的意味空間上で遺伝的アルゴリズム（GA: Genetic Algorithm）を用いて

ユーザプロファイル（ユーザの興味）を作成し、低次元化による推薦精度への影響について検討する。低次元化を行なう際には、寄与率の低い成分を除去する。このとき、累積寄与率の変化による推薦精度の推移についても検討する。さらに、得られた潜在的意味について考察を行なう。

4.2　従来手法

LSA によって得られる特異ベクトルは、ドキュメント中の共起する単語の関係を表わしており、LSA では、それらの潜在的意味空間上へドキュメントベクトルを写像している。

ICA と SVD を組み合わせてトピックを取得する研究としては、文献 [15] や文献 [18] がある。これらの研究では、SVD により低次元化を行なった後、ICA を適用して得られる混合行列がトピックを表わしていると報告されている。また、上記の列ベクトルは独立性という特徴は持たない。本章では、上述の混合行列を用いる手法ではなく、トピックの独立性という観点に基づき、文献 [14] に述べられている ICA の独立成分をトピックとみなす手法に対して、SVD を組み合わせ潜在的意味を抽出する手法を提案する。また、その得られた潜在的意味を情報フィルタリングに応用することを考える。

4.3　提案手法

本章では、SVD と ICA を組み合わせ、ドキュメントから潜在的意味を抽出し、ユーザプロファイルを作成する手法を提案する。ICA のみを用いた場合、ドキュメントベクトルの次元数が大きく、ICA の処理に多大きな時間を要する。これを回避するために、ICA を実行する前に SVD を用いて、ドキュメントベクトルの低次元化を図る。つぎに、潜在的意味空間上でユーザプロファイルを作成し、低次元化による推薦精度への影響を検討する。

以下に、SVD と ICA を組み合わせて潜在的意味を抽出し、それらを用いてドキュメント

を表現する手法を説明する。

まず、2.4.1 で説明を行なった、文書-単語モデルに対する ICA に焦点を当てる。ICA では、データの次元数が大きいと、探索空間が大きくなり過ぎ、独立な成分を探索するのに時間がかかる。一方、ベクトルの次元数に比べ、ドキュメント数が少ないという点から、より低次元の空間でドキュメントは表現することができると考えられる。そこで、SVD により低次元化し、ICA の処理時間を短縮させる。SVD によって得られた分散の小さい軸を削除した空間に、ドキュメントベクトルを写像して低次元化を実現する。SVD は、式（4.1）のように D を分解する。

$$D = U\Sigma V^t \tag{4.1}$$

ここで、U、V の列ベクトルはそれぞれ左特異ベクトル、右特異ベクトルと呼ばれ、D の直交基底になっている。また、Σ は式（4.2）に示す特異値 σ_i を対角成分にもつ対角行列である。

$$\Sigma = diag[\sigma_1, \sigma_2, \cdots, \sigma_p] \tag{4.2}$$

p は D のランク数を示し、$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$ を満す。本手法では、D を SVD によって得られた V の張る空間へ写像することにより、D の低次元化を行なう。その際、V に含まれる特異ベクトルのうち、対応する特異値の小さな特異ベクトルを除去することは、ノイズ低減を意味する。式（4.3）に特異ベクトルを選ぶ際の基準となる累積寄与率 r_k を定義する。

$$r_k = \sum_i^k \sigma_i^2 / \sum_i^p \sigma_i^2 \tag{4.3}$$

いま、V の特異ベクトルの中で、式（4.3）の累積寄与率 r_k がある割合 γ 以下となる条件で、σ_1 から順に、最大 k 個のそれぞれの特異値に対応する特異ベクトルを選択し、それらからなる行列を V_k とする。そのとき、D の変換は式（4.4） のようになる。

$$D_k = DV_k \tag{4.4}$$

求められた D_r を用いて ICA を適用し、独立成分 S を求める。なお、独立成分は Fast ICA アルゴリズムを用いて求める。その際、用いる式（2.16）の非ガウス性のコスト関数は tanh(ν)

とした。この S は右特異ベクトルの張る空間上の特徴であり、索引語を用いて潜在的意味として解釈するためには、もとの空間へ戻す必要がある。そこで、S を式（4.5）を用いてもとの単語による空間へ戻し、それを潜在的意味 T と呼ぶ。

$$T = S V_k^t \tag{4.5}$$

T を用いて、D を式（4.6）のように変換する。

$$\hat{D} = T^t D \tag{4.6}$$

上記の $\hat{D}$ を用いてユーザプロファイルを作成する。

以下に提案手法の手順をまとめる。

1. 特異値分解 $D = U\Sigma V^t$ の計算をする。
2. 累積寄与率 r_k に基づいて V_k を決定する。
3. $D_k = DV_k$ と変換する。
4. Fast ICA により $D_k = AS$ を満足する S を決定する。
5. 潜在的意味 $T = S V_k^t$ を抽出する。
6. D を $\hat{D} = T^t D$ と変換する。
7. $\hat{D}$ よりユーザプロファイルを作成する。

4.4 実験と結果

本節では、提案手法によって得られる潜在的意味を用いて作成したユーザプロファイルの有効性を確認するために行なった評価実験について説明する。

4.4.1 実験方法

実験データとして日本語のテストコレクション NTCIR2 から、タスク No.109（TCP の高速化）、タスク No.110（情報検索の可視化）、タスク No.121（ネットワークを用いた VoD システム）、タスク No.148（有限要素法による応力解析）の 4 つのタスクのデータを用いた。表 4.1 に各データの詳細を示す。なお、Doc. はドキュメントを表わす。

まず、3.3.1 節に示した方法でドキュメントベクトルを作成した。また、これらのドキュメントに対して、SVD と ICA を適用し、ドキュメントを潜在的意味空間上で表現した。なお、SVD では累積寄与率 r_k を 0.6 から 1 まで 0.1 刻みで変化させ、右特異ベクトル V_k を作成した。各データセットの r_k の変化による特異ベクトルの数を表 4.2 に示す。なお、r_k が 1 の場合の特異ベクトルの個数が、D のランク数と一致する。

また、潜在的意味空間上で、交叉検定によりユーザプロファイルの作成を行ない、11 点平均適合率により評価を行なった。

ユーザプロファイルの作成には 2.3.2 節で示した単純 GA を利用する。GA の適合度関数として、式（2.5）を用いた。他のパラメータを表 4.3 に示す。GA による結果にはばらつきが生じるため、同じデータに対して、それぞれ、5 回ユーザプロファイルの作成を行ない、

表 4.1 実験データ

Task No.	Doc. #	Interesting Doc. #	Indexing Word #
109	453	21	6,348
110	625	34	5,548
121	500	63	1,462
148	480	14	6,990

表 4.2　r_k による特異ベクトル数

Task No. \ r_k	0.6	0.7	0.8	0.9	1.0
No.109	111	151	204	280	453
No.110	158	218	295	400	623
No.121	52	83	127	201	497
No.148	40	81	142	240	477

表 4.3　GA のパラメータ

Generation	Population	Crossover	Mutation
10000	100	1	0.005

推薦精度の平均値を結果とした。ここでは、SVD による次元削減の効果を確認するため、ICA のみを用いた実験（ICA）、SVD によって低次元化を行なった実験（SVD+ICA）を行なった。この 2 つの実験に対して、$r_k = 0.7$ として処理時間を計測した。なお、SVD を適用しないで ICA を行なった場合、文書行列 D のランク数だけトピックが得られる。

また、提案手法によるユーザプロファイル作成の有効性を検討するために、LSA との比較実験を行なった。その際、LSA においても r_k を 0.6 から 1.0 まで変化させ実験を行なった。

4.4.2　実験結果

ICA の処理時間を表 4.4 に示す。SVD+ICA では、SVD の処理を行なう時間も含んでいる。これらの処理は Pentium4 3.2GHz、メモリ 1GB のマシンで Matlab ver. 6.1 を用いて行なった。つぎに、$r_k = 0.7$ における、ICA と提案手法の 11 点平均適合率の比較結果を表 4.5 に示す。各手法のそれぞれの累積寄与率 r_k における提案手法と LSA の 11 点平均適合率の

推移を図 4.1 から図 4.4 に示す。

表 4.4　処理時間の比較

Task No.	SVD+ICA(min)	ICA(min)
109	4	262
110	10	333
121	1	130
148	3	120

表 4.5　11 点平均適合率の比較

Topic No.	ICA	SVD+ICA
109	0.259	**0.407**
110	0.125	**0.149**
121	0.228	**0.260**
148	**0.056**	0.054

