サポートベクターマシンによる適合性フィードバックを用いた情報検索

柘植 覚[†] 獅々堀 正幹[†]
黒岩 眞吾[†] 北 研二[††]

近年のインターネット技術の発展により，World Wide Web（WWW）を代表とする個人で扱えるオンラインテキストデータの量が増加している．それにともない，莫大なテキストデータ中から必要な情報を検索する機会も増え，情報検索に関する研究への関心が高まっている．情報検索システムとして，検索対象文書と検索質問を多次元ベクトルで表現するベクトル空間モデル（VSM: Vector Space Model）が広く使用されている．VSMを用いた検索システムの精度を改善する手法の1つとして，適合性フィードバック手法（Relevance Feedback）が提案されている．この手法は，VSMを用いた1次検索結果に対し，利用者が適合・不適合の判断を行いその情報をシステムにフィードバックし，再検索を行うことで検索精度を向上させている．本論文では，この利用者からのフィードバック情報を検索対象文書全体の適合・不適合の判別に用いた．判別を行う識別器として，従来手法より，判別の能力が高く，汎化性に優れたサポートベクターマシン（SVM: Support Vector Machine）を用いた．このフィードバック手法をサポートベクターマシンによる適合性フィードバックとして本論文で提案する．日本語テストコレクション（BMIR-J2）を用いた類似文書検索実験において，提案手法は従来手法と比較し，利用者が判断し，システムにフィードバックされる文書数が50の場合，24.0%の検索精度改善を得ることが可能であった．

Relevance Feedback Using Support Vector Machine for Information Retrieval

Satoru Tsuge,[†] Masami Shishibori,[†] Shingo Kuroiwa[†] and Kenji Kita[††]

With the rapid growth of online information, e.g., the World Wide Web (WWW), a large collection of full-text documents is available and opportunity for getting a useful piece of information is increased. Information Retrieval (IR) is now becoming one of the most important issues for handling large text data. Relevance feedback is a technique that improves retrieval performance based on relevance judgments from the user. Here, we propose the relevance feedback method using Support Vector Machine (SVM). Experiment results on Japanese test collection BMIR-J2 show that the proposed method is useful feedback method comparing to the conventional feedback method. Especially, the proposed method improved the performance of IR system.

1. はじめに

近年のインターネット技術の発展により，World Wide Web（WWW）を代表とする個人で扱えるオンラインテキストデータの量が増加している．それにともない，莫大なテキストデータ中から必要な情報を検索する機会も増え，情報検索に関する研究への関心が高まっている．これらの研究は，米国におけるTREC（Text Retrieval Conference）[1]や，日本におけるIREX（Information Retrieval and Extraction Exercise）[2]，NTCIR（NII-NACSIS Test Collection for IR Systems）[3]のワークショップを中心に広く行われている．

情報検索システムとして，検索対象文書と検索質問を多次元ベクトルで表現するベクトル空間モデル（VSM: Vector Space Model）[4]が広く使用されている．このモデルを用いた情報検索システムは質問ベクトルと文書ベクトル間の類似度（コサイン尺度や内積が広く用いられる）を計算し，その値により順位付け

† 徳島大学工学部
Faculty of Engineering, Tokushima University
†† 徳島大学高度情報化基盤センター
Center for Advanced Information Technology, Tokushima University

を行い，検索結果として出力する．しかし，VSMを用いた情報検索システムの検索性能は十分とはいえない．そのため，検索精度を改善する手法が多く提案されている．その中の一手法として，適合性フィードバック（Relevance Feedback）☆がある．この手法は，利用者が提示された検索結果に対して適合・不適合の判断を行い，その結果をシステムにフィードバックし，再検索を行うことにより検索精度を向上させる手法である．利用者からのフィードバック情報をシステムに反映させる際の修正対象として，検索モデル，検索質問，検索文書の3種類が考えられる[5]．

フィードバック情報を用いて検索モデルの変更や文書中の索引語の重みを変更する手法は，文献6)をはじめ数多く提案されている．このフィードバック手法は，特定の利用者が使用する場合には有効な手法であると報告されている．しかし，複数人で利用した場合，特定利用者の影響が他の利用者に悪影響を及ぼす可能性があるという欠点を持っている．また，検索質問を拡張する手法は文献7)をはじめ数多く提案されている．しかし，この手法ではフィードバック情報を検索質問拡張のみにしか使用しておらず，検索対象文書に関してフィードバックが行われていない．

