非適合プロファイルを利用した文書フィルタリング手法

帆足 啓一郎[†] 松本 一則[†]
井ノ上 直己[†] 橋本 和夫[†]

テキスト文書の流れの中からユーザの要求（プロファイル）を満たした文書を選択する文書フィルタリングのシステムでは，多くの場合，プロファイルと検索対象文書との類似度を計算し，その類似度が閾値を超えた文書を選択する手法がとられている．しかし，このような類似度に基づいた手法では，閾値を高く設定した場合多くの適合文書が見逃されてしまい，また，逆に閾値を低く設定した場合は多くの非適合文書が誤って選択されてしまうなど，十分なフィルタリング精度が得られていないのが現状である．そこで本論文では従来のプロファイルに加え，非適合文書から抽出された情報に基づいた非適合プロファイルを利用する新たなフィルタリング手法を提案する．TREC データに対する評価実験の結果，提案手法の適用によって誤って選択される非適合文書の数が減り，フィルタリング精度に向上がみられた．

Document Filtering Method Using Non-relevant Information Profile

Keiichiro Hoashi,[†] Kazunori Matsumoto,[†] Naomi Inoue[†] and Kazuo Hashimoto[†]

Document filtering is a task to retrieve documents relevant to a user's profile from a flow of documents. Generally, filtering systems calculate the similarity between the profile and each incoming document, and retrieve documents with similarity higher than a threshold. However, many systems set a relatively high threshold to reduce retrieval of non-relevant documents, which results in the ignorance of many relevant documents. In this paper, we propose the use of a non-relevant information profile to reduce the mistaken retrieval of non-relevant documents. Results from experiments show that this filter has successfully rejected a sufficient number of non-relevant documents, resulting in an improvement of filtering performance.

1. はじめに

文書フィルタリングとは，継続的に流れてくる文書の中からユーザの要求と合致した文書を抽出し，ユーザに出力するタスクである．個々の文書がユーザの要求に合致しているかどうかの判断には，テキスト自動分類技術を適用しているシステム[1)]も発表されてはいるものの，多くのシステムではユーザの要求を**プロファイル**としてシステム内に表し，そのプロファイルと流れてくる個々の文書との類似度を算出し，類似度が高い文書をユーザに出力する手法が取り入れられている．

以上の処理からも明らかなように，文書フィルタリングと情報検索のタスクは類似しているため，文書フィルタリングでは情報検索分野で開発された技術が適用されることが多い．たとえば，プロファイルやフィルタリングの対象となる個々の文書はベクトル空間モデルによって表すことができ，プロファイルと文書との類似度は両者のベクトルのコサイン値を算出することにより求められる．また，情報検索では適合フィードバックを利用して検索式の情報を拡張する検索式拡張（query expansion）が広く利用されているが，これに対し，文書フィルタリングでは，文書の流れの中から抽出された文書に対する適合フィードバックに基づいてプロファイルを更新することにより，それ以降の文書に対するフィルタリング精度向上を図るプロファイル更新という技術が利用されている．

しかし，文書フィルタリングと情報検索の間にはいくつかの相違点もある．情報検索では，与えられた要求に対する検索対象はあらかじめ蓄積されている静的な文書集合である．一方，文書フィルタリングではあ

† 株式会社 KDD 研究所
KDD R&D Laboratories, Inc.

らかじめユーザからの要求が与えられ，その検索対象は次々に到着する文書の流れである．したがって，検索対象が動的に変化する．

また，情報検索では要求にどれだけ類似しているかをランク付けした文書集合を検索結果として出力するのに対し，文書フィルタリングでは1つ1つの文書が要求に合っているか否かの判断のみを行う．検索対象とする文書は時間ごとに到着し，それらを保持することはできないため，情報検索のように検索対象文書集合全体における各文書のランク付けを行うことは不可能である．したがって，プロファイルとの類似度を基準に文書フィルタリングを行う場合，類似度の閾値の設定が重要であることは明らかである．閾値を低く設定した場合，多くの適合文書を選択することができるが，同時に誤って選択される非適合文書も増加する．逆に，閾値を高く設定すれば非適合文書の誤り選択は減少するが，見逃される適合文書も多くなってしまう．近年，TREC[2]の Filtering Track[3]などで数々の文書フィルタリングに関する研究が発表されているが，フィルタリングの進行にともなって閾値を上げることにより，誤り選択の減少を図るシステムが多い．

本論文では，文書の流れの中からより多くの適合文書を選択しつつ，非適合文書の選択を減少させるため，ユーザの要求を表すプロファイルに加え，過去に誤って選択された文書の特徴を表す非適合文書プロファイルを利用した新たな文書フィルタリング手法を提案する．まず，既存のプロファイル更新手法による従来手法の評価実験を行い，従来手法の問題点を明らかにする．次に提案手法について説明し，TREC データに基づく評価実験とその結果を示し，提案手法の有効性を実証する．

2. 従来手法

文書フィルタリングのプロファイル更新には情報検索における検索式拡張手法を適用する手法が多く用いられている．ここでは，一般に広く利用されている Rocchio のアルゴリズムに基づいたプロファイル更新手法と，単語寄与度を利用した検索式拡張手法に基づいたプロファイル更新手法のそれぞれについて説明し，これらの手法の評価実験を行う．

2.1 Rocchio のアルゴリズムに基づくプロファイル更新手法

現在，最も有効な検索式拡張手法の 1 つとして，Rocchio のアルゴリズムに基づく手法があげられる[4]．Rocchio のアルゴリズムはベクトル空間モデルを前提とした語の重み付け手法であり，検索式のベクトルを適合文書のベクトルに近づけつつ非適合文書から遠ざけることを目的とする[5]．拡張前の検索式と拡張後の検索式をそれぞれ Q_{org}，Q_{new} で表すとすると，この手法は式 (1) で表される．

$$\vec{Q}_{new} = \alpha \times \vec{Q}_{org} + \beta \times \frac{1}{R} \sum_{D \in Rel} \vec{D} - \gamma \times \frac{1}{N} \sum_{D \in nRel} \vec{D} \quad (1)$$

ただし，R，N はそれぞれ適合文書ならびに非適合文書の数を表し，α，β，γ は任意の係数である．本手法を用いた検索式拡張は，上記ベクトル変換の結果，元の検索式に含まれない語のうち重みの値が高い語を抽出し，その重みとともに元の検索式のベクトルに加えるという手法で実現される．TREC-7 では $\alpha = 3$，$\beta = 2$，$\gamma = 2$ の係数値で上記検索式拡張手法を採用した検索システム “SMART” が発表され，高い検索精度が得られている[6]．