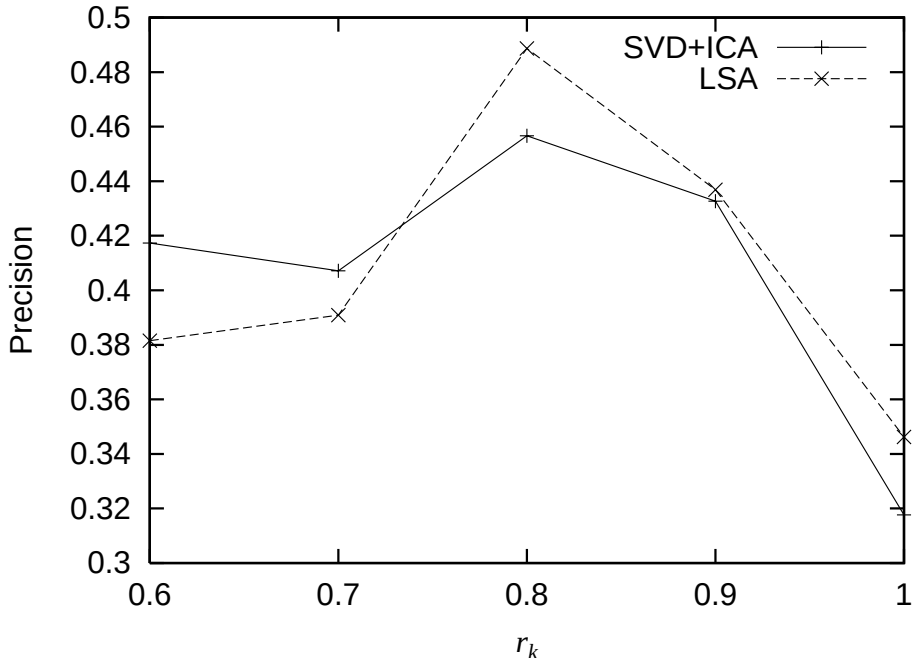

図 4.1　No.109 の r_k による 11 点平均適合率の推移

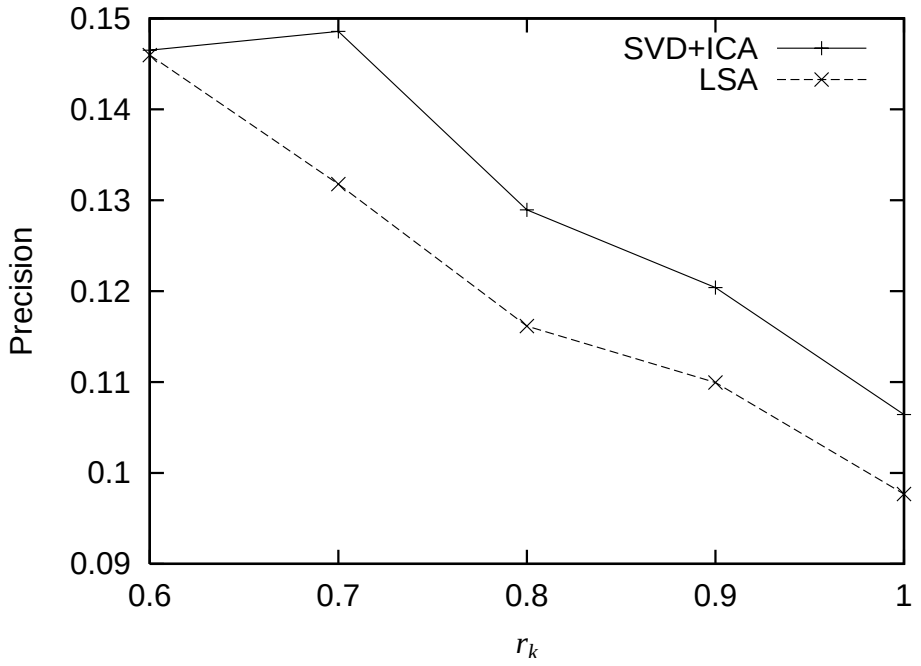

図 4.2　No.110 の r_k による 11 点平均適合率の推移

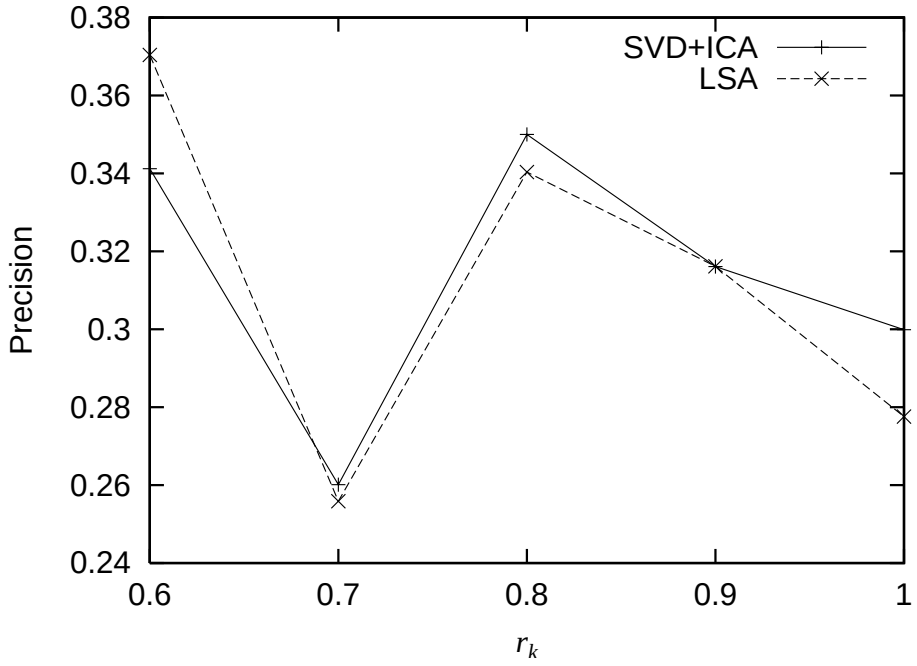

図 4.3 No.121 の r_k による 11 点平均適合率の推移

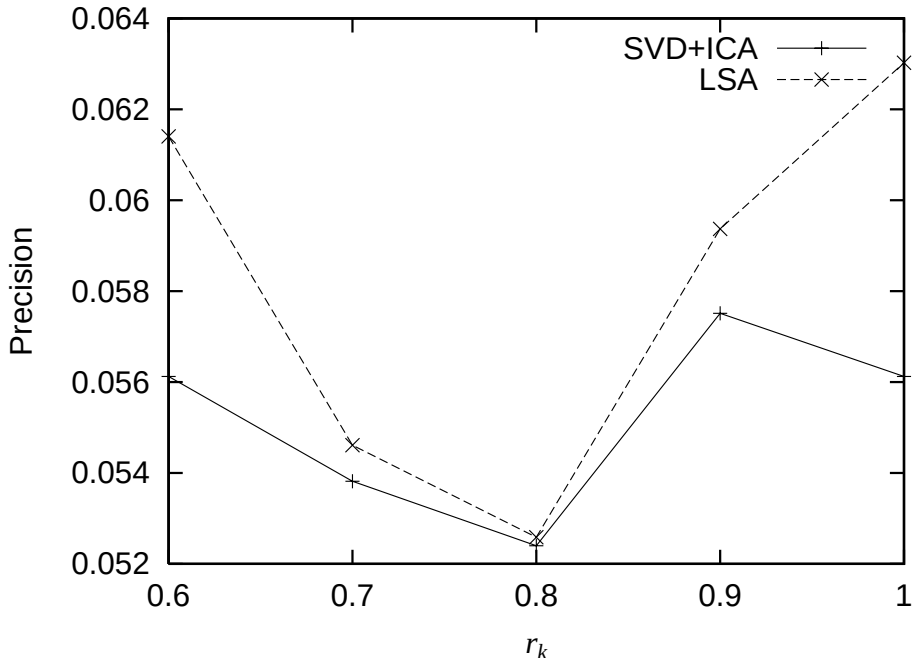

図 4.4 No.148 の r_k による 11 点平均適合率の推移

最後に、$r_k = 0.7$ として得られた提案手法と LSA の潜在的意味を表 4.6 に示す。表 4.6 では、各潜在的意味の重みの高い単語を示している。LSA の潜在的意味としては、第 1 主成分を取り上げた。提案手法の潜在的意味は、LSA の第 1 主成分の中でもっとも重みの高い単語が共通して現れる成分を選んだ。

表 4.6　潜在的意味の例

Task No.	Method	Top words
No.109	LSA	翼、インターフェース、計算
	SVD+ICA	翼、アルゴリズム、計算
		翼, 干渉, パケット
No.110	LSA	シーケンス、ファジィ、国際様式
	SVD+ICA	シーケンス、ファジィ検索、ファジィ
		シーケンス, アニメーション, アニメーションアルゴリズム
No.121	LSA	イオン、酸素, 樹脂
	SVD+ICA	イオン、プラスチック、樹脂
		イオン, アンケート, 酸素
No.148	LSA	弧、マントル、地震
	SVD+ICA	弧、マントル、対流
		弧、導波路, 地震

4.5　検討と考察

表 4.4 より、本手法の SVD による低次元化処理により、ICA の処理時間が大幅に短縮できていることが分かる。No.110 の処理時間が他のタスクに比べて長いのは、推定すべき独立成分数が多かったためであると考えられる。また、表 4.5 より、トピック No.148 を除い

て提案手法の方が通常の ICA に比べて 11 点平均適合率が改善していることが分かる。11 点平均適合率の低下と処理時間の短縮を検討すると、本手法は有効である。

図 4.1 を見ると、No.109 のデータに関しては、r_k が低い位置において、提案手法では LSA より 11 点平均適合率の改善が見られる。しかし、r_k が高くなると、LSA の方が適合率が良い。このタスクのタイトルは “TCP の高速化”である。LSA では、航空力学に関する単語が上位に現れている。これは、このタスクには高速艇に関するドキュメントが含まれており、その中の水中翼に関する部分からこのような索引語が得られたと考えられる。一方、提案手法で得られた潜在的意味には r_k の低い場合においても、これら通信関連の単語が上位に含まれていたため、r_k の低い位置における改善が見られたと考えられる。しかし、取得する潜在的意味の数を多くすることにより、LSA でも通信に関する単語を含む成分が取得できたため、r_k を高くすると 11 点平均適合率の逆転現象が現れたと考えられる。

No.110 のデータでは、図 4.2 に見られるように、LSA に比べて、全体的に提案手法の平均適合率が常に 1% から 2% 改善している。これは、潜在的意味として適切なトピックが取得できたためと考えられる。また、r_k により次元削減を行なうことで 11 点平均適合率の改善が見られる。

No.121 のデータでは、図 4.3 に示すように、提案手法と従来手法では r_k の変化に対して、ほぼ同じように振る舞う。これは、このタスクが “ネットワークを用いた VoD システム”というテーマであるにも関わらず、LSA、提案手法共にそれらのトピックを的確に追従できず、同じような潜在的意味が提案手法でも得られたためであると考えられる。

No.148 のデータでは、図 4.4 に示すように、提案手法は全体的に LSA に比べて 11 点平均適合率が悪くなっている。このデータの特徴として、元々、11 点平均適合率が他のデータに比べて低いことが挙げられる。潜在的意味を見ると、地質学に関するドキュメントを集めたデータであることが推察されるが、実際このタスクは有限要素法を用いたコンクリート構造物の応力解析に関するドキュメントが集られている。また、それらのドキュメントを実際

の主なトピックと異なった潜在的意味で再表現したことにより、一層分離が困難になったと考えられる。

以上、4 つのタスクに対して実験を行なったが、それぞれのタスクごとに提案手法の有効性が異っている。まず、LSA と提案手法の差異が余り現れていない No.121 の特徴として他のタスクに比べて索引語数が少ないことが考えられる。それゆえ、索引語数が少ない場合には、本手法は LSA と同じような結果が得られると考えられる。また、No.148 では索引語数は多いが、提案手法の効果が得られていない。この原因としては、このタスクでは興味のあるドキュメント数が少ないことが考えられる。興味のないドキュメントが少ないということは、タスクのトピックに関連の薄いドキュメントが文書集合中に多いことを意味している。それゆえ、得られる潜在的意味がそれらの関連度の低いドキュメントの影響を強く受る。また、本実験では $r_k = 0.7 \sim 0.8$ が妥当な値であることが示された。

表 4.6 により得られた潜在的意味を検討すると、LSA で得られる主成分の軸から、更に拡張ないし細分化されたトピックが、提案手法で取得できていると考えられる。No.109 の主成分では “翼”や “干渉”、“計算”といった単語が見られ、提案手法で得た潜在的意味では、“計算”から派生した “アルゴリズム”などの単語が見られる。また、“パケット”などの単語が現れていることから、“干渉”という単語が情報科学における干渉と捉えることができる。これより、通信に関するトピックを追従できていると思われる。

No.110 の主成分では、“シーケンス”、“ファジィ”といった単語と同時に、“国際様式”といった関係のない成分が含まれている。提案手法で得られた潜在的意味においては、“ファジィ”から “ファジィ検索”と拡張したものや、“シーケンス”から “アニメーション”へと広げていった例が見られた。このタスクは、情報検索の視覚化というタスクである。その中において、昨今のインターネットの普及による生活様式などの変化について言及したドキュメントが存在した。それらの影響から、第 1 主成分において、“国際様式”といった単語が出現したと考えられる。一方、提案手法の潜在的意味としては、“ファジィ検索”や、視覚化という

意味での “アニメーション”などトピックに対して的確に追従した成分が得られたと考えられる。

No.121 の主成分では、“イオン”や “酸素”、“樹脂”といった単語が現れていた。さらに、潜在的意味では、“樹脂”から “プラスチック”へ化学分野でより細分化されたり、“アンケート”といった関係性の薄いトピックと混じってしまうことが見られた。