一方，フィードバック情報を検索文書全体に反映させる方法として，利用者からのフィードバック情報を学習データとして用いて学習した識別器により検索対象文書全体を2つのクラス（適合・不適合）に分類する手法が考えられる．システムにフィードバックされる利用者からの判別結果が多ければ，識別器の学習データが増加し，識別器の識別能力が高まり，検索精度向上が期待できる．しかし，システムへのフィードバック情報を増加させることは，利用者への負担を増加させてしまうという問題を持つ．そのため，検索対象文書全体を判別する識別器には少数のデータで頑健に判別できる能力が要求される．

本論文では検索対象文書の適合・不適合の判別を行う識別器としてサポートベクターマシン（SVM: Support Vector Machine）[8)~11)]を用いた**サポートベクターマシンによる適合性フィードバック手法**を提案し，日本語テキストコレクションBMIR-J2を用いた情報検索実験で本手法の有効性を検証する．SVMは従来の識別器と比較し，識別能力が高くかつ汎化能力が高いという特徴があり，従来のフィードバック手法より少ないフィードバックで高い性能が期待できる[12)]．

SVMは近年多くの分野で研究が進められている．例として，画像識別[13)]，テキスト分類[14)~16)]，音声認識[17),18)]などに用いられている．特に，文献14)~16)において，SVMはテキスト分類に有効であると報告されているため，検索対象文書を適合・不適合に分類する手法としても有効であると考えられる．

以下，次章でSVMの概要を述べ，3章において，本論文で提案するSVMを用いた適合性フィードバック手法について述べる．提案手法の有効性を確かめるため，4章において，日本語テキストコレクション（BMIR-J2）を用いた情報検索実験について述べ，実験結果，考察を5章で述べる．最後に，6章において本論文のまとめと今後の課題を述べる．

2. サポートベクターマシン

SVMは，1960年代にVapnikらにより提案された汎化能力が高い統計的パターン認識手法であり，Optimal Separating Hyperplaneがその源流となっている．1990年代にカーネル手法と組み合わされ非線形識別手法に拡張されることで適応範囲を飛躍的に拡大した．SVMは機械学習における誤りが最小となる仮説を見つける構造的リスク最小化に基づく手法であり，学習データに対しリスク最小となる分離平面を見つけ，データ集合全体を分類するものである．

2.1 最適分離平面の決定

SVMによる2分離のクラスタ化タスクに対する最適分離平面の決定手法を文献19)に従って説明する．2値のクラス正解ラベル（$y \in \{+1, -1\}$）を持つデータ $\boldsymbol{x}_i (i = 1, \cdots, m)$ に対する正解ラベル決定関数を

$$f(\boldsymbol{x}) = sgn(< \boldsymbol{w}, \boldsymbol{x} > +b) \quad (1)$$

とする．ここで，$\boldsymbol{w} = w_1, \cdots, w_m \in \boldsymbol{R}$, $b \in R$, $<,>$ は内積を表す．$sgn(u)$ は，$u > 0$ で1を，$u \leq 0$ で -1 を出力する符号関数である．式(1)は，入力空間である $\boldsymbol{R}$ を $< \boldsymbol{w}, \boldsymbol{x} > +b = 0$ で定義される超平面で2つに分離し，一方に1，逆を -1 に対応させるといえる．入力データが線形分離可能ならばすべてのデータが，

$$y_i \cdot (< \boldsymbol{w}, \boldsymbol{x}_i > +b) \geq 1 \quad (i = 1, \cdots, m) \quad (2)$$

を満足するような $\boldsymbol{w}$，b が存在する．

これより，式(2)に属さない領域がマージンの領域となり，分離平面からデータ $\boldsymbol{x}_i$ までの距離（マージン）$d(\boldsymbol{w}, b; \boldsymbol{x}_i)$ は，

$$d(\boldsymbol{w}, b; \boldsymbol{x}_i) = \frac{|\boldsymbol{w} \cdot \boldsymbol{x}_i + b|}{||\boldsymbol{w}||} \quad (3)$$

から求められる．ここで，マージンとは分離平面と分離平面に最も近い入力データの距離である．

式(2)によって分離される領域のマージンは $1/||\boldsymbol{w}||$

☆ 関連性フィードバック，適合フィードバックなどとも訳される．

である．このマージンを最大にすることにより，与えられたデータに対し最適な分離平面を得ることができる．

マージンが最大となる最適な分離平面は，次の制約付き最適化問題を解くことにより得ることができる．

- 目的関数：

$$L(\boldsymbol{w}) = ||\boldsymbol{w}||^2 \rightarrow \text{最小} \tag{4}$$

- 制約条件：

$$y_i \cdot (<\boldsymbol{w}, \boldsymbol{x}_i> + b) \geq 1 \quad (i=1, \cdots, m) \tag{5}$$