このアルゴリズムに基づいた検索式拡張手法をプロファイル更新に適用したシステムとして，“CAFES” が TREC-8 において報告されている[7]．CAFES では，式 (1) で $\alpha = 1$，$\beta = 0.1$，$\gamma = 0$ とし，適合文書の情報のみを利用してプロファイル更新を行っている．

本手法ではユーザからのフィードバック情報があらかじめ決めた n 件蓄積されるごとにプロファイルを更新している．n 件の文書に対する適合フィードバックを得るごとに，式 (1) により，その間の選択した文書中の単語の重みを求める．そして，重みを算出した単語のうち重みの大きい単語上位 20 個をプロファイル更新に利用する．

以上の処理により抽出された単語とそのスコアを，元のプロファイルに加えることによりプロファイルを更新する．

2.2 単語寄与度に基づくプロファイル更新手法

ここでは情報検索においてその有効性が確認されている，単語寄与度に基づいた検索式拡張手法をプロファイル更新に適用した手法について述べる．

2.2.1 単語寄与度に基づく検索式拡張手法

単語寄与度とは，文書間の類似度における各単語の影響を数値化した尺度である．ユーザの要求を検索式 q で表すとすると，ある検索式 q と検索対象文書 d との間の類似度における単語 w_i の単語寄与度を式 (2) によって定義する[8]．

$$Cont(w_i, q, d) = Sim(q, d) - Sim(q'(w_i), d'(w_i)) \quad (2)$$

ただし，$Sim(q, d)$ は q，d 間の類似度を表し，$q'(w_i)$ は q から単語 w_i を除いた検索式を，$d'(w_i)$ は d か

ら単語 w_i を除いた文書を表すとする．すなわち，単語寄与度 $Cont(w_i, q, d)$ とは，q と d との類似度と単語 w_i が存在しない場合の q と d との類似度との差である．したがって，q と d に出現するすべての単語のうち，類似度を向上させる単語の寄与度は正であり，逆に類似度を下げる単語の寄与度は負である．

参考文献 8) によれば，出現単語の多くの寄与度は 0 に近く，類似度に有意な影響を与えている単語は少ない．そのうち，寄与度が大きく正の値を持つ単語は，検索式と検索対象文書の両方に存在する単語である．一方，大きく負の値の寄与度を持つ単語は一方の文書にのみ存在し，かつ，その文書の特徴を顕著に表す単語であると考えられる．そこで，単語寄与度に基づいた検索式拡張手法では以下のように検索式の拡張を行っている．

まず，検索式 q と適合している文書群 $D_{rel}(q)$ 中の各文書に出現するすべての単語の寄与度を求め，各適合文書から単語寄与度の低い単語を N 個抽出する．次に抽出された各単語の寄与度の総和に重み wgt をかけ，これを単語 w_i に対するスコアとする．単語 w_i の q と文書 d の類似度に対する寄与度を $Cont(w_i, q, d)$ とすると，単語 w_i のスコア $Score(w_i)$ は式 (3) によって表される．

$$Score(w_i) = wgt \times \sum_{d \in D_{rel}(q)} Cont(w_i, q, d) \quad (3)$$

次に，抽出された単語のうち元の検索式に含まれていない単語を検索式に加えることで検索式拡張を実現する．ある単語 w_i を検索式のベクトルに加える際には，式 (3) で計算されたスコア $Score(w_i)$ を単語 w_i が検索式に出現する頻度 tf_i と見なし，検索式のベクトル内で単語 w_i を表す要素の値を計算する．ベクトルの各要素が TF∗IDF によって計算されている場合，$Score(w_i)$ に単語 w_i の IDF をかけ，その結果得られた TF∗IDF 値を検索式のベクトルの単語 w_i を表す要素に入れることにより，検索式拡張を行う．

以上の検索式拡張手法を用いた検索実験において，Rocchio のアルゴリズムに基づく検索式拡張を用いる手法を上回る高い精度が得られている．

2.2.2 単語寄与度に基づくプロファイル更新

単語寄与度による検索式拡張では，初期検索の結果に対するフィードバックにより得られた適合文書集合中の各文書から寄与度に基づいて抽出された単語の寄与度の総和に重みを掛けることで，各単語に対するスコアを求めた．ここでは，フィルタリング中に選択された文書 1 つごとに単語寄与度に基づき文書中から単語を抽出し，その情報を直前のプロファイルに加えることで，随時プロファイルを更新する[9]．

まず，選択された文書が適合文書の場合には，その文書から抽出された単語 w_i に対するスコア $Score_{rel}(w_i)$ を式 (4) により算出し，選択した文書が非適合文書中の場合には，抽出された単語 w_i のスコア $Score_{nrel}(w_i)$ を式 (5) により算出する．負の単語寄与度を持つ単語を使用するため，パラメータ wgt_{rel_R}，wgt_{nrel_R} は負の値を持った重みである．

$$Score_{rel}(w_i) = wgt_{rel_R} \times Cont(w_i, q, d) \quad (4)$$

$$Score_{nrel}(w_i) = wgt_{nrel_R} \times Cont(w_i, q, d) \quad (5)$$

上記の式によって求めた各単語のスコアを単語 w_i が出現する頻度 tf_i として扱い，TF∗IDF 法により各単語の重みを算出する．そして，抽出された単語が適合文書に含まれていた単語の場合はその単語と重みを元のプロファイルに加え，非適合文書に含まれていた単語の場合は単語と重みを元のプロファイルから引く．なお，この処理により負の重みを持った単語は，類似度計算に使用されない．

以上の処理により，プロファイルを表すベクトルの，値を持たなかった次元が正の値を持つようになり，プロファイルの情報が拡張されることになる．また，適合文書と非適合文書に共起する単語の重みが抑制され，適合文書のみに出現する単語の重みが強調される．

2.3 評価実験

上記の 2 つの検索式拡張手法に基づいたプロファイル更新手法の評価を行うために以下の実験を行った．

2.3.1 タスク

本論文での評価実験はすべて TREC Filtering Track で用意されているタスクに基づいて行われている．TREC Filtering Track では，文書フィルタリングをシミュレートするため，プロファイルの元となる入力文として ad hoc task で使用された topic を利用し，topic 同様 ad hoc task で使用された文書データを検索対象文書として時間順にフィルタリングシステムに 1 つずつ入力する．そして，各 topic に合致する文書の一覧（以下，正解文書データ）を利用し，ユーザからの適合フィードバックをシミュレートする．