最後に、No.148 は、“弧”や “地震”、“マントル”など地質学的なトピックと考えられるが、このデータにおいて得られた潜在的意味では、“マントル”の “対流”や、“地震”の “導波路”といったトピックが得られた。

4.6 まとめ

本章では、SVD によりドキュメントベクトルを低次元化し、ICA による潜在的意味抽出の処理時間を短縮する手法を提案した。さらに、得られたトピックを利用し、ユーザプロファイルの作成を行なった。その結果、ICA だけによる処理とほぼ同等かそれ以上の精度を持つユーザプロファイルを作成しつつ、ICA の処理時間を短縮することができた。また、LSA との比較を行ない、本手法の有用性を確認した。最後に、本章で用いた潜在的意味を考察することにより、それらの潜在的意味がドキュメント中のトピックを表わしていることも確認した。今後は、従来手法に比べてユーザプロファイルの精度が悪くなっている部分に対してさらなる検討を行なうとともに、SVD 以外のベクトル低次元化手法についても検討を行なっていく予定である。また、Web ページなどの大規模な実運用を想定した場合には、さらなる高速化が求められると考えられる。

第５章

独立成分分析における混合行列を用いた情報フィルタリング

5.1　はじめに

前章までは、SVD と ICA を組み合わせて、ICA の処理時間を短縮すると同時にフィルタリング精度の改善に取り組んだ。一方、4.2 節で紹介した文献 [15] や文献 [18] では、ICA を行列分解の一手法とみなし、得られる混合行列をトピックとして取得する研究が報告されている。本章では、上記の手法で得られる ICA の混合行列による潜在的意味空間上で、ユーザの興味を用いた情報フィルタリングを実現することを提案する。

この手法では、前章までのように基底の独立性といった性質は満たさないが、トピックとしてより直感的に理解しやすいものが得られるという特徴がある。また、SVD を適用する段階において累積寄与率を用いたノイズ低減も行なうことができる。

本章では、混合行列のトピックを用いてドキュメントを再表現し、GA によってユーザプロファイルを作成して、それらのトピックの情報フィルタリングへの応用を提案する。また、NTCIR2 を用いた評価実験を行ない、その有用性について検討を行なう。さらに、得ら

れるトピックについても考察を行なう。

5.2　提案手法

この節では、混合行列によるトピックの抽出手法と、そのトピックを用いたドキュメントの再表現方法について述べる。

5.2.1　SVD による低次元化

本章の手法では、前章までの手法で用いた文書-単語行列ではなく、LSA で用いる SVD と同様に行列分解の一手法と考え、単語-文書行列を用いる。いま、n 個のドキュメントベクトルを $\mathbf{d}_1, \cdots, \mathbf{d}_n$ とすると、単語-文書行列 $D^\dagger$ は、以下のように表わせる。

$$D^\dagger = [\mathbf{d}_1 \quad \mathbf{d}_2 \quad \cdots \quad \mathbf{d}_n] \tag{5.1}$$

$$= D^t \tag{5.2}$$

さらに、SVD を用いると、$D^\dagger$ は次のように分解される。

$$D^\dagger = U\Sigma V^t \tag{5.3}$$

U と V の列ベクトルは $D^\dagger$ の直交基底であり、特異ベクトルと呼ばれる。Σ は、特異値 σ_i を対角成分にもつ対角行列である。なお、σ_i は、$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$ を満す。p は、$D^\dagger$ のランク数を示す。ここでは、対応する特異値の大きい k 個の特異ベクトルを選択する。k の決定には、式（4.3）の累積寄与率を用いる。それらの k 個の特異ベクトルを並べた行列 U_k を作成し、U_k を用いて、次式のように単語-文書行列を変換する。

$$D_k^\dagger = U_k^t D^\dagger \tag{5.4}$$

5.2.2 ICA の混合行列による特徴抽出

ドキュメントに ICA を適用する場合、入力の時系列信号がドキュメントベクトルに対応する。いま、m 個の独立成分を、$\mathbf{s}_1, \cdots, \mathbf{s}_m$ とし、それらを並べた行列 S を次のように定義する。

$$S^{\dagger} = [\mathbf{s}_1 \quad \mathbf{s}_2 \quad \cdots \quad \mathbf{s}_m] \tag{5.5}$$

ここで、$\mathbf{s}_i$ は、要素数 k の列ベクトルである。このとき、5.2.1 節 で述べた、$D_k^{\dagger}$ は次のように表現できる。

$$D_k^{\dagger} = MS^{\dagger} \tag{5.6}$$

ここで、M は、$k \times k$ の混合行列である。

ICA では、M の逆行列 W を推定する。そのとき、式（5.6）は、次のようになる。

$$Y = WD_k^{\dagger} \tag{5.7}$$

本章では、混合行列 M、つまり、W の逆行列の列ベクトルをドキュメントの潜在的意味と考える。ただし、ICA の処理の前に、特異ベクトルの行列 U_k によって空間変換を行なっているので、式（5.8）によって、それらの潜在的意味をもとの空間に戻す。

$$F = U_k M \tag{5.8}$$

さらに、得られたもとの空間上の潜在的意味 F を用いて、ドキュメントを表現する。そのために、次の処理を行なう。

$$\hat{D} = DF \tag{5.9}$$

5.3　実験と結果

この節では、手法の有効性を確認するために行なった実験について説明し、結果について検討する。

5.3.1　実験環境と実験方法

実験データとして、テストコレクション NTCIR2 から、タスク No.110（情報の可視化）に含まれる 625 件のドキュメントを用いた。これらのドキュメントには、タスクテーマとの関連度合いが既に付与されている。また、625 件のうちが、34 件が関連度が高いとされている。

まず、3.3.1 節に示した方法で、各ドキュメントをドキュメントベクトルで表現した。それらのドキュメントベクトルを並べた、単語-文書行列に対して、SVD を適用し、累積寄与率が 0.8 となるまで特異ベクトル数を増加させ、その特異ベクトルが張る空間へ単語-文書行列を写像した。なお、その際の特異ベクトルの本数は 409 本になった。その後、ICA を適用し、特徴を取り出した。なお、ICA の処理では、Fast ICA アルゴリズムを用いた。その際の式（2.16）中の非ガウス性のコスト関数には $\tanh(\nu)$ を用いた。

最後に、得られた潜在的意味に写像したドキュメントを用いて、ユーザプロファイルの作成を行なった。ユーザプロファイルの評価のため交叉検定を用いた。625 件のドキュメントを 5 つに分割し、そのうち 3 つをユーザプロファイルの作成のための学習データ、残り 2 つをそのユーザプロファイルを評価するための評価データとした。5 つに分割したデータの詳細を表 5.1 に示す。

また、ユーザプロファイルの作成には、単純 GA を用いた。単純 GA の評価関数は 2.3.2 節に述べたものを利用した。得られたユーザプロファイルを用いて、評価用ドキュメントの

表 5.1 実験データ

データセット	ALL	Set1	Set2	Set3	Set4	Set5
興味あり	34	7	13	5	3	6
興味なし	591	118	112	120	122	119

推薦を行ない、再現率・適合率曲線と 11 点平均適合率により評価を行なった。推薦を行なう際には、式（2.6）の類似度を用いた。

なお、ユーザプロファイルは各データに対して 5 回ずつ作成し、それらの平均を結果とした。

以下に、実験の流れをまとめる。

1. ドキュメントベクトルを作成する。
2. SVD を単語-文書行列に適用し、特異ベクトルにより空間を変換する。
3. ICA を適用する。
4. 得られた混合行列をもとの空間へ変換する。
5. 得られた潜在的意味によりドキュメントベクトルを変換する。
6. GA によりユーザプロファイルを作成する。
7. 再現率・適合率曲線、11 点平均適合率を用いてユーザプロファイルを評価する。

さらに、比較手法として変換を行なわない場合（Original）、ICA のみの場合（ICA）についてユーザプロファイルを作成し検討を行なった。GA のパラメータについては、表 5.2 に示す。

表 5.2　GA のパラメータ

	世代数	交叉率	突然変異率
Original	10000	1	0.005
ICA	10000	1	0.05
Proposed	10000	1	0.05

5.3.2　実験結果

この節では、実験結果について述べる。図 5.1 は、再現率・適合率曲線を表わしている。表 5.3 は、11 点平均適合率を示している。表 5.4 は、得られた潜在的意味の例を示している。表 5.4 では、各潜在的意味に対して、索引語をその重みによって並べ替え、そのうち重みの大きい索引語 10 単語を例示した。表 5.5 は、本章で用いた潜在的意味と ICA によって得られるトピックの比較である。表 5.4 同様、重みの大きな索引語 10 個を例示している。

表 5.3　11 点平均適合率.

	Original	ICA	Proposed
11 点平均適合率	0.118	0.125	0.163

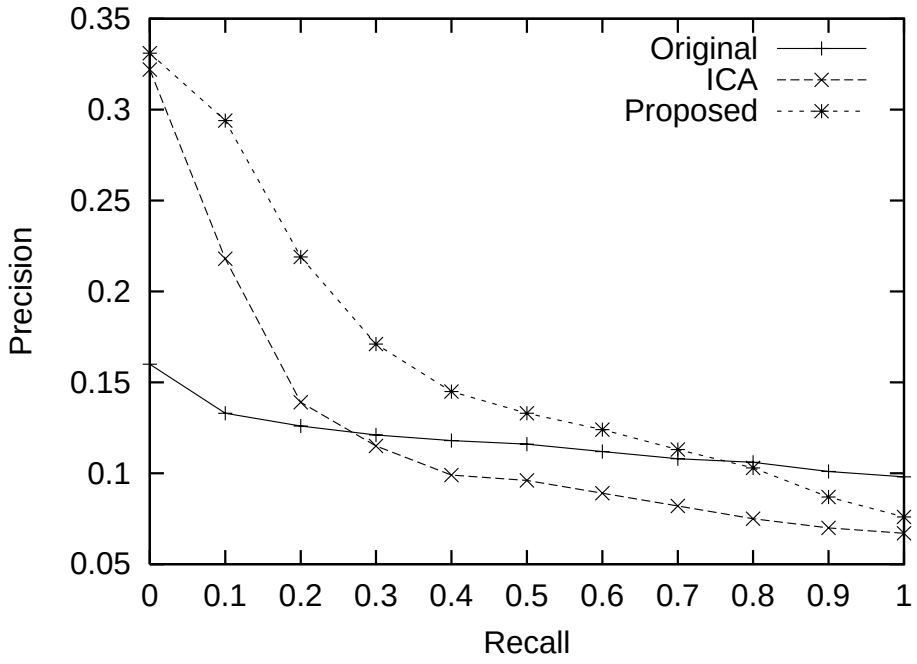

図 5.1　再現率・適合率曲線.