ここで，この最適化問題を解くために，ラグランジュ乗数（$\alpha_i \geq 0 \quad (i = 1, \cdots, m)$）を導入し，式 (4) を

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}||\boldsymbol{w}|| - \sum_{i=1}^{m} \alpha_i (y_i \cdot (<\boldsymbol{x}_i, \boldsymbol{w}> + b) - 1) \tag{6}$$

と変形する．

さらに，b と $\boldsymbol{w}$ の停留点における関係を用いて，式 (6) を変形することにより次の双対問題（Wolfe dual）に帰着させることができる．

- 目的関数：

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j <\boldsymbol{x}_i, \boldsymbol{x}_j> \rightarrow \text{最大} \tag{7}$$

- 制約条件：

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{8}$$

双対問題では，各サンプルに対する α_i に対して最適化となる．その解において $\alpha_i > 0$ となる $\boldsymbol{x}_i$ はサポートベクターと呼ばれる．

上述の式を解くことにより，識別関数は

$$f(\boldsymbol{x}) = sgn \left(\sum_{i; \boldsymbol{x}_i \in SVs}^{m} y_i \cdot \alpha_i \cdot <\boldsymbol{x}, \boldsymbol{x}_i> + b \right) \tag{9}$$

で与えられる．ここで，SVs はサポートベクターの集合を示す．

2.2 カーネル関数

前節では学習データを線形分離可能である場合について考えたが，実際には線形分離不可能な場合が数多く存在する．そこで，学習データを非線形変換することにより高次元空間に写像し，その空間において線形識別をする手法が考えられる．しかし，学習データを非線形変換し高次元に写像するには，莫大な計算量が必要となる．SVM の学習，判別において，入力データに対し必要な演算は内積計算のみである．そこで，高次元空間の内積計算をカーネル関数（$K(\boldsymbol{x}, \boldsymbol{y})$）の計算だけに抑えるカーネル手法を適用し，識別関数（式 (9)）を

$$f(\boldsymbol{x}) = sgn \left(\sum_{i; \boldsymbol{x}_i \in SVs}^{m} y_i \cdot \alpha_i \cdot K(\boldsymbol{x}, \boldsymbol{x}_i) + b \right) \tag{10}$$

とする．同様に目的関数（式 (7)）も変形する．

広く一般に用いられるカーネル関数としては，次の 2 つがあげられる．

- Polynomial カーネル関数

$$K(\boldsymbol{x}, \boldsymbol{y}) = (<\boldsymbol{x}, \boldsymbol{y}> + c)^p \tag{11}$$

- Radial baisis カーネル関数

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp \left(\frac{-||\boldsymbol{x} - \boldsymbol{y}||^2}{2\sigma^2} \right) \tag{12}$$

3. SVM による適合性フィードバック手法

本章では，本論文で提案する SVM による適合性フィードバック手法を用いた情報検索システムについて述べる．

3.1 SVM 適用のためのベクトル作成

上述したとおり SVM は，学習データにより決定された最適分離平面を用い，入力データ（ベクトル）をその分離平面で分離された 2 クラスのいずれかと判別する手法である．そのため，SVM を類似文書検索に用いることが可能ならば，検索対象文書全体を検索質問に対し，適合・不適合に分類することができる．適合と分類された文書のみから再検索を行うことにより，高精度の情報検索ができると考えた．

しかし，SVM は多次元ベクトルに対しクラス分類を行う手法であるため，テキスト形式で記述された文書をそのまま使用することはできない．そこで，類似文書検索手法として広く使用されている VSM を用い，テキスト形式の文書を索引語の多次元ベクトルとして表現する．これにより，類似文書検索に SVM を適用することが可能となる．

また，SVM を用い各検索質問ごとに検索対象文書全体を適合・不適合に分類するためには，各質問ごとに SVM の学習が必要となる．SVM の学習には，従来より情報検索に用いられている適合性フィードバック手法に従い，各検索質問ごとに利用者が適合・不適合の判別を行った情報を用いる．学習された SVM を用い，検索文書全体を適合・不適合に分類し，適合に