具体的な処理の流れは以下のとおりである．まず，システムは，入力される検索対象文書が topic に適合しているか否かの判断を行う．ここでは，topic から生成されるプロファイルと検索対象文書との類似度を算出し，その類似度が閾値を超えた場合のみ，検索対象文書が適合していると判断し，出力する．システム

から出力された文書を topic の正解文書データと照合することにより，出力文書の適合性を判断する．出力された文書が正解文書データに含まれている場合，その文書は適合であり，逆に正解文書データに出力文書が含まれていない場合はその文書は非適合であるとする．この照合結果はシステムにフィードバックされ，プロファイル更新の際に利用される．そして，更新されたプロファイルを利用し，以降入力される検索対象文書に対するフィルタリングを行う．

TREC Filtering Track では，システムに入力されていない検索対象文書に含まれる情報を利用することはできない．たとえば，未入力の検索対象文書に含まれている単語情報を利用して出現単語の df 情報などを作成することは禁止されている．しかし，システムに入力された文書に含まれる情報は利用することが可能である．

なお，評価実験で使用した評価データの詳細については後述する．

2.3.2 使用システム

本論文では検索対象文書とプロファイルの両方をベクトル空間モデルを用いて表し，両者間の類似度を計算することで各文書に対するフィルタリングを実現した．

ベクトル空間モデルで表現する際，各文書をおよびプロファイル表すベクトルの各要素の重みは TF*IDF 法により算出する．本実験では，最も有効な情報検索システムの 1 つである SMART において使用されているアルゴリズムに基づき TF および IDF の計算式を使用した．本実験で使用した TF および IDF の計算式を，式 (6)，式 (7) に示す[6]．

- TF factor
$$\log(1+tf_i) \tag{6}$$
- IDF factor
$$\log\left(\frac{M}{df_i}\right) \tag{7}$$

ただし，tf_i は文書 d 内の単語 w_i の出現頻度，df_i は単語 w_i が出現する文書数，M は df のリストの作成に使用した文書数とする．TF の計算の際，tf_i に 1 を加えた値を使用しているが，これは単語寄与度によるプロファイル更新の際に tf_i が 1 未満になる（すなわち，$\log(tf_i)$ が負になる）単語に対処するためである．

また，類似度は式 (8) で定義されるプロファイル q と検索対象文書 d のベクトルのコサイン値を算出することにより求めた[10]．

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}||\vec{d}|} \tag{8}$$

表 1 TREC-7 評価データ詳細
Table 1 Details of TREC-7 experiment data.

データ種類	使用データ	件数
入力文	Topics 1-50	50
検索対象	*Associated Press* (1988〜1990)	242918
初期 *df* 作成	*Federal Register*, *Foreign Broadcast Information Service*, *LA Times*, *Financial Times*	528150

表 2 TREC-8 評価データ詳細
Table 2 Details of TREC-8 experiment data.

データ種類	使用データ	件数
入力文	Topics 351-400	50
検索対象	*Financial Times* (1992〜1994)	204790
初期 *df* 作成	*Federal Register*, *Foreign Broadcast Information Service*, *LA Times*	317993

ただし，$\vec{q}$, $\vec{d}$ はそれぞれ入力文書と検索対象文書を表すベクトルとし，$|\vec{q}|$ は $\vec{q}$ のユークリッド長とする．

2.3.3 評価データ

本論文の評価実験では，TREC-7 および TREC-8 の Filtering Track の評価データを使用する．各評価データには，入力文として TREC の Topic 50 件が用意されている．また，検索対象文書としては，TREC-7 では *Associated Press*の 3 年分（1988〜1990）の記事，TREC-8 では *Financial Times*の 3 年分（1992〜1994）の記事をそれぞれ利用した．

また，2.3.1 項で述べたとおり，システムに未入力の検索対象文書からは各単語の *df* 情報を求めることはできないため，各評価データごとに，検索対象文書以外の文書データに基づき，各単語の初期 *df* 情報を作成する．さらに，フィルタリングが 10000 件進むごとにフィルタリング済みの文書 10000 件の情報を追加して *df* 情報の更新を行う．

TREC-7 および TREC-8 の評価データの詳細をそれぞれ**表 1** および**表 2** に記述する．

2.3.4 単語寄与度によるプロファイル更新

以下，単語寄与度に基づくプロファイル更新の具体的な方法について説明する．

式 (4)，(5) によって求めた各単語 w_i のスコアを tf_i として扱い，式 (6)，(7) により TF および IDF を計算する．適合文書中の単語の TF*IDF 値を $Value_{rel}(w_i)$，非適合文書中の単語の TF*IDF 値を $Value_{nrel}(w_i)$ とすると，各単語の TF*IDF 値は式 (9)，(10) により算出される．

$$Value_{rel}(w_i) = \log(1 + Score_{rel}(w_i)) \times \log \frac{M}{df_i} \quad (9)$$

$$Value_{nrel}(w_i) = \log(1 + Score_{nrel}(w_i)) \times \log \frac{M}{df_i} \quad (10)$$

ただし，df_i は単語 w_i が出現する文書数，M は df データの作成に使用された文書数とする．

プロファイル p を式 (11) のように表すとする．

$$\vec{p} = (p_1, p_2, \ldots, p_k) \quad (11)$$

ただし，$p_1, \ldots, p_k$ はプロファイル中の単語の重みを表し，k はベクトルの次元数を表す．

更新後のプロファイルを式 (12) で表すとすると，抽出された各単語 w_i について，適合文書中の単語の場合は式 (13) により表され，非適合文書中の単語の場合は式 (14) により表される．

$$\vec{p_{new}} = (p'_1, p'_2, \ldots, p'_k) \quad (12)$$

$$p'_i = p_i + Value_{rel}(w_i) \quad (13)$$

$$p'_i = p_i - Value_{nrel}(w_i) \quad (14)$$

すなわち，適合文書から選択された各単語の要素を元のプロファイルの要素に加え，非適合文書から選択された各単語の要素を元のプロファイルの要素から引くという処理を行う．なお，この処理により負の重みを持った単語は，類似度計算に使用されない．