表 5.4　潜在的意味の例

潜在的意味 1	潜在的意味 2
横断言語	アレイ
情報アクセス	プロセッサアレイ
利用分野	プロセッサ
一覧段階	マッピング
文書選択	物理アレイ
翻訳表示	ルーティング
実利用	サイズ
技術情報	故障プロセッサ
国際交流	論理アレイ
ゲート	マッピングアルゴリズム

表 5.5　トピックの比較

混合行列	独立成分
顔部分空間	顔画像
顔領域	対象画像
切り出し	特徴マップ
顔切り出し	写真画像
顔認識	焼き付け
顔画像	顔画像の抽出
対象画像	ノイズ
有名人	ニューラルネットワーク
人名辞典	フィルタ処理
音声合成	シミュレーション

5.4　検討と考察

図 5.1 から、本章で提案した手法は、ICA だけを用いたものよりもフィルタリング精度を全体的に改善させていることが分かる。とくに、再現率の低い場所において、ICA 単体のものよりも改善幅が大きい。Original が全体的に適合率が低いのは、次元数の多さのために、GA による十分な探索が行なえなかったためと考えられる。計算機の処理能力の制約の観点などからも、空間を変換して次元を削減することは有効であることが、この結果からも分かる。また、表 5.3 に示した、11 点平均適合率からも、提案手法は他の手法に比べ改善が見られる。これは、潜在的意味のトピックとしての精度が改善されたためと考えられる。

つぎに、その潜在的意味について検討を行なう。表 5.4 を見ると、潜在的意味 1 は、“横

断言語”や “翻訳表示”などの単語から、言語横断検索に関するトピックであると推定できる。また、潜在的意味 2 は、“プロセッサ”や “アレイ”といった索引語が多く並んでおり、演算処理装置に関するトピックであると分かる。表 5.5 を見ると、本章で用いた潜在的意味も、ICA によって得られたトピックのどちらも顔画像に関するトピックを表わしていると考えられる。しかしながら、本章で用いた潜在的意味の方が、より明確に “顔認識”に関するトピックであると推定できる。このように、ICA によって得られるトピックよりも、より人の直感に近いトピックを得ることが可能であると考えられる。

5.5　まとめ

本章では、文献 [15] や文献 [18] で提案されているトピックを情報フィルタリングに応用することを検討した。また、それがフィルタリング精度の改善に効果があることを確認した。さらに、得られるトピックについても検討を行ない、従来述べてきた手法に比べ、より直感に近いトピックが得られることを確認した。

第 6 章

独立成分分析による索引語選別を用いた情報フィルタリング

6.1　はじめに

ドキュメントを対象とした情報フィルタリングにおいて、ベクトル空間法によるドキュメントの表現方法がしばしば用いられる。ドキュメントベクトルは、ドキュメント中に含まれる語（索引語）に対する重みを要素とするベクトルである。一般的に、取り扱うドキュメント数が多くなるにつれて索引語は増加していくため、ドキュメントベクトルの次元数も大きくなり、処理に必要なメモリや時間も膨大となる。そこで、索引語を選別することにより、ドキュメントベクトルを低次元化することを考える。

文献 [51] や文献 [52] では、単語間の共起関係に着目して、単一ドキュメントや、Web ドキュメント中から重要な語を抽出する手法が述べられている。これらの手法では、ドキュメントないしドキュメント集合中における高頻度語とそれらと共起関係が強い単語を選別している。この手法は、基準となる語によって抽出される単語が異なってくるので、基準となる語の選別の過程は重要である。本章では、単語頻度という基準ではなく、ICA によって得ら

れるドキュメント集合に含まれるトピックを用いて、それらの基準となる語を選別することを考える。基準となる語をトピックという概念で捉えることにより、ドキュメント集合中の小さなトピックに含まれる索引語も選別できると考えられる。情報フィルタリングでは、抽出すべきドキュメント数は、ドキュメント集合全体の要素数に対して少ない場合が多く、高頻度語が示す大まかなトピックだけではなく、様々なトピックを捉える必要がある。

本章では、これまでに述べた ICA によって得られるトピックにおける重要語を用いて文献 [51] 同様、χ^2 値により索引語を選別する。また、得られた索引語を用いてドキュメントベクトルを再構築し、重心を用いたユーザプロファイルによるフィルタリングを行ない、提案手法の有効性を確認する。

6.2 関連研究

本章で提案する索引語選別は、文献 [51] にヒントを得たものである。文献 [51] では、単一ドキュメントから、重要語を選別する手法を提案している。具体的には、ドキュメント中の高頻度語を基準にし、その高頻度語と際だって共起する語をドキュメント中における重要語としている。その共起の偏り具合を判定するために χ^2 値を用いている。一方、文献 [52] では、Web 上のページから重要語を抽出する手法を提案している。ここでも、高頻度語を基準とし、それらに関連性の高い単語を抽出している。また、基準語として用いる語の出現頻度が少なすぎると、的確な重要語が抽出できないことも述べられている。

文献 [51] では、単一ドキュメント中から重要語を選別するため、高頻度語をその文書の特徴語と考えることができた。しかし、それをドキュメント集合に適用する場合、コーパス内に存在する様々なトピックを抽出できないと考えられる。そこで、本章では、上述の基準となる語の選別に ICA によって得られるトピックを利用する。一般的な情報フィルタリングにおいて、必要なドキュメントの数は、不必要なドキュメントの数に比べて少ない場合が

多く、それら少ないドキュメントの傾向を捉えることが重要であると考えられる。したがって、高頻度語ではなく含まれるトピックによって選別することは、有用であると考えられる。

6.3 提案手法

本章では、ICA によって得られるトピックを用いて、索引語を決定する。また、得られた索引語を用いてドキュメントベクトルを作成する。最後に、重心を用いて作成したユーザプロファイルを用いて推薦を行ない、提案した索引語選別の有効性を検討する。本節では、ICA を利用したトピックから基準語を抽出する方法と ICA によって得られた基準語から索引語を選別する方法について説明する。

6.3.1 ICA を用いた基準語抽出

ICA によって得られたトピックは、2.4.1 節で示したように、単語に対する重みで構成されている。いま、重みの大きい単語が、そのトピックにおける重要な語と考えられる。そこで、本章では、その重要な語の中でも、もっともトピックを特徴づける単語、つまり、各成分中においてもっとも重みの大きな単語を基準語 g として抽出する。なお、ICA によって得られる独立成分には符号の任意性が存在するため、本手法では、各成分の重みの絶対値が最大の単語を基準語とする。また、そのようにして得られた基準語の集合を G と表記する。

6.3.2 χ^2 値を用いた索引語候補の選別

基準語に基づいて索引語候補を選別するために、本章では、基準語 g と各単語の共起確率分布と各単語の出現分布の違いに着目する。それらの分布の偏りは χ^2 値を用いて評価する。ここで、基準語以外の単語を t とする。もし、単語 t とすべての基準語 G との共起確率分布と t の出現確率分布の違いが大きければ、χ^2 値は大きくなる。逆に、単語 t がどの基準語

ともまんべんなく共起していたならば、χ^2 値は小さいものとなる。ドキュメントの特徴を良く表わす単語は、そのドキュメントの持つトピックの基準語と共起すると考えられる。χ^2 値の高い単語を選べば、全基準語の出現確率分布から外れた特徴的な単語を抽出できると考えられる。したがって、それらの特徴的な単語を索引語として用いる。

p_g を基準語 G 中の g の出現確率とすると、次のように表わせる。

$$p_g = \frac{g \text{ を含むドキュメント数}}{G \text{ を含む全ドキュメント数}} \tag{6.1}$$

また、n_t を単語 t と基準語集合 G の共起頻度、$freq(t, g)$ を単語 t と基準語 g との共起頻度とすると、単語 t の χ^2 値は次のようになる。

$$\chi^2(t) = \sum_{g \in G} \frac{(freq(t, g) - n_t p_g)^2}{n_t p_g} \tag{6.2}$$

なお、本章では、t と g が共起するとは、同一ドキュメント中にそれら両方が含まれていると定義する。また、χ^2 値の高い単語と基準語を索引語とする。

6.4　実験と結果

本節では、提案手法によって得られる索引語を用いて作成したユーザプロファイルの有効性を確認するために行なった評価実験について説明する。

6.4.1　実験方法

まず、本手法の有効性を確認するために、小規模なデータにおいて実験を行なった。実験データは、4 件の正解を含む 100 件のドキュメントを、それぞれ、正解を 2 件ずつ含む、80 件の学習データと 20 件の評価データに分割して作成した。その後、さらなる有効性の検討のため、テストコレクション NTCIR2 から、トピック No.110、トピック No.114、トピック No.115、トピック No.148 の 4 つの各データを用いて、それぞれ、実験を行なった。表

表 6.1 実験データ

Topic No.	Doc.#	Interest Doc.#	Index words #
110	625	34	5,448
114	612	20	8,709
115	930	94	12,055
148	480	14	6,990

6.1 に各データの詳細を示す。なお、興味の有無のラベル付けは、2.5.1 節と同様の方法で行なった。

まず、3.3.1 に述べた手順でドキュメントベクトルを作成した。これらのドキュメントに対して、ICA を適用し、各独立成分から一つずつ基準語を抽出した。また、上記の手順で得られた全基準語から、約 30% と 10% に基準語を削減した場合の実験も行なった。なお、30%、10% の場合については、独立成分の出現順序には任意性があるので、ドキュメント集合中における基準語の出現頻度の多いものを選んだ。表 6.2 に、各実験におけるデータごとの基準語数を示す。

表 6.2 各実験における基準語数

Topic No.	10%	30%	ALL
No.110	60	205	608
No.114	62	185	597
No.115	95	273	889
No.148	46	140	462

そのようにして得られた基準語を用いて、χ^2 値により索引語候補を選別、基準語と合わせ

て索引語とした。なお、取得する索引語数は全索引語の約 50%、30%、10% について実験を行なった。

最後に、得られた索引語を用いてドキュメントベクトルを作成し直し、学習ドキュメントの重心を用いてユーザプロファイル **u** を作成した。ユーザプロファイルの作成には、式 (2.5) を利用した。なお、式 (2.5) 中の係数 α と β は、式（2.21）を用いた。また、作成したユーザプロファイルを用いて推薦を行ない、再現率・適合率曲線と、11 点平均適合率により評価を行なった。さらに、提案手法の索引語選別による低次元化の有効性を検討するために、索引語選別を行なわない場合（ORIG）と、ドキュメントベクトルの低次元化で用いられる潜在的意味解析 (LSA) を用いた場合の実験を行なった。

6.4.2　実験結果

まず、小規模で行なった実験の結果を表 6.3 に示す。表 6.3 中の索引語数は全単語数に対する使用した索引語数の割合を示し、基準語数は、全トピックから得られる基準語数に対する使用した基準語数の割合を示している。

表 6.3　小規模実験の結果

索引語数 ＼ 基準語数	30%	50%
10%	0.45	0.63
30%	0.85	0.85
50%	0.67	0.85
Original	0.63	

つぎに、NTCIR2 の Topic No.110 における、実験結果を示す。図 6.1 から、図 6.3 はそれぞれ基準語数が 10% 、30% 、ランク数 のときの各索引語数に対する適合率・再現率曲線である。また、表 6.4 は、各実験の 11 点平均適合率を示す。図 6.4 から、図 6.6 は Topic No.114 のデータにおける基準語数を 10% 、30%、ランク数と変化させたときの各索引語数に対する適合率・再現率曲線をそれぞれ示している。また、表 6.5 に Topic No.114 における 11 点平均適合率を示している。さらに、Topic No.115 における実験結果を示す。図 6.7 から図 6.8 は、基準語数を 10% 、30% 、ランク数としたときの各索引語数に対する適合率・再現率曲線をそれぞれ示しており、表 6.6 は各実験の 11 点平均適合率を示している。Topic No.148 における実験結果は以下のとおりである。図 6.10 から図 6.12 が、それぞれ基準語数を変えた時の適合率・再現率曲線を表わしている。11 点平均適合率は表 6.7 に示されている。最後に、11 点平均適合率の 4 つのデータにおける平均値を表 6.8 に示す。