分類された文書集合に対し再度検索を行う．本論文では，このフィードバック手法を**サポートベクターマシンによる適合性フィードバック手法**として提案する．

3.2 SVMを用いた適合性フィードバックによる情報検索

SVMを用いた適合性フィードバック手法の手順を以下に示す．

Step 1： 初期検索
利用者により要求された検索質問に対し，VSMを用い検索を行い，類似度の高い上位 N 文書を利用者に提示する（以下，利用者に提示し，システムにフィードバックされる文書をフィードバック文書とする）．

Step 2： 利用者による判断
Step 1 で提示された文書に対し，利用者は適合・不適合の判断を行い，各文書に対し，適合ならば1，不適合ならば -1 のラベルをつける．

Step 3： SVMの学習（最適分離平面の決定）
利用者が判別した文書を用いSVMの学習を行い，検索対象文書全体を適合・不適合に分類する最適分離平面を決定する．

Step 4： 検索結果出力
決定された分離平面により適合と分類された文書に対し，再度VSMを用い検索を行い，**Step 1** で要求された検索質問と類似度の高い上位 M 文書をシステムからの検索結果とする．

ただし，**Step 4** において，SVMにより適合と判断された文書数が M 文書に至らない場合，適合と判断された文書すべてを検索結果として利用者に提示する．

4. 検索実験

本論文で提案するSVMを用いた適合性フィードバック手法の有効性を検討するため，情報検索のための日本語テストコレクション，BMIR-J2[20]を用いた情報検索実験を行った．

4.1 実験条件

BMIR-J2は1994年の毎日新聞から経済と工学に関する記事を取り出した5,080文書からなる検索対象文書とテスト用検索質問60文書から構成される．1検索質問あたりの平均適合文書数は28.4である．

〈前処理〉 テストコレクションに対し，茶筌[21]を用い形態素解析を行い，名詞と判断されかつ2文書以上に出現した単語を索引語として用いた．この処理により索引語数は18,969となった．

〈索引語の重み付け〉 前処理により得られた索引語に対し，VSMを用いて索引語文書行列を作成した．索引語文書行列の各要素に対する重み付けには対数エントロピー法[22]を用いた．この重みは，

- ローカル重み：
$$L_{ij} = log(1 + f_{ij}) \tag{13}$$
- グローバル重み：
$$G_i = 1 + log\left(\sum_j \frac{p_{ij} \cdot log(p_{ij})}{log(n)}\right) \tag{14}$$

であたえられる．ここで，i，j は索引語番号，文書番号を表す．n はテストコレクション中の文書数，f_{ij} は文書 j における索引語 i の出現頻度を表す．また，$p_{ij} = f_{ij} / \sum_j f_{ij}$ である．重み付けにより，索引語文書行列の各要素 d_{ij} は

$$d_{ij} = L_{ij} * G_i \tag{15}$$

で与えられる．

〈SVM〉 SVMの学習，入力データに対するクラス識別にはSVMLight[23]を用いた．SVMの学習データとなる利用者からのフィードバック文書数には，文書数による検索精度の比較を行うため，10，20，30，40，50の5種類を用いた．これらの文書数は，3.2節の **Step 1** で示したフィードバック文書（N）に対応する．

〈従来手法〉 提案手法との比較を行うため，検索質問拡張の適合性フィードバック手法として広く用いられているRocchio-basedフィードバック手法[7]を従来手法として用いた．Rocchio-basedフィードバック手法は，検索質問（Q_i）を

$$Q_{i+1} = Q_i + \alpha \sum_{\boldsymbol{x} \in R_r} \boldsymbol{x} - \beta \sum_{\boldsymbol{x} \in R_n} \boldsymbol{x}, \tag{16}$$

と拡張させ，検索精度を向上させる手法である．ここで，R_r は i 回目の検索結果のうち利用者に適合と判別された文書集合，R_n は不適合と判別された文書集合である．また，α，β は，それぞれ適合，不適合文書をどの程度重要視するかを調整するパラメータである．本論文では，本実験と同条件で行った予備実験において，最も検索精度が高かった $\alpha = 1.0$，$\beta = 0.5$ を用いた．広く知られているように，このフィードバック手法は繰返しを行うごとに検索精度が高くなるといわれている[24]．そのため，繰返しを最大4回まで行い，各繰返しごとに検索精度を調べた．また，フィードバック自体の有効性を調べるため，フィードバックを行わないVSM情報検索システムを用い検索実験を行った．