2.3.5 評価基準

情報検索では一般的に，選択された文書中の適合文書の割合を表す適合率（precision），全適合文書中の選択された適合文書の割合を表す再現率（recall）などによって評価が行われている．しかし，上記評価基準を用いるためにはシステムの出力が順位付けされた文書集合である必要があり，文書フィルタリングのように検索対象文書 1 つ 1 つに対しての判断を返すタスクの評価に使用することはできない．また，適合文書が存在しないプロファイルに関しては再現率を算出することは不可能である．たとえば，このようなプロファイルに対して 1 件の文書を選択するシステムと 1000 件の文書を選択するシステムの Precision はともに 0 となるが，前者の方が精度の高いシステムであることは明らかである．

このように，情報検索において使用される評価基準を文書フィルタリングの評価に使用することは妥当ではないと考えられるため，文書フィルタリングの評価基準としては，式 (15) に示す utility[3] が使用される．

$$u(S,p) = A \times R_+ + B \times N_+ + C \times R_- + D \times N_- \quad (15)$$

ただし，

- $u(S,T)$：システム S におけるプロファイル p の utility
- R_+：選択された適合文書数
- R_-：選択されなかった適合文書数
- N_+：選択された非適合文書数
- N_-：選択されなかった非適合文書数

とする．本論文の評価には TREC-8 Filtering Track において使用されたパラメータ設定である $A = 3$, $B = -2$, $C = D = 0$ を使用した．

一方で，各プロファイルにより適合文書の数が違うことから，utility の理論的な上限は異なってくる．この点を考慮した評価基準として，utility を正規化した scaled utility[3] があげられる．式 (16) に scaled utility の計算式を示す．

$$u^*_s(S,p) = \frac{\max(u(S,p), U(s)) - U(s)}{MaxU(p) - U(s)} \quad (16)$$

ただし，

- $u^*_s(S,p)$：システム S におけるプロファイル p の scaled utility
- $u(S,p)$：システム S におけるプロファイル p の utility
- $U(s)$：s 個の非適合文書のみを選択した場合の utility
- $MaxU(p)$：プロファイル p の utility の理論的最大値

とする．

scaled utility の算出の際は，$U(s)$ 以下の utility はすべて $U(s)$ とされる．そのため，すべての utility が $U(s)$ から $MaxU(p)$ の間の値となり，0 から 1 の間で正規化された値が得られる．

2.3.6 実　　験

ここでは，従来手法の評価のため，TREC-7 ならびに TREC-8 の評価データを利用し，Rocchio および単語寄与度を利用したプロファイル更新手法の評価実験を行った．

Rocchio のアルゴリズムに基づいたプロファイル更新手法では，TREC-8 において報告されている実験[7] で使用されたパラメータ，$\alpha = 1$, $\beta = 0.1$, $\gamma = 0$, $n = 2$ を用いて実験を行った．文書フィルタリングにおいては Rocchio の手法で $\gamma \neq 0$ とした実験も報告されているが，$\gamma = 0$ と比較した場合の優位性に関する報告はない．Rocchio のパラメータの最適化は本実験の目的ではないため，ここでは参考文献 7) で報告されている値をそのまま使用した．

図 1 に，類似度の閾値を 0.1 とした場合のある 1 つのプロファイルに対する選択文書の類似度を，適合文

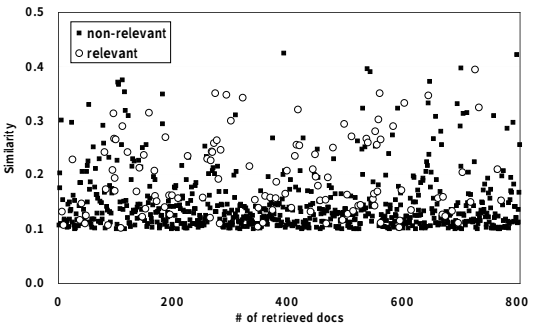

図 1 選択された文書の類似度（Rocchio）
Fig. 1 Similarity of selected documents (Rocchio).

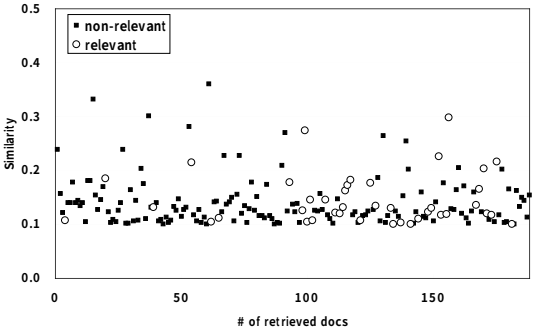

図 2 選択された文書の類似度（単語寄与度）
Fig. 2 Similarity of selected documents (WC).

書・非適合文書に区別して示す．

単語寄与度に基づいたプロファイル更新手法では，$wgt_{rel_R} = \{-200, -400, -800\}$，$wgt_{nrel_R} = \{-100, -200, -400, -800\}$ と変化させて実験を行った．**図 2** に，$wgt_{rel_R} = -200$，$wgt_{nrel_R} = -800$ としたときのある 1 つのプロファイルに対する選択文書の類似度を，適合文書・非適合文書に区別して示す．また，単語寄与度に基づいた手法において各 wgt_{rel_R}，wgt_{nrel_R} でのすべてのプロファイルの平均 scaled utility，および Rocchio のアルゴリズムでの平均 scaled utility を，TREC-7 および TREC-8 についてそれぞれ**表 3**，**表 4** に示す．

図 1，図 2 より，閾値を大きく超える類似度を持った非適合文書は少ないものの，閾値近辺の類似度では適合文書と非適合文書が混在している様子が分かる．したがって，閾値を下げることで多くの適合文書を選択しようとすると非適合文書も多く選択されてしまうため，多くの適合文書を選択することが困難であることが分かる．逆に，閾値を大きく設定することで，誤って選択される非適合文書数を減少させることは可能だが，その場合取得される適合文書数が減少する．

このことから，検索式拡張に使用される手法を適用して更新されたプロファイルとの類似度のみによるフィルタリングを行う手法では高い精度が得られないという問題点が明らかになった．したがって，高精度のフィルタリングを行うためには，単純にプロファイルとの類似度のみに基づいたフィルタリングを行うのではなく，新たな手法を用いて誤って選択される非適合文書を減少させる必要がある．

表 3 従来手法の平均 scaled utility（TREC-7）
Table 3 Average scaled utility of existing methods (TREC-7).