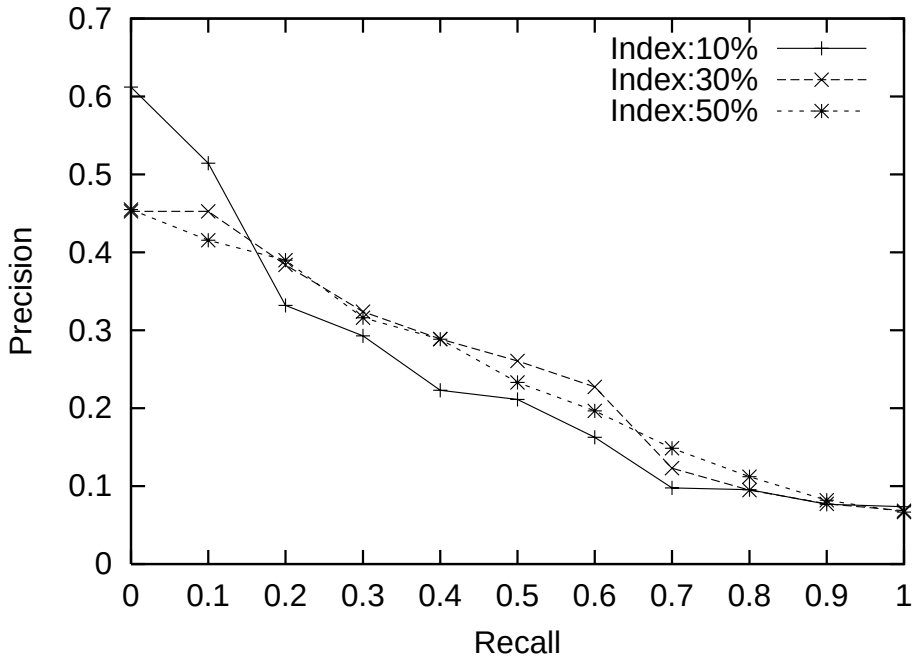

図 6.1　基準語数 10% 時の適合率の推移

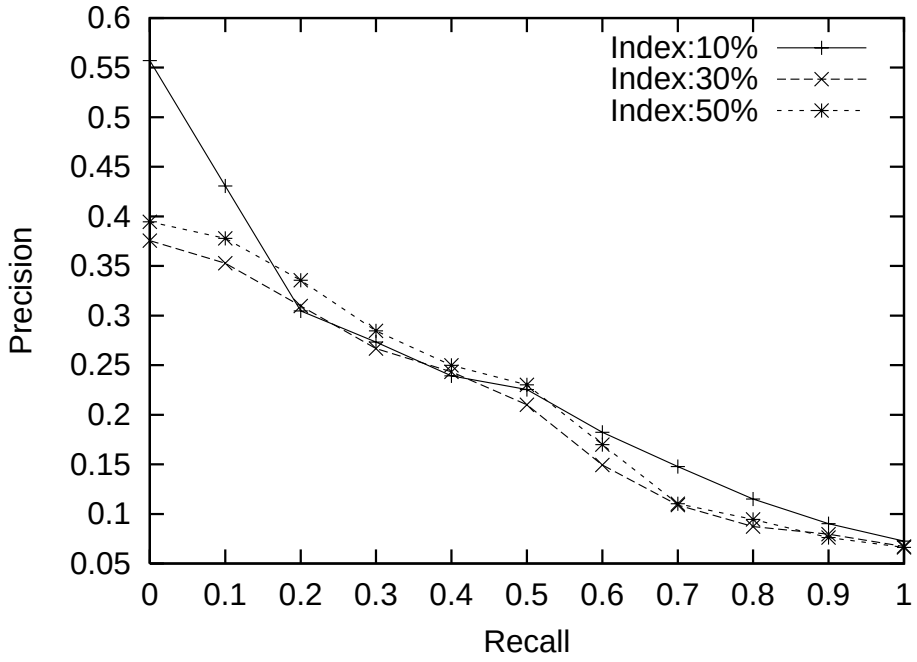

図 6.2　基準語数 30% 時の適合率の推移

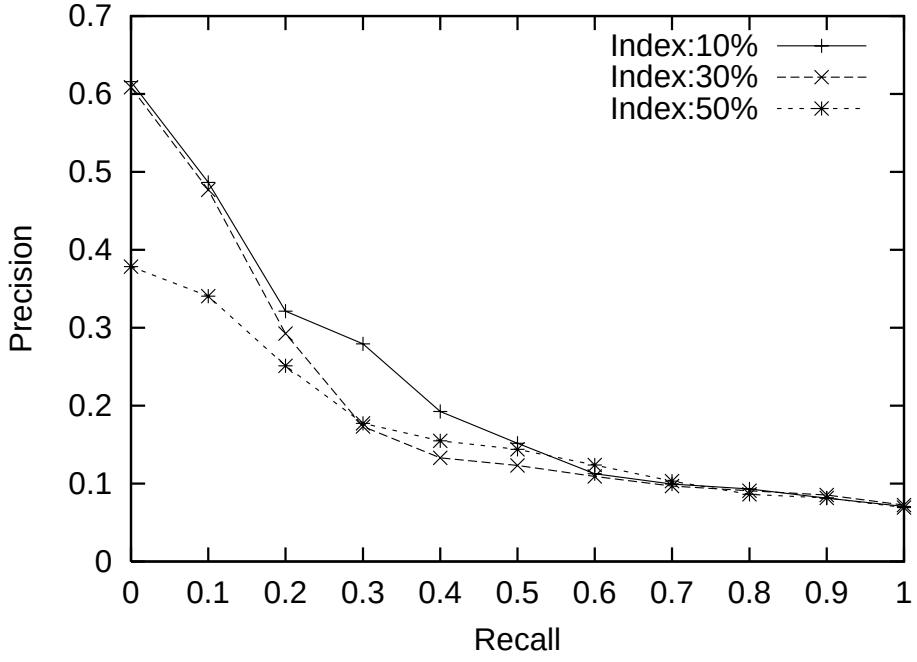

図 6.3　基準語数がランク数時の適合率の推移

表 6.4　Data110 の 11 点平均適合率

基準語数 \ 索引語数	10%	30%	50%
10%	0.24	0.25	0.25
30%	0.24	0.20	0.22
50%	0.23	0.21	0.17

Original	0.25	LSA	0.25

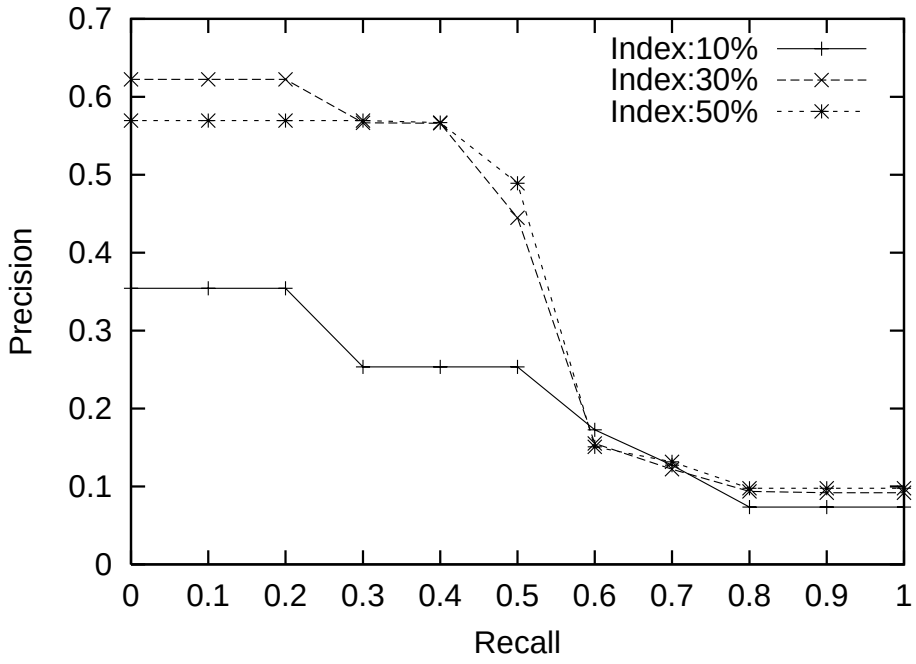

図 6.4　基準語数 10% 時の適合率の推移

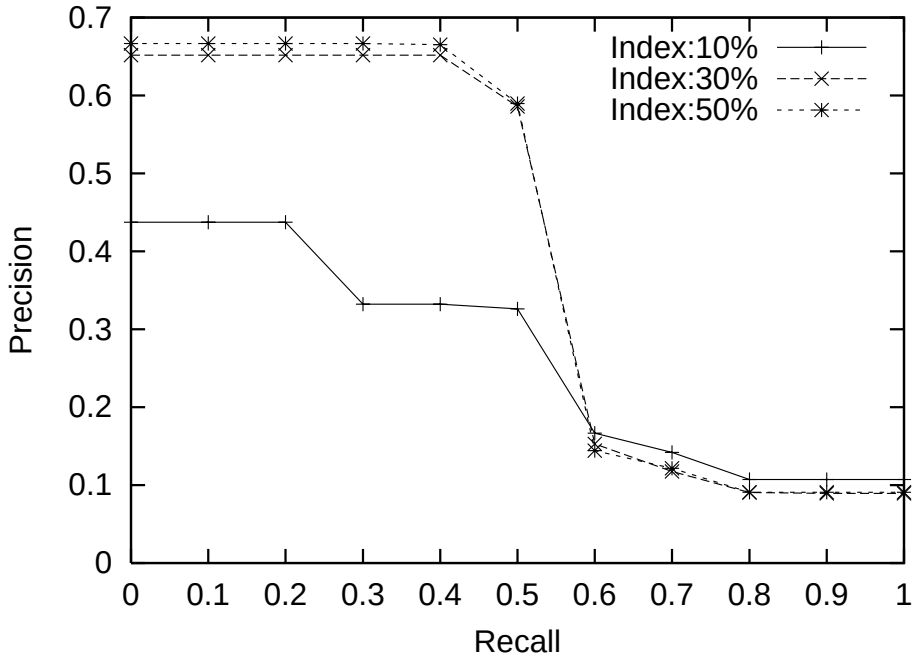

図 6.5　基準語数 30% 時の適合率の推移

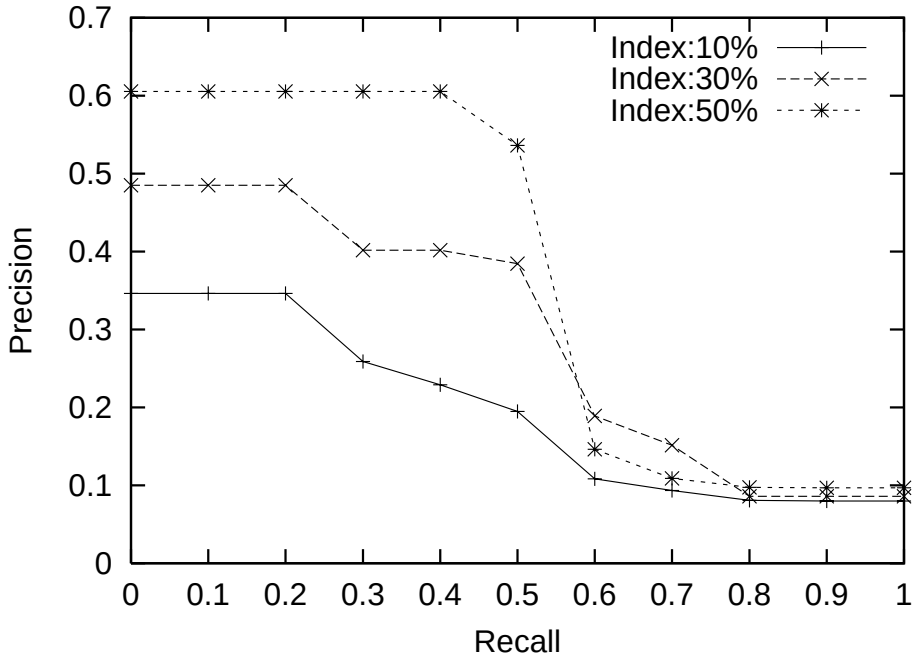

図 6.6　基準語数がランク数時の適合率の推移

表 6.5　Data114 の 11 点平均適合率

基準語数＼索引語数	10%	30%	50%
10%	0.21	0.36	0.36
30%	0.27	0.40	0.41
50%	0.20	0.29	0.37
Original	0.38	LSA	0.19