〈適合・不適合の判別〉 本実験では情報検索用テストセットを用いている．そのため，適合・不適合の判別は，利用者が行ったのではなく，テストセットで与

えられる適合文書を用いて行った．

〈類似度・評価〉 検索を行う際のベクトル間の類似度はコサイン尺度を用いた．検索システムの精度評価はシステムにより出力された類似度が高い上位 50 文書に対し行った．評価尺度には，一般にランク付け検索システムの評価に用いられる再現率・適合率曲線，平均適合率[25),26)]を用いた．これらの計算には “trec_eval” [1)] を用いた．

適合率，再現率の計算は，

$$\text{適合率} = \frac{\text{検索できた関連文書数}}{\text{検索文書数}}$$

$$\text{再現率} = \frac{\text{検索できた関連文書数}}{\text{関連文書数}}$$

で行った．適合率・再現率曲線の各点は，各検索質問ごとに補完適合率を用いて 11 点平均適合率を計算し，全検索質問で平均した値を用いた[27)]．本論文の評価には類似度の高い上位 50 文書を用いているため，各質問ごとの 11 点平均適合率の計算の際に，再現率が高い値の適合率が計算できない可能性がある．“trec_eval” では計算不可能な再現率の平均適合率は 0.0 とし，適合率・再現率曲線の計算を行っている．

また，平均適合率は，各再現率レベルでの適合率の平均値（適合文書が検索された時点での適合率の平均値）である[27)]．この値は，検索できた関連文書がランク上位に検索されていれば，その値は高く，下位になれば低くなる．そのため，検索された文書内に含まれる不適合文書数も考えることができ，再現率も考慮した評価が可能となる．また，関連文書数で除算を行っているため，検索文書数を制限し，関連文書が検索されなかった場合においても適切な評価が可能である．

5. 検索結果・考察

5.1 カーネル関数を用いたフィードバック手法

先に述べたとおり SVM の判別関数には，最適分離平面により線形にデータを判別する線形判別関数とカーネル関数により非線形にデータを判別する非線形判別関数がある．そこで，判別関数の違いが情報検索精度に与える影響について調べた．

本実験では判別関数として，式 (9) に示した線形判別関数，式 (11) に示した Polynomial カーネル関数による非線形判別関数，式 (12) に示した Radial baisis カーネル関数による非線形判別関数を用いた．それぞれのカーネル関数のパラメータは，Polynomial カーネル関数の場合 $p = 3$，$c = 1$，Radial baisis カーネル関数の場合 $1/2\sigma^2 = 0.6$ を用いた．これらのパラメータは，同実験条件で個々のパラメータを変化させ

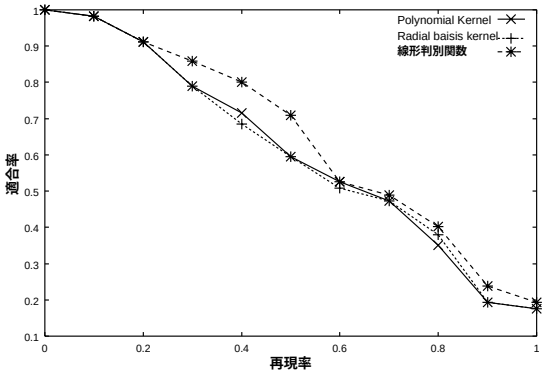

図 1 カーネル関数を用いた場合の検索性能比較
Fig. 1 Retrieval performances using different discriminated functions.

表 1 SVM フィードバック手法におけるフィードバック文書数と平均適合率の比較
Table 1 Average precision using SVM feedback.

フィードバック文書数	平均適合率
フィードッバックなし	0.3291
10	0.3863
20	0.4671
30	0.5340
40	0.5863
50	0.6156

最も検索精度が高かった値である．また，利用者が判別し，システムにフィードバックする文書数は 50 とした．

図 1 にそれぞれの判別関数を用いた場合の情報検索結果を示す．この図より，線形判別を行った場合とカーネル関数を用いて非線形判別を行った場合の情報検索精度には顕著な差が現れないことが分かる．さらに，カーネル関数の違いは検索精度にあまり影響しないことが分かる．

VSM により多次元ベクトル表現された文書ベクトルは十分高次元であり，しかも要素に 0 が多いスパースなベクトルである．そのため，高次元に射影することなく線形分離が可能であったと考えられる．この結果より，以下の実験では，SVM の判別関数は計算量を低く抑えるために線形判別関数を用いる．

5.2 検索精度とフィードバック文書数との関連

本節では，利用者が判断を行いシステムにフィードバックする文書数（フィードバック文書数）とそのフィードバック文書を用いた検索精度との関連性について述べる．**表 1** に，提案手法におけるフィードバック文書数に対する検索精度（平均適合率）を示す．また，従来のフィードバック手法との比較のため，**表 2** に Rocchio-based フィードバック手法を用いた場合の

表 2 Rocchio-based フィードバック手法におけるフィードバック文書数と平均適合率の比較
Table 2 Average precision using Rocchio-based feedback.