	wgt_{nrel_R}			
wgt_{rel_R}	-100	-200	-400	-800
-200	0.4053	0.4406	0.4669	0.4820
-400	0.3791	0.4122	0.4518	0.4720
-800	0.3478	0.3894	0.4229	0.4525
Rocchio	0.3923			

表 4 従来手法の平均 scaled utility（TREC-8）
Table 4 Average scaled utility of existing methods (TREC-8).

	wgt_{nrel_R}			
wgt_{rel_R}	-100	-200	-400	-800
-200	0.4558	0.4840	0.5091	0.5257
-400	0.4172	0.4777	0.5107	0.5184
-800	0.3815	0.4349	0.4842	0.5100
Rocchio	0.3730			

3. 非適合プロファイルを利用した文書フィルタリング

本章では，従来の適合文書を表すプロファイルとは逆に，誤って選択された非適合文書の特徴を表す **“非適合プロファイル”** を作成し，非適合プロファイルとの類似度が高い文書は選択しないという手法を提案する．

なお，ユーザの要求を表すプロファイルの更新は前章の単語寄与度に基づいたプロファイル更新手法を用い，非適合プロファイルの作成においても単語寄与度に基づくアルゴリズムを利用する．

3.1 提 案 手 法

従来のプロファイル（以下，p_R）は，ユーザの要求を表すよう，適合文書の特徴を取り入れて更新してきた．しかし，前章の実験から，p_R に類似している文書を選択するだけでは，非適合文書も多く選択されてしまうということが明らかになった．したがって，プロファイルとの類似度が高いが実際には要求に適合していない非適合文書を選択しないようにすれば，精度を向上させることができると考えられる．

そこで，過去に p_R との類似度が高いために誤って選択された非適合文書の特徴を表すプロファイルを非

適合プロファイル p_N として作成する．この p_N との類似度が高い文書は，過去に誤って選択された非適合文書に類似しているものと考えられる．したがって，このような文書を選択しないようにすれば，これまで誤って選択されていた非適合文書を除外することができると期待される．

以下，本手法の具体的説明を行う．

まず，2.2.2 項で述べた p_R の更新同様，選択された文書から単語寄与度に基づき単語を抽出する．そして，抽出された単語が適合文書中の単語の場合には，単語に対するスコア $Score_{rel_N}(w_i)$ を式 (17) により算出し，非適合文書中の単語の場合には $Score_{nrel_N}(w_i)$ を式 (18) によって算出する．ここで，wgt_{rel_N}，wgt_{nrel_N} は負の値を持った重みである．

$$Score_{rel_N}(w_i) = wgt_{rel_N} \times Cont(w_i, p, d) \quad (17)$$

$$Score_{nrel_N}(w_i) = wgt_{nrel_N} \times Cont(w_i, p, d) \quad (18)$$

次に，上記の式によって求めた各単語 w_i のスコアを単語頻度 tf_i として扱い，以下の式 (19)，(20) により各単語 w_i の TF∗IDF 値を算出する．

$$Value_{rel_N}(w_i) = \log(1 + Score_{rel_N}(w_i)) \times \log \frac{M}{df_i} \quad (19)$$

$$Value_{nrel_N}(w_i) = \log(1 + Score_{nrel_N}(w_i)) \times \log \frac{M}{df_i} \quad (20)$$

ただし，df_i は単語 w_i が出現する文書数，M は df のリスト作成に使用された文書数とする．

また，p_N および更新後の p_N を式 (21)，式 (22) で表すとする．

$$\vec{p_N} = (p_{N_1}, p_{N_2}, \ldots, p_{N_k}) \quad (21)$$

$$\vec{p_{N_{new}}} = (p'_{N_1}, p'_{N_2}, \ldots, p'_{N_k}) \quad (22)$$

そして，p_R の更新時とは逆に，抽出された単語 w_i が適合文書中の単語の場合は p_N を表すベクトルから $Value_{rel_N}(w_i)$ を引き（式 (23)），非適合文書中の単語の場合は $Value_{nrel_N}(w_i)$ を加える（式 (24)）ことで，非適合文書の特徴を表すように p_N を更新する．

$$p'_{N_i} = p_{N_i} - Value_{rel_N}(w_i) \quad (23)$$

$$p'_{N_i} = p_{N_i} + Value_{nrel_N}(w_i) \quad (24)$$

そして，p_R との類似度が閾値を超えた文書について，p_N との類似度を計算し，あらかじめ設定した閾値を超えた文書は過去に誤って選択した非適合文書に類似していると判断し，選択しない．

以上の処理の流れを**図 3** に示す．ただし，p_R，p_N との類似度を sim_R，sim_N とし，それぞれの閾値を $Thres_R$，$Thres_N$ とする．

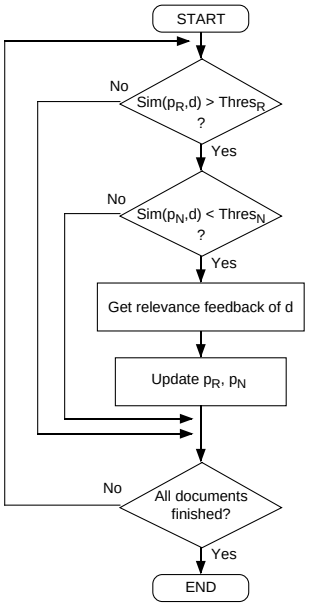

図 3 非適合プロファイルを利用したフィルタリングの流れ
Fig. 3 Filtering using non-relevant information profile.