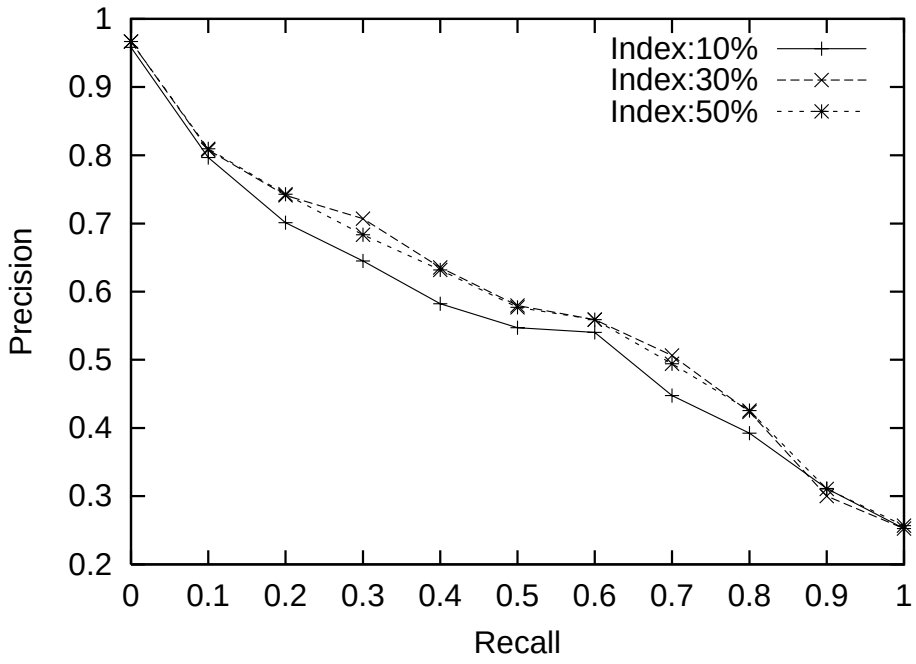

図 6.7　基準語数 10% 時の適合率の推移

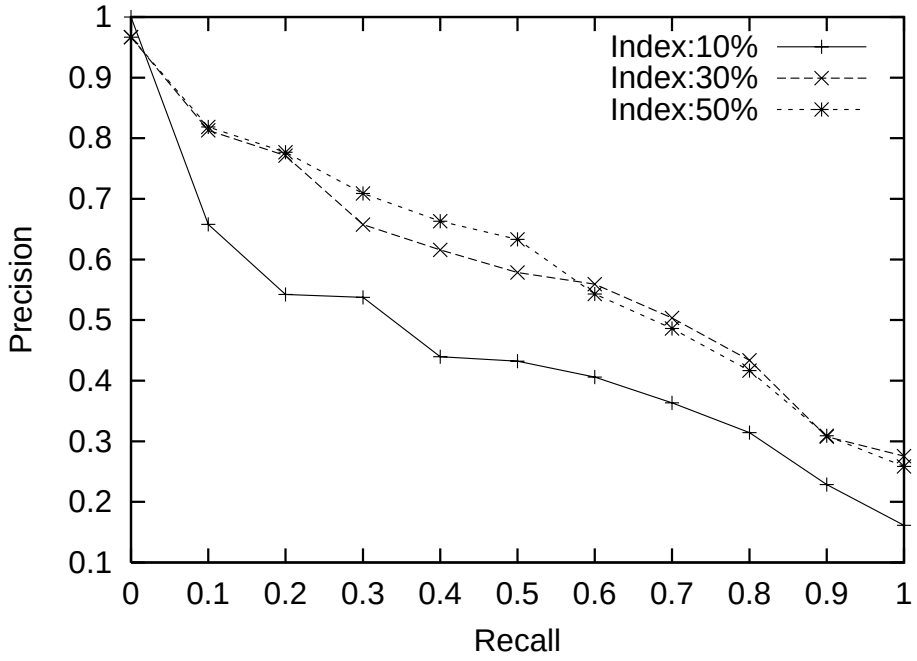

図 6.8　基準語数 30% 時の適合率の推移

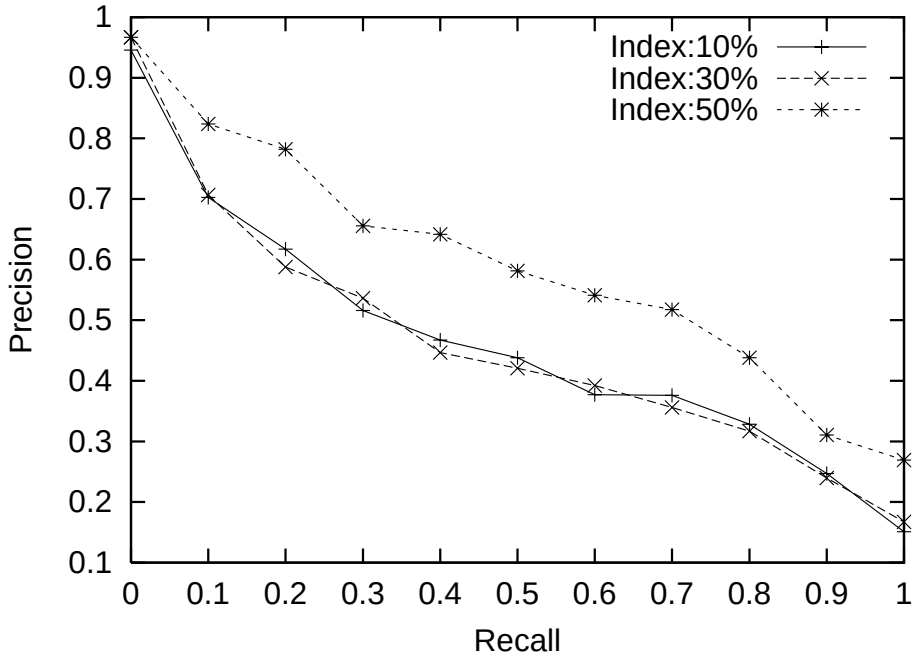

図 6.9　基準語数がランク数時の適合率の推移

表 6.6　Data115 の 11 点平均適合率

基準語数＼索引語数	10%	30%	50%
10%	0.56	0.59	0.59
30%	0.46	0.59	0.60
50%	0.47	0.47	0.60
Original	0.61	LSA	0.61