フィードバック文書数	平均適合率（重複を除いたフィードバック文書数） 繰返し回数			
	1	2	3	4
10	0.4173 (10.0)	0.4401 (6.0)	0.4501 (4.7)	0.4568 (3.7)
20	0.4756 (20.0)	0.5207 (13.1)	0.5194 (10.2)	0.5327 (8.4)
30	0.4867 (30.0)	0.5338 (21.1)	0.5409 (16.6)	0.5839 (14.0)
40	0.5159 (40.0)	0.5820 (30.0)	0.5475 (23.2)	0.5711 (20.3)
50	0.4940 (50.0)	0.5606 (38.5)	0.5230 (30.5)	0.5806 (27.0)

フィードバック文書数に対する平均適合率を示す．従来手法は，フィードバックを繰り返すことにより検索精度が向上することが知られている[24]．そこで，フィードバックを繰返し行った場合の検索精度も同表に示す．繰返しを行った場合，フィードバック文書中に前回までのフィードバックの際に利用者が判別した文書が含まれる可能性がある．そのため，重複するフィードバック文書を除いた考察を行うため，表中の括弧に重複を除いた平均フィードバック文書数を示した．

表 1 と表 2 の繰返し回数 1 回の比較において，提案手法はフィードバック文書数が 30 以上の場合に従来手法より高い検索精度を示すことが分かる．特にフィードバック文書数 50 の場合，フィードバックを行わない場合と比較して 42.7%，従来手法のフィードバック文書数 50 と比較して 24.0%の平均適合率の改善が見られた．しかし，フィードバック文書数が 20 以下の場合，従来手法より提案手法は検索精度が低くなることが分かる．これは，フィードバック文書数が 10，20 と非常に少ないため，適切に SVM が学習できなかったためと考えられる．実際，フィードバック文書中に適合文書が 1 つも含まれない検索質問がフィードバック文書数 10 の場合 8，20 の場合 5 質問存在した．これらの検索質問を除いた平均適合率はフィードバック文書数 10 の場合 0.4458，20 の場合 0.5095 となり，従来手法とほぼ同等の結果となった．

また，提案手法は，従来手法において最も検索精度が高かったフィードバック文書数 30，繰返し回数 4 回（合計フィードバック文書数 120，重複を除く場合 81.7）の平均適合率（0.5839）より高い平均適合率をフィードバック文書数 40 以上（フィードバック文書数 40：0.5863，フィードバック文書数 50：0.6156）で示した．フィードバック文書から重複を除いた場合においても，同様に提案手法はフィードバック文書数 40 以上において，従来手法の最高平均適合率より高い値を示した．この結果より，提案手法は，従来手法を繰返し適用した場合の合計フィードバック文書数より少ないフィードバック文書数 40，50 で，従来手法と同等の検索精度を示すといえる．さらに，合計フィードバック文書数が 30 以上の場合，同数のフィードバック文書に対する検索精度は，フィードバック文書の重複にかかわらず，提案手法が従来手法より上回っていることが分かる．

さらに，表 1 の結果より，提案手法はフィードバック文書数の増加にともない，検索精度が向上していることが分かる．一方，従来手法では，繰返し回数 1 回におけるフィードバック文書数 40 から 50 の場合など，フィードバック文書数が増加しているにもかかわらず，検索精度の劣化が見られる．これは，従来手法がフィードバックにより検索質問を変動させているため，フィードバック文書数が増加した場合に不適合の文書が増加し，検索質問の拡張が局所的に検索精度を劣化させる方向に変動しているのではないかと考えられる．これに対して提案手法は，フィードバック文書数の増加にともない SVM の学習データが増加し，文書全体を適合・不適合に分離するより適した分離平面を求めることが可能となり，検索精度の向上につながったと考えられる．そのため，利用者が多くの文書に対し適合，不適合の判別を行うことにより，フィードバック文書数に応じたより良い検索結果が得られるといえる．