この手法により，p_R に類似していると判断された文書から，過去に誤って選択された非適合文書に類似している文書を除外し，選択される非適合文書を減少させることができると考えられる．

3.2 実　　験

非適合プロファイルを利用したフィルタリング手法の評価のために以下の実験を行った．

3.2.1 実験条件

$Thres_R = 0.1$ とし，2.3.6 項の実験より，適合文書を表すプロファイルのみを用いた手法において TREC-7，TREC-8 ともに scaled utility が最も高かったパラメータ設定である，$wgt_{rel_R} = -200$，$wgt_{nrel_R} = -800$ を用いた．

$Thres_N$ は 0.1 と 0.25 の 2 つの値において実験を行い，それぞれ wgt_{rel_N}=$\{-200, -400, -800\}$，wgt_{nrel_N}=$\{-100, -200, -400, -800\}$ の各組合せで実験を行った．また，評価データとして，TREC-7 および TREC-8 の両方のデータを利用した．

3.2.2 結　　果

各 $Thres_N$ について，wgt_{rel_N}，wgt_{nrel_N} の各組合せにおける全プロファイルの平均 scaled utility を示す．**表 5**，**表 6** には，TREC-7 の評価データに対し，$Thres_N = 0.1, 0.25$ の結果をそれぞれ示す．また，**表 7**，**表 8** には，TREC-8 の評価データに対する結果を示す．

表 5 平均 Scaled Utility（TREC-7, $Thres_N = 0.1$）
Table 5 Average scaled utility (TREC-7, $Thres_N = 0.1$)

	wgt_{nrel_N}			
wgt_{rel_N}	−100	−200	−400	−800
−200	0.5228	0.5262	0.5167	0.5164
−400	0.5216	0.5240	0.5125	0.5220
−800	0.5224	0.5236	0.5250	0.5280

表 6 平均 Scaled Utility（TREC-7, $Thres_N = 0.25$）
Table 6 Average scaled utility (TREC-7, $Thres_N = 0.25$).

	wgt_{nrel_N}			
wgt_{rel_N}	−100	−200	−400	−800
−200	0.4961	0.4979	0.5003	0.5021
−400	0.4960	0.4972	0.4995	0.5015
−800	0.4957	0.4970	0.4984	0.5000

表 7 平均 Scaled Utility（TREC-8, $Thres_N = 0.1$）
Table 7 Average scaled utility (TREC-8, $Thres_N = 0.1$).

	wgt_{nrel_N}			
wgt_{rel_N}	−100	−200	−400	−800
−200	0.5660	0.5755	0.5814	0.5852
−400	0.5667	0.5702	0.5810	0.5858
−800	0.5690	0.5728	0.5743	0.5863

表 8 平鋸 rcaled Utility（TREC-8, $Thres_N = 0.25$）
Table 8 Average scaled utility (TREC-8, $Thres_N = 0.25$).

	wgt_{nrel_N}			
wgt_{rel_N}	−100	−200	−400	−800
−200	0.5448	0.5464	0.5448	0.5508
−400	0.5448	0.5466	0.5491	0.5466
−800	0.5408	0.5466	0.5484	0.5505

表 5 より，TREC-7 の実験では従来のプロファイル p_R のみを用いた手法と比較して scaled utility が 2.8%から 4.2%向上していることが分かる．また，表 6 より，$Thres_N = 0.1$ とすることで，より scaled utility の向上は大きくなり，6.3%から 9.5%の向上が見られることが分かる．また，表 7，8 に示された結果より，TREC-8 の実験でも $Thres_N = 0.25$ の場合は 2.9%から 4.8%，$Thres_N = 0.1$ の場合は 7.7%から 12%の scaled utility の向上が見られることが分かる．

3.3 考　　察

$Thres_N$ の変化による影響を調べるため，$Thres_N = 0.1$，$Thres_N = 0.25$ のそれぞれにおいて，p_R との類似度 sim_R が $Thres_R$ を超えた文書の p_N との類似度 sim_N が，適合文書・非適合文書によってどのように分布しているかを調査した．TREC-8 の実験において，p_R との類似度 sim_R が $Thres_R$ を超えたすべての文書（適合文書および非適合文書）の sim_R，sim_N の関係を，$Thres_N = 0.1$ および $Thres_N = 0.25$ の条件ごとにそれぞれ**図 4**，**図 5** に示す．なお，図 4，5 に示された実験での wgt_{rel_N}，wgt_{nrel_N} の値はそれぞれ −800，−200 である．

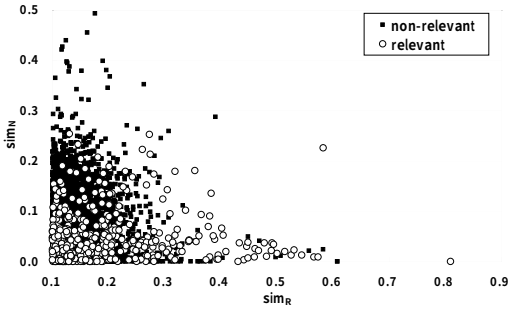

図 4 各文書と p_R，p_N との類似度（$Thres_N = 0.1$）
Fig. 4 Similarity of p_R, p_N and each document ($Thres_N = 0.1$).

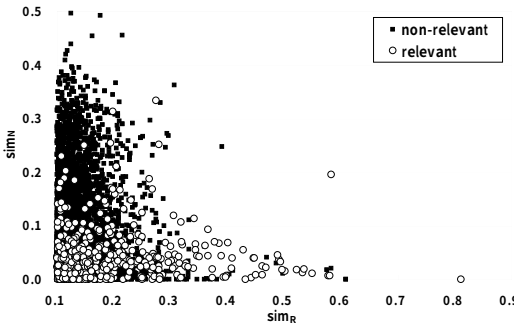

図 5 各文書と p_R，p_N との類似度（$Thres_N = 0.25$）
Fig. 5 Similarity of p_R, p_N and each document ($Thres_N = 0.25$).

図 4，図 5 より，$Thres_R$ を超える sim_R を持つ文書には適合文書と非適合文書が混在している様子が分かる．このことからも，p_R との類似度のみでは適切な閾値を設定することが困難であるという，既存手法の問題点が明らかである．

一方 sim_N に関しては，図 5 より $Thres_N = 0.25$ としたとき，適合文書は全体的に sim_N が小さいものが多く，非適合文書は sim_N の高い位置にまで分布していることが分かる．このことから，sim_N に適切な閾値を設定することで，誤って選択される非適合文書を減らし，より精度の高いフィルタリングを実現することが可能であるといえる．しかし，sim_N が 0.25 を超える非適合文書は少なく，$Thres_N = 0.25$ では非適合文書選択数を大幅に減少させる結果とはなっておらず，結果として scaled utility が $Thres_N = 0.1$ のときに比べ，下がっている．したがって，$Thres_N = 0.25$ で

は非適合文書選択数を大幅に減少させる結果とはなっていない．

一方で図 4 より，$Thres_N = 0.1$ とすると，sim_N に関して，$Thres_N = 0.25$ の場合と比較して適合文書と非適合文書が混在している様子が分かる．この 2 つの実験の違いは $Thres_N$ の値である．$Thres_N$ を 0.1 と，厳しく設定した結果，選択された文書の数，すなわち適合フィードバックが得られる文書の数が減少する．その結果，図 4 から明らかなように，p_N が高い精度で非適合文書を識別するためのフィードバック情報が不足していたと考えられる．