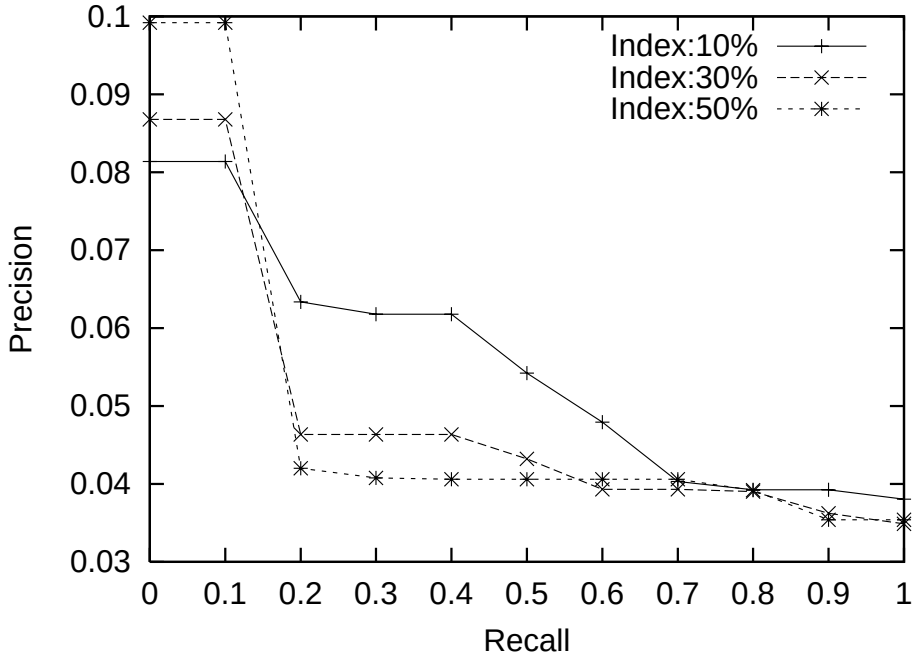

図 6.10　基準語数 10% 時の適合率の推移

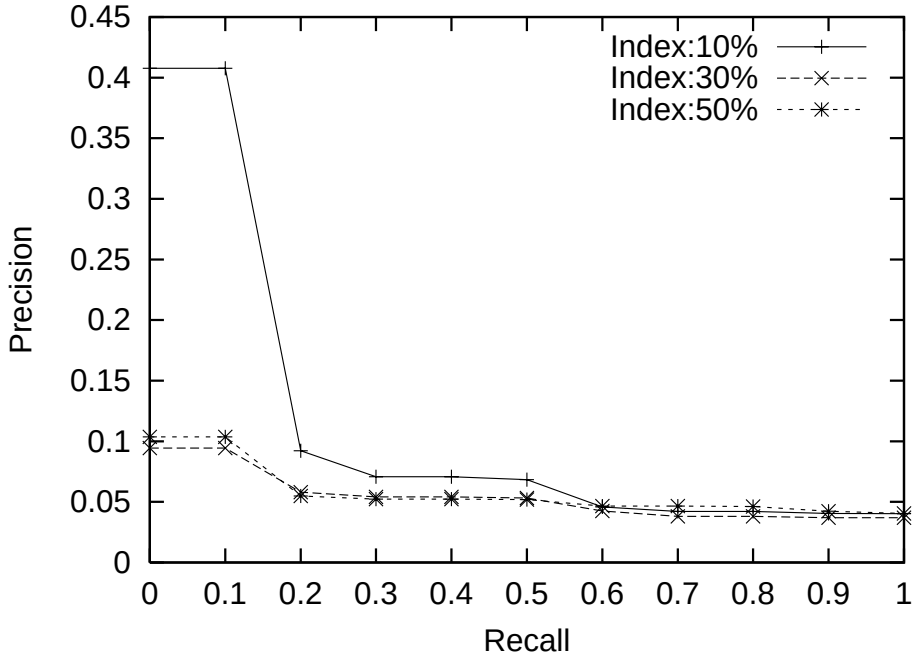

図 6.11　基準語数 30% 時の適合率の推移

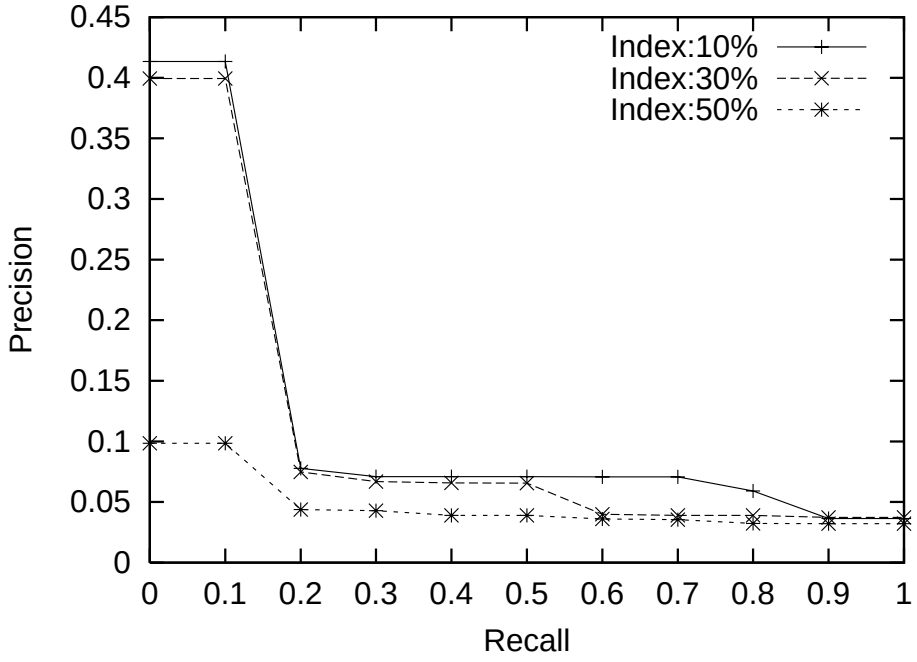

図 6.12　基準語数がランク数時の適合率の推移

表 6.7　Data148 の 11 点平均適合率

基準語数＼索引語数	10%	30%	50%
10%	0.06	0.05	0.05
30%	0.12	0.06	0.06
50%	0.13	0.12	0.05
Original	0.05	LSA	0.08

表 6.8　4 つのデータにおける 11 点平均適合率の平均値

基準語数＼索引語数	10%	30%	50%
10%	0.27	0.31	0.31
30%	0.27	0.31	0.32
50%	0.26	0.27	0.30
Original	0.32	LSA	0.28

6.5　検討と考察

まず、表 6.3 に示されている小規模で行なった実験の結果について検討する。このデータの特徴として、正解ドキュメント数が少ないことが挙げられる。これは、情報フィルタリングの一般的な場合、つまり、抽出すべきドキュメントが少ない場合に相当すると考えられる。表 6.3 を見ると、独立成分によって得られる全基準語を用いた場合に、索引語選別を行なわないものに比べて 11 点平均適合率が改善していることが分かる。よって、本手法は、

正解ドキュメントが少ない文書集合において、改善を見込める手法であると考えられる。つぎに、テストコレクションを用いた実験の結果について検討する。

まず、本章で用いた 4 つのデータにおける平均値を示した表 6.8 について考察を行なう。表 6.8 によると、基準語数を 10% から 30% くらいに設定すると、索引語数を削減しても、索引語選別を行なわないもとのデータとほぼ同程度の推薦精度を得られることが分かる。

以下、各データごとの結果の振る舞いについて検討する。最初は、Topic No.110 について考察する。図 6.1 から図 6.3 を見ると、どの場合においても索引語数を減らすことで、再現率の低い場所において、適合率の改善が見られる。しかしながら、再現率の高いところにおいては、さほど変化がないか、もしくは低下していることが分かる。11 点平均適合率を示した表 6.4 を見ても、索引語を削減しても Original と同程度の結果が見られる。これは、このドキュメントを表現するためにはそれほどの索引語数が必要でなかったと考えられる。

つぎに、Topic No.114 について述べる。図 6.4 から図 6.6 を見ると、基準語数に関わらず、索引語数を全索引語数の 10% まで減らすと、再現率の低い場所において特に適合率の低下が見られる。これは、索引語数が全単語数の 10% では、索引語数が少なすぎるために、すべてのドキュメントを表現することができず、分類が行なえなかったと考えられる。表 6.5 を見ると、本手法による削減で、索引語数を全体の 30% ほどにすることで、Original に比べて、同等ないし、適合率の精度を改善することができている。また、従来手法である LSA に比べても本手法が有効であることが分かる。

Topic No.115 に関しては、図 6.7 から図 6.9 を見ると、基準語数が少ない（全単語数の 10% ）のときは、索引語数によらず同じような推薦精度を示しているが、基準語数が増加したときに、索引語数が減ることで適合率の低下が見られる。表 6.5 を見ると、基準語数が 10% のときは、Original とさほど変わらない 11 点平均適合率が得られている。基準語数が 30% のときは、索引語数が全単語数の 30% 以上にすることで、Original と同程度の適合率が得られるが、索引語数を減らすと、適合率の低下が見られる。また、基準語数が 50% の

ときは、Original と同程度の適合率を得るためには、さらに多くの索引語が必要となっている。これは、No.115 のデータは、ドキュメントの集合に含まれているトピックが少なく、基準語として用いるトピックを絞り込むことと不必要なトピックが除去されるが、反対に基準語を増やすと、共起する単語が増えるため索引語としては効果的でない単語の χ^2 値も高くなる。索引語数を減らした場合は、それらが選択されることで、必要な索引語が選択されなかったのではないかと考えられる。

最後に、Topic No.148 のに対する実験結果の考察を行なう。図 6.10 から図 6.12 を見ると、索引語数を絞ることで、とくに、再現率の低い場所において推薦精度の改善が見られる。表 6.7 を見ると、上述の適合率改善の影響で、全体的な 11 点平均適合率が押し上げられ、Original や LSA に比べて改善されていることが分かる。

このように、本手法は索引語選別を行なっても、ユーザプロファイルによる推薦精度は、選別を行なわない場合に比べて、ほぼ同程度、場合によっては改善が見込めることが分かった。また、改善が顕著であったデータ No.148 の特徴として、Original の 11 点平均適合率が他のデータセットに比べて低いことがあげられる。これは、推薦すべき正解ドキュメント数が全ドキュメント数に対して少ないことが理由として挙げられる。提案手法では、χ^2 値を用いて、基準語と各語の共起頻度の偏りにより索引語を選別しており、No.148 のドキュメント集合において数の少ない正解ドキュメントの特徴が的確に抽出できたと考えられる。これは小規模で行なった実験の結果を裏付けるものとなった。以上より、とくに正解ドキュメント数の少ないデータにおいて提案手法は特に有効であると考えられる。

6.6 まとめ

本章では、ICA によって得られるトピックに基づいて、索引語を抽出を選別する手法を提案した。さらに、それらの索引語からドキュメントベクトルを作成し、重心を用いたユーザ

プロファイルによる推薦を行なってその有効性を確認した。その結果、索引語を削減しても、推薦精度はほぼ同程度ないし、場合によっては改善が見込めることが分かった。とくに、正解ドキュメント数が少なく、推薦精度が低いデータセットにおいて、提案手法が特に有効であることを確認した。また、この手法で得られるドキュメントベクトルは索引語によるので、さらに潜在的意味を解析するなどの応用も考えられる。今後は基準語を選別するために、トピック選別を行なう必要があると考えられる。

第 7 章

結論

本論文では、ユーザの興味情報（ユーザプロファイル）による情報フィルタリングにおいて、ドキュメントに対して ICA を適用した際に得られる潜在的意味（トピック）を用いることにより、その推薦精度を改善する手法について検討を行なった。本論文で得られた結果は以下のとおりである。

第 2 章では、まず、ICA によって得られるトピックを用いてドキュメントを表現し、それがユーザプロファイルの精度改善に有効であることを確認した。この章では、従来ユーザプロファイルの作成で用いられている重心を用いた手法と、遺伝的アルゴリズムを用いた手法の 2 つの手法によりトピックによって表現されたドキュメントからユーザプロファイルを作成した。その結果、どちらの手法においても、ICA によって得られるトピックでドキュメントを表現することが、ユーザプロファイルの推薦精度と作成時間の点において有効であることを確認した。

第 3 章では、ICA によって得られるトピックの中には、不必要なトピック（ノイズ）が含まれていると考え、それを除去することにより、フィルタリングの精度改善を試みた。LSA 等では、ノイズ低減の指標として累積寄与率などが利用されているが、ICA ではそのような基準は検討されていないため、この章では、MDA を用いてトピックのクラスタリングを行

ない、トピックを整理することで不必要な成分を除去した。また、ICA によって得られるトピックにはスケールと順序の任意性という性質が存在するため、従来クラスタリングを行なう際に良く用いられるユークリッド距離では、そのトピックの類似度を測ることができないという問題点があった。その問題点に対して、JS 情報量を導入した MDA を導入し、クラスタリングを行なった。また、クラスタリングを行ない、不必要な成分を除去した結果、第 2 章で得られた結果に比べさらなる改善が確認できた。

第 4 章では、SVD と ICA を組み合わせて、ICA を適用する前にドキュメントベクトルを低次元化し、ICA による潜在的意味抽出の処理時間を短縮する手法を提案した。ここでは、SVD によって得られる特異ベクトルを累積寄与率により選択し、その選択された特異ベクトルが張る空間上で ICA を行なった。また、低次元化によりノイズを低減し、ユーザプロファイルの精度改善も試みた。累積寄与率を変えることによるユーザプロファイルの精度の変化や、本手法によって得られる潜在的意味について検討を行なった。その結果、ICA の処理時間を短縮しつつ、ICA だけによる処理とほぼ同程度か、それ以上の精度を持つユーザプロファイルの作成に成功した。

第 5 章では、ICA を行列分解の一手法と見なし、混合行列を潜在的意味として用いてドキュメントベクトルを再構成し、ユーザプロファイルの精度改善を試みた。ここでは、NTCIR2 を用いた実験において、従来よりもより直感に近いトピックを得ると共に、そのトピック空間上でユーザプロファイルを作成することで、フィルタリングの精度改善が実現できることを確認した。

第 6 章では、これまでのドキュメントを ICA によって得られる基底で表現するのではなく、そのトピックを用いて索引語を選別することによって、ユーザプロファイルの精度を改善することを試みた。ここでは、トピックをもっとも表わすと考えられる単語と、頻繁に共起する単語を、そのトピックの重要語と位置づけ、それらを索引語として選別した。また、その選別した索引語を用いてユーザプロファイルを作成し、フィルタリングの精度改善に効