5.3 検索精度とフィードバック文書中の関連文書数との関係

提案手法とフィードバックを行わない VSM の比較

前節において，提案した SVM を用いた適合性フィードバックは情報検索精度向上に有効であることを示した．本節では，フィードバック文書中に含まれる適合文書数の割合に対する検索精度を調べ，提案手法がどの程度フィードバック文書に適合文書が含まれた場合に有効であるかを検討する．**図 2** に，提案手法を用い検索実験を行った場合とフィードバックを行わない VSM を用い検索実験を行った場合の各検索質問ごとのフィードバック文書中に含まれる適合文書の割合に対する平均適合率を示す．図中の ×，+ は，提案手法と VSM を用いた場合の各質問ごとの平均適合率を

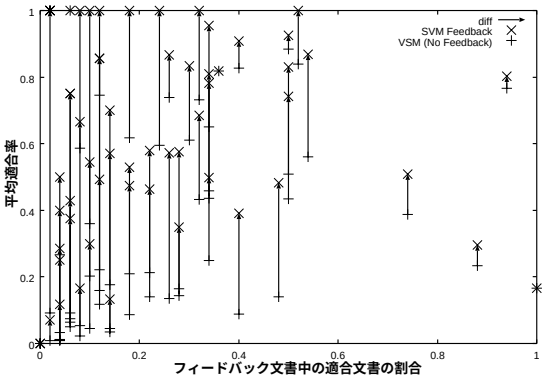

図 2 フィードバックに用いた文書中の適合文書の割合と平均適合率（VSM vs. SVM フィードバック）

Fig. 2 Relationship between the average precision and the proportion of the relevant documents to the feedback documents (VSM vs. SVM feedback).

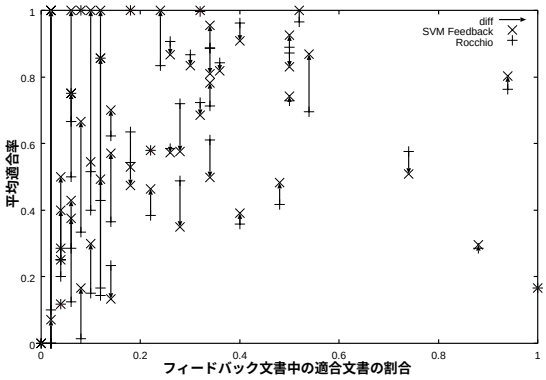

図 3 フィードバックに用いた文書中の適合文書の割合と平均適合率（Rocchio vs. SVM）

Fig. 3 Relationship between the average precision and the proportion of the relevant documents to the feedback documents (Rocchio-based feedback vs. SVM feedback).

示し，矢印（→）は，VSM から提案手法への平均適合率の差を示す．提案手法の平均適合率は，フィードバック文書を 50 文書用いた場合の実験結果より計算した．

図 2 より，提案手法は，適合文書がまったく含まれない場合と全文書が適合している場合を除き，平均適合率を向上させていることが分かる．これらの傾向は，他のフィードバック文書数に対しても同様の結果であり，適合文書がまったく含まれない場合と全文書が適合している場合を除き，平均適合率を向上可能であった．

フィードバック文書中に適合文書が存在しない場合，これらのデータで学習した SVM は全文書を不適合と判別する．そのため，フィードバックが有効に働かず検索精度向上が見られなかったと考えられる．逆に，全文書が適合と判断された場合，全文書が SVM により適合と判別されるため，フィードバックを行わない結果と同様の検索結果しか得られない．今後，全文書適合，不適合であった場合のフィードバック手法の検討が必要である．

提案手法と従来のフィードバック手法との比較

図 2 と同様に，提案手法と Rocchio-based フィードバック手法のフィードバック文書中に含まれる適合文書の割合に対する平均適合率を**図 3** に示す．図中の ×，＋ は，提案手法と従来手法を用いた場合の各検索質問ごとの平均適合率を示し，矢印（→）は，従来手法から提案手法への平均適合率の差を示す．提案手法の平均適合率は，フィードバック文書を 50 文書用いた場合の実験結果より計算した．また，従来手法はフィードバック文書数 50，繰返し回数 1 回の実験結果を用いた．