以上より，効果的な非適合プロファイルを作成するためにはフィードバック情報が多いほど効果的であるが，多くのフィードバック情報を得るために閾値を低く設定すると，除外される非適合文書数が減少するということが分かった．一方で多くの非適合文書を除外するために閾値を厳しく設定すると，効果的な非適合プロファイルを作成するために必要なフィードバック情報が得られないということも明らかになった．すなわち，厳しい閾値の設定と非適合プロファイル更新のためのフィードバック情報量の間にはトレードオフが存在するということである．

4. pseudo feedback による非適合プロファイル更新

3.2 節の評価実験の結果から，非適合プロファイルの効果を高めるためには，非適合プロファイルの閾値を厳しく設定し，かつ非適合プロファイルに対するフィードバック情報を十分に与える必要があることが明らかになった．そこで，pseudo feedback に基づいたフィードバック手法を使用することにより，非適合プロファイルとの類似度の閾値を厳しくした際のフィードバック情報の減少を補う手法を提案する．

4.1 提案手法

情報検索では，検索式拡張に利用する情報を得る適合フィードバックの手法として，初期検索に対する適合性の判断をユーザが返す manual feedback と，初期検索の結果から文書の適合性を仮に設定し，その情報をシステムに返す pseudo feedback[11] の 2 つの手法が提案されている．

これまで述べられたプロファイル更新手法でのフィードバックは，あらかじめ与えられた各プロファイルごとの正解文書データを使用しているため，manual feedback の一種であるといえる．ここでは pseudo feedback を取り入れることで，非適合プロファイルの更新に利用する文書情報を増やす手法を提案する．

ここで提案するフィードバック手法は p_R との類似度が閾値を超え，p_N によるフィルタリングの結果選択されなかった文書をすべて非適合文書であると見なし，システムにフィードバックするという手法である．そしてこの pseudo feedback による情報も非適合プロファイル更新に利用することで，非適合文書の更新に利用する情報を増やす．これにより，p_N との類似度の閾値を厳しくした際の，選択文書の減少にともなう p_N 更新に利用するフィードバック情報の減少を補う．

以上のように pseudo feedback を取り入れることで，非適合プロファイルとの閾値を厳しくすることによって非適合プロファイルの効果を活かし，かつ，多くの文書情報をもとに有効な非適合プロファイルを作成できると期待される．

4.2 実験

pseudo feedback を導入したフィルタリング手法の評価のために以下の実験を行った．

4.2.1 実験条件

3.2 節における実験と同様の手法で，$Thres_R = 0.1$，$Thres_N = 0.1$，$wgt_{rel_R} = -200$，$wgt_{nrel_R} = -800$ とし，$wgt_{rel_N} = \{-200, -400, -800\}$，$wgt_{nrel_N} = \{-100, -200, -400, -800\}$ の各値において実験を行った．また評価データは TREC-8 のものを使用した．

4.2.2 結果

図 6 に，$wgt_{rel_N} = -800$，$wgt_{nrel_N} = -200$ において p_R との類似度が $Thres_R$ を超えた文書の，sim_R，sim_N を適合文書・非適合文書で区別して示す．また，wgt_{rel_N}，wgt_{nrel_N} の各組合せにおける全プロファイルの平均 scaled utility を**表 9** に示す．

図 6 より，3.2 節における実験で $Thres_N = 0.1$ とした場合と比較して，非適合文書の sim_N が高くなっていることが分かる．これは，厳しい $Thres_N$ を設定しながらも，pseudo feedback で得られた情報をプロファイル更新に使用することで，有効な非適合プロ

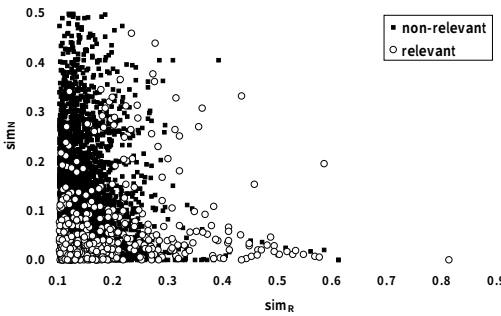

図 6 各文書と p_R，p_N との類似度（pseudo, $Thres_N = 0.1$）
Fig. 6 Similarity of p_R, p_N and each document (pseudo, $Thres_N = 0.1$)

表 9 平均 Scaled Utility（TREC-8, pseudo）
Table 9 Average scaled utility (TREC-8, pseudo).

	wgt_{nrel_N}			
wgt_{rel_N}	−100	−200	−400	−800
−200	0.5752	0.5799	0.5900	0.5927
−400	0.5779	0.5803	0.5859	0.5954
−800	0.5790	0.5813	0.5862	0.5896

ファイルが作成されたことを示す．このことは，表 9 に示されるように，pseudo feedback を用いることで scaled utility が 0.6%から 2.1%向上していることからも明らかである．

以上より，pseudo feedback を利用したフィルタリング手法の有効性が確認された．

4.3 考　　察

pseudo feedback を行うことで，非適合文書の sim_N が全体的に高くなり，多くの非適合文書を除外することが可能となった一方で，図 6 から明らかなように一部の適合文書の sim_N も上昇しており，適合文書の選択数が減少する結果となっている．提案手法では，sim_N が $Thres_N$ を超えたものをすべて非適合文書として扱ったが，非適合文書としてフィードバックされた文書中に適合文書も含まれていたことが原因と考えられる．

フィードバック情報の信頼性を上げる方法の 1 つとして，“疑わしい” と考えられる情報は使用しないという方法がある．非適合プロファイルとの類似度が閾値をわずかに超えた程度の文書は，非適合プロファイルとの類似度が高い文書と比較して，実際に非適合文書である確率が低いと考えられる．そこで，このような文書から得られる情報はフィードバックせずに，非適合プロファイルとの類似度が特に高い文書の情報だけを非適合プロファイル更新に使用することで，pseudo feedback により得られる情報の信頼性を上げることができる．

また，もう 1 つの解決策としては，各文書と非適合プロファイルとの間の類似度に基づき，pseudo feedback で得られるフィードバック情報に重み付けを行うという方法がある．単に “疑わしい” 情報は pseudo feedback に利用しないとするのではなく，疑わしいと考えられるような情報には小さな重みを掛けるようにし，一方で非適合文書である確率が高いと考えられる文書の情報には大きな重みを掛けることにより，フィードバック情報の信頼性向上を図る．具体的には，非適合プロファイルとの類似度の高い文書には大きい重みを掛けたうえでその情報を利用することで，フィードバック情報の質を向上させ，より効果的な pseudo feedback を行うことができると考えられる．