果があることを確認した。

以上のように、本論文では、ICA をドキュメントに対して適用して得られる潜在的意味を大きく分けて、“潜在的意味によりドキュメントを表現する方法”と “トピックを用いて索引語そのものを選別する方法”という 2 つのアプローチによりユーザプロファイルによる情報フィルタリングに用いることを提案した。また、その上でフィルタリングの精度を改善するための手法を検討し、それらに対して NTCIR2 を用いた評価実験を行なうことで有効性を確認した。

今後はより大規模な実験データに対しての適用が課題と考えられる。第 4 章でも述べたように、ICA の処理は大変コストのかかる処理であるため、このままでは、Web データのような大規模なデータに、直接適用することは困難であると考えられる。したがって、それらの問題に対処する必要が、今後この研究を実社会で役立てていくために必要である。

謝辞

本論文をまとめるにあたり、終始丁寧なるご指導とご助言を賜りました大阪府立大学大学院工学研究科電気・情報系専攻知能情報工学分野 大松 繁 教授に深く感謝するとともに厚く御礼申し上げます。同教授には公私ともに多大なご配慮を頂きました。同分野の 黄瀬 浩一 教授、 石渕 久夫 教授には、本研究の遂行に対する配慮のみならず、副査として本論文を審査して頂き、有意義なご助言を賜り深く感謝致します。また、本研究の遂行に関して、有益なご助言とご配慮を賜りました福永 邦雄 教授、汐崎 陽 教授、辻 洋 教授、松本 啓之亮 教授、市橋 秀友 教授に感謝致します。

筆者が研究室に配属された頃より、本研究の遂行に際して、終始暖かいご指導とご助言を賜りました柳本 豪一助手に心より感謝致します。同助手にも公私ともにさまざまな場面においてお世話になりました。また、藤中 透 助教授、吉岡 理文 助教授には日常より本研究に対する有意義なご指摘を頂き感謝致します。

さらに、筆者が所属する研究室において、伊藤 征嗣博士（現広島工業大学講師）を始めとする皆様方には、本研究を進めるにあたり、惜しみない協力を頂きありがとうございました。

最後に、筆者の学位取得を応援し続けてくれた父と母に深い感謝の意を表して、謝辞と致します。

平成 19 年 1 月

参考文献

[1] 森田 昌弘, 速水 治夫: 情報フィルタリングシステム; 情報洪水への処方箋, 情報処理学会学会誌, Vol. 37, No. 8, pp. 751–758(1996).

[2] S. Loeb and D. Terry: Information Filtering, *Communications of the ACM*, Vol. 35, No. 12, pp. 26–28(1992).

[3] P. Resnick and H. Varian: Recommender Systems, *Communications of the ACM*, Vol. 40, No. 3, pp. 56–58(1997).

[4] M. Balabanovic and Y. Shoham: Fab; Content-based Collaborative Recommendation, *Communications of the ACM*, Vol. 40, No. 3, pp. 66–72(1997).

[5] T. Malone, K. Grant, F. Turbak, S. Brosbst and M. Cohen: Intelligent Information-sharing Systems, *Communications of the ACM*, Vol. 30, No. 5, pp. 390–402(1987).

[6] T. Yan and H. Garia-Molina: The SIFT Information Dissemination System, *ACM Transaction on Database Systems*, Vol. 24, No. 4, pp. 529–565(1999).

[7] B. Sheth and P. Maes: Evolving Agents for Personalized Information Filtering, *Proceedings of IEEE Conference on Artificial Intelligence for Applications*, pp. 342–352, Orlando, USA(1993).

[8] M. Pazzani and D. Billsus: Learning and Revising User Profiles; The Identification of Interesting Web Sites, *Machine Learning*, Vol. 27, pp. 313–331(1997).

[9] P. Kantor, E. Boros, B. Melamed, V. Mekov, B. Shapira and D. Neu: Capturing Human Intelligence in the Net, *Communications of the ACM*, Vol. 43, No. 8, pp. 112–115(2000).

[10] G.Salton and M. McGill: *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company(1983).

[11] S. Deerwest, T. Dumais, T. Landauer, W. Furnas, A. Harshman.A: Indexing by Latent Semantic Analysis, *Journal of the Society for Information Science*, Vol. 41, No. 6, pp. 391–497(1996).

[12] A.Hyvārinen and E.Oja: Independent Component Analysis; A Tutorial, *Neural Network*, Vol. 13, No. 6, pp. 411–430(2000).

[13] 村田 昇: 入門 独立成分分析, 東京電機大学出版局 (2004).

[14] T. Kolenda, L. Hansen, and S. Sigurdsson: Indepedent Components in Text, *Advances in Independent Component Analysis*, pp. 229–250, Springer-Verlag(2000).

[15] T. Kolenda, L. Hansen and J. Larsen: Signal Detection Using ICA; Application to Chat Room Topic Spotting, *Third International Conference on Independent Component Analysis and Blind Source Separation*, pp. 540–545(2001).

[16] Y. H. Kim and B. T. Zhang: Document Indexing Using Independent Topic Extraction, *Proceedings of the International Conference on Independent Component Analysis and Signal Separation(ICA2001)*, pp. 557–562(2001).

[17] A. Kabán, E. Bingham and M. Girolami: Topic Identification in Dynamical Text by Complexity Pursuit, *Neural Processing Letters*, Vol. 17, pp. 69–83, Springer(2003).

[18] 濱本 雅史, 北川 博之, Jia-Yu Pan, Christos Faloutsos: 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).

[19] R. Reinhard: Mining Text for Word Senses Using Independent Component Analysis, *Pro-*

ceedings of the 4th SIAM International Conference on Data Mining (SIAM DM'04)(2004).

[20] http://research.nii.ac.jp/ntcir/index-en.html.

[21] T. Yokoi, H. Yanagimoto and S. Omatu: The Proposal for the Way to Recommend Information with ICA, *Proceedings of the Ninth International Symposium on Artificial Life and Robotics*, pp. 694–697, Oita, Japan(2004).

[22] T. Yokoi, H. Yanagimoto and S. Omatu: Information Recomendation using ICA, *Journal of Artificial Life and Robotics*, Vol. 9, No. 3, pp.103–105(2005).

[23] T. Yokoi, H. Yanagimoto and S. Omatu: Improvement of Information Filtering using Topic Selection, *Proceedings of the Tenth International Symposium on Artificial Life and Robotics*, pp. 83–86, Oita, Japan(2005).

[24] 横井 健, 柳本 豪一, 大松 繁: 独立成分の選択による情報推薦の改良, 電気学会論文誌 C, Vol. 126, No. 4, pp. 492–497(2006).

[25] T. Yokoi, H. Yanagimoto and S. Omatu: Information Filtering using SVD and ICA, *Proceedings of the Tenth International Symposium on Artificial Life and Robotics*, pp. 79–82, Oita, Japan(2005).

[26] 横井 健, 柳本 豪一, 大松 繁: 潜在的意味を用いた情報フィルタリング, 電気学会論文誌 C, Vol. 126, No. 7, pp. 865–870(2006).

[27] T. Yokoi, H. Yanagimoto and S. Omatu: Information filtering using SVD and ICA, *Journal of Artificial Life and Robotics*, Vol. 10, No. 2, pp.116–119(2006).

[28] T. Yokoi, H. Yanagimoto and S. Omatu: Index Words Selection with ICA, *Proceedings of the 2006 IEEE International Conference on Systems, Man and Cybernetics*, pp. 3348–3353, Taipei, Taiwan(2006).

[29] 横井 健, 柳本 豪一, 大松 繁: ICA による索引語抽出を用いた情報フィルタリング, 電気学会論文誌 C（投稿中）.

[30] K. Kageura and B. Umino: Methods of Automatic Term Recognition, *Terminology*, Vol. 3, No. 2, pp. 259–289(1996).

[31] T. Noreault, M. McGill and M. Koll: *A Performance Evaluation of Similarity Measure, Document Term Weighting Schemes and Representations in a Boolean Environment*, Butterworths, London(1997).

[32] 徳永 健伸: 情報検索と言語処理, 東京大学出版 (1999).

[33] 高村 大也, 松本 裕治: 独立成分分析を用いた文書分類; SVM のための素性空間再構成, IPSJ SIG notes(2001).

[34] U. Shardanand and P. Maes: Social Information Filtering: Algorithms for Automating ”World of Mouth”, *Proceedings of ACM CHI’95 Conference on Human Factors in Computing Systems*, Vol. 1, pp. 210–217, Colorado, USA(1995).

[35] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm and J. Riedl: Grouplens; An Open Architecture for Collaborative Filtering of Netnews, *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pp. 175–186, North Carolina, USA(1994).

[36] http://www.amazon.com.

[37] 長尾 真: 自然言語処理, 第 11 章, 岩波書店 (1996).

[38] 伊庭 斉志: 遺伝的アルゴリズムの基礎; GA の謎を解く, オーム社 (1994).

[39] 柳本 豪一, 藤中 透, 吉岡 理文, 大松 繁: 情報フィルタリングにおける遺伝的アルゴリズムを用いたユーザプロファイルの作成手法, 電気学会論文誌 C, Vol. 121-C, No. 7, pp. 1277–1282(2001).

[40] P. Baclace: Competitive Agents for Information Filtering, *Communications of the ACM*, Vol. 35, No. 12, p. 50(1992).

[41] 北 研二, 津田 和彦, 獅々堀 正幹: 情報検索アルゴリズム, 共立出版 (2002).

[42] A. Hyvārinen and E. Oja: A Fast Fixed-point Algorithm for Independent Component

Analysis, *Neural Computation*, Vol. 9, No. 7, pp. 1483–1492(1997).

[43] Y.Matsumoto: Japanese Morphological Analysis System; Chasen, Technical Report, Nara Institute of Science Technology, Information Science Technical Report NAIST-IS-TR97007(1997).

[44] 森 辰則: 検索結果表示向け文書要約における情報利得比に基づく語の重要度計算, 自然言語処理, Vol. 9, No. 4, pp. 3–32(2002).

[45] J. Lin: Divergence Measures Based on the Shannon Entropy, *IEEE Transaction on Information Theory*, Vol. 37, No. 1, pp.145–151(1991).

[46] L.Lee: On the Effectiveness of the Skew Divergence for Statistical Language Analysis, *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics*, pp. 65–72(2001).

[47] 長尾 真: パターン情報処理, 電子通信学会大学シリーズ, コロナ社 (1983).

[48] J. Tou and R. Gonzalez: *Pattern Recognition Principles*, Addison-Wesley(1974).

[49] S. Kullback and R. Leiber: On Information and Sufficiency, *Annals of Mathematical Statistics*, Vol. 22, pp. 79–86(1951).

[50] 鳥脇 純一郎: 工学のための確率論, オーム社 (2002)

[51] 松尾 豊, 石塚 満: 語の共起の統計情報に基づくドキュメントからのキーワード抽出アルゴリズム, 人工知能学会誌, Vol. 17, No. 3, pp. 213–227(2002).

[52] 大澤 幸生, N.Benson, 谷内田 正彦: Keygraph; 語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会論文誌, J82-D- , No. 2, pp. 391–400(1999).