5. 結　　論

本論文では，検索式拡張手法を適用して更新されたプロファイルとの類似度のみによりフィルタリングを行う手法の問題点を提起し，過去の非適合文書情報によって構成した新たなフィルタを導入する手法を提案した．

情報検索における検索式拡張手法をプロファイル更新に適用し，プロファイルとの類似度によりフィルタリングを行う従来手法では，多くの非適合文書が誤って選択されてしまうという問題点が明らかになった．そこで，過去に選択された非適合文書から抽出された情報に基づき作成された非適合プロファイルを利用したフィルタリング手法を提案した．評価実験において，従来の適合文書を表すプロファイルとの類似度が閾値を超えたすべての文書に対し，非適合プロファイルとの類似度を測定したところ，非適合文書の類似度は適合文書と比較して高く，また，scaled utility にも向上が見られたことから，非適合プロファイルを利用した手法の有効性が確認された．

さらに，非適合プロファイルとの類似度の閾値を変化させ，比較を行った．その結果，フィードバックされる情報が多いほどより効果的な非適合プロファイルを作成することが可能であるが，そのために閾値を緩くすると非適合プロファイルにより除外される非適合文書数が減少するということが分かった．そこで，pseudo feedback を行い，非適合プロファイル更新に利用するフィードバック情報を増やす手法を取り入れ，その評価を行った．その結果，非適合文書の非適合プロファイルとの類似度が全体に上昇したことから，pseudo feedback を取り入れる手法の有効性が確認された．

文書フィルタリングで有効性が確認されている手法の 1 つに，プロファイルとの類似度の閾値を動的に調整する手法[7]があげられる．本論文では非適合プロファイルの有効性を検証することが目的であるため，このような閾値の調整手法を取り入れた評価実験については報告していない．しかし，従来のプロファイルならびに非適合プロファイルのそれぞれに閾値調整を適用することにより，さらにフィルタリング精度が向上する可能性は高く，今後検討を行う必要がある．

謝辞 日頃ご指導いただく KDD 研究所秋葉所長に感謝いたします．また，本論文においてご指導いただいた早稲田大学白井克彦教授，および本論文の評価実験において多大なご協力をいただいた早稲田大学の

大西亜希子氏ならびにスウェーデン・Uppsala 大学の Rickard Johansson 氏に感謝申し上げます.

参 考 文 献

1) Ng, Ang, Soon: DSO at TREC-8: A Hybrid Algorithm for the Routing Task, *Proc. 8th Text REtrieval Conference* (2000).
2) Voorhees, E. and Harman, D.: *7th Text REtrieval Conference*, NIST SP 500-240 (1997).
3) Hull, D.A.: The TREC-7 Filtering Track: Description and Analysis, *7th Text REtrieval Conference*, pp.33–56 (1999).
4) Rocchio, J.: Relevance Feedback in Information Retrieval, *SMART Retrieval System Experiments in Automatic Document Processing*, pp.313–323, Prentice Hall Inc. (1971).
5) Buckley, C. and Salton, G.: Optimization of Relevance Feedback Weights, *Proc. SIGIR'95*, pp.351–357 (1995).
6) Singhal, A., Choi, J., Hindle, D., Lewis, D. and Pereira, F.: AT&T at TREC-7, *The 7th Text REtrieval Conference*, pp.239–251 (1999).
7) Zhai, C., Jansen, P., Roma, N., Stoica, E. and Evans, D.A.: Optimization in CLARIT Adaptive Filtering, *Proc. 8th Text REtrieval Conference* (2000).
8) 帆足，松本，井ノ上，橋本：文書間の類似度における単語寄与度を利用した検索式拡張手法，情報処理学会論文誌：データベース，Vol.40, No.SIG 8 (TOD 4), pp.63–73 (1999).
9) 帆足，松本，井ノ上，橋本：文書フィルタリングにおけるプロファイル更新手法の検討，電子情報通信学会「知識発見のための自然言語処理」シンポジウム (1999).
http://www.pluto.ai.kyutech.ac.jp/plt/inui_lab/pub/NLP_Sympo99/hoashi/
10) Witten, I., Moffat, A. and Bell, T.: *Managing Gigabytes: Compressing and Indexing Documents and Images*, Van Nostrand Reinhold (1994).
11) Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. and Gatford, M.: Okapi at TREC-3, *Overview of the 3rd Text REtrieval Conference*, pp.109–125 (1994).

(平成 12 年 5 月 24 日受付)
(平成 13 年 1 月 11 日採録)

帆足啓一郎（正会員）

平成 7 年早稲田大学理工学部情報学科卒業．平成 9 年同大学大学院修士課程修了．同年国際電信電話（株）入社．現在,（株）KDD 研究所インターネットアプリケーショングループにおいて情報検索，情報フィルタリング等の研究に従事．

松本　一則（正会員）

昭和 59 年京都大学工学部情報工学科卒業．昭和 61 年同大学大学院修士課程修了．同年国際電信電話（株）入社，研究所所属．現在，同研究所インターネットアプリケーショングループにて，時系列データ処理，類似検索の研究開発に従事．特に実例からの知識獲得手法に興味を持つ．電子情報通信学会会員．

井ノ上直己（正会員）

昭和 57 年京都大学工学部電子工学科卒業．昭和 59 年同大学大学院修士課程修了．同年国際電信電話（株）入社．昭和 62〜平成 3 年 ATR 自動翻訳電話研究所に出向．知識ベース，自然言語処理の研究に従事．平成 3 年より，KDD 研究所において機械翻訳，音声認識，情報検索の研究に従事．工学博士．平成 3 年度学術奨励賞受賞．平成 7 年度日本音響学会技術開発賞受賞．電子情報通信学会，日本音響学会各会員．

橋本　和夫（正会員）

昭和 59 年東北大学工学部電子工学科卒業．昭和 54 年同大学大学院修士課程修了．同年国際電信電話（株）入社，研究所所属．現在，同研究所インターネットアプリケーショングループ（リーダ）．自然言語処理，知識表現，エキスパートシステム等の研究開発に従事．電子情報通信学会，人工知能学会各会員．