図 3 より，本提案手法は，従来のフィードバック手法と比較して，フィードバック文書に含まれる適合文書数が少ない場合に特に有効であることが分かる．実際，提案手法の検索精度が従来手法より高い検索質問数は 33 であり，それらのフィードバック文書に含まれた適合文書数の割合は 0.23 であった．逆に検索精度が劣化した検索質問数は 15 であり，それらのフィードバック文書に含まれた適合文書数の割合は 0.33 であった．これは，従来手法ではフィードバック文書中に適合文書が少ない場合に精度向上が低いことに起因する．図 2 におけるフィードバックを行わない VSM の精度と図 3 における Rocchio-based フィードバック手法の精度を比較すると，フィードバック文書中の適合文書の割合が少なくなるほど精度向上率が低い．これに対して，提案手法は，SVM により不適合と判別された文書は再検索が行われないため，検索結果中に適合文書の割合が増加し，従来手法に比べても精度向上が得られたと考えられる．同様の考察を提案手法の有効性が確認された最小のフィードバック文書数 30 について行ったところ，本考察とほぼ同等の結果となった．実際，提案手法の検索精度が従来手法より高い検索質問数は 23 であり，それらのフィードバック文書に含まれた適合文書数の割合は 0.36 であった．逆に検索精度が劣化した検索質問数は 18 であり，それらのフィードバック文書に含まれた適合文書数の割合は 0.40 であった．

6. ま と め

本論文では情報検索精度向上を行うためのサポート

ベクターマシン（SVM: Support Vector Machine）を用いた適合性フィードバック手法を提案した．SVMは高い識別能力と汎化能力をあわせ持つため，利用者からの少ないフィードバック情報で検索文書全体を適合・不適合に判別可能である．

日本語テストコレクション（BMIR-J2）を用いた実験において，本提案手法の有効性を示した．実験結果より，提案手法は従来手法を繰返し適用した場合の最高検索精度（0.5839，フィードバック文書数81.8）を約半分のフィードバック文書数40で実現できた．また，合計のフィードバック文書が同一ならば，フィードバック文書数30以上で提案手法は従来手法より高い検索精度を示した．

本論文において，SVMは繰返しを行わずフィードバックをさせたが，文献28)においてSVMの繰返し手法が提案されている．そのため，今後，本提案手法を繰返し適用させるアルゴリズムの検討を行う予定である．また，SVMを単に文書集合全体の適合・不適合の判別に用いたが，適切な分離平面が決定された場合，分離平面から各文書への距離を類似度計算に利用することが可能であると思われる．今後，類似度計算に分離平面からの距離を利用することを検討する．

実験結果より，適合・不適合の判別は線形判別が可能であることが分かった．そのため，他の識別器を適合性フィードバックに利用することを今後行う予定である．また，本論文において，計算量に関する考察を行っていないが，実システムなどに応用した場合には計算量も非常に大きい要因であるため，今後，計算量の考察，計算量を削減する手法などの検討を行う予定である．

参 考 文 献

1) TREC Homepage. http://trec.nist.gov/
2) IREX Homepage. http://cs.nyu.edu/cs/projects/proteus/irex/
3) NTCIR Homepage. http://www.rd.nacsis.ac.jp/~ntcadm/
4) Salton, G. and McGill, J.: *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company (1983).
5) 徳永健伸：情報検索と言語処理，東京大学出版会 (1999).
6) Tai, X., Sasaki, M., Tanaka, Y. and Kita, K.: Improvement of Vector Space Information Retrieval Model based on Supervised Learning, *Proc. Information Retrieval with Asian Languages* (*IRAL*), pp.69–74 (2000).
7) Rocchio, J.: Relevance Feedback in Information Retrieval, *The SMART Retrieval System*, Salton, G. (Ed.), pp.313–323, Englewood Cliffs, N.J., Prentice Hall (1971).
8) Cortes, C. and Vapnik, V.: Support-Vector Networks, *Machine Learning*, Vol.20, pp.273–297 (1995).
9) Scholkoph, B., Burges, C. and Smola, A.: *Advance in Kernel Methods — Support Vector Learning*, The MIT Press (1999).
10) 麻生英樹：サポートベクトルマシン入門，人工知能学会研究会，Vol.SIG-CII, pp.22–25 (2000).
11) 前田英作：痛快! サポートベクトルマシン —古くて新しいパターン認識手法，情報処理学会誌，Vol.42, No.7, pp.676–683 (2001).
12) Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer-Verlag (1995).
13) Chapelle, O., Haffner, P. and Vapnik, V.: SVMs for Histogram-Based Image Classification, *IEEE Trans. Nerural Networks*, Vol.10, No.5, pp.1055–1065 (1999).
14) Joachims, T.: Text categorization with Support Vector Machines: Learning with many relevant features, *Proc. European Conference on Machine Learning* (*ECML*), pp.137–142 (1998).
15) 平 博順，春野雅彦：Support Vector Machineによるテキスト分類における属性選択，情報処理学会論文誌，Vol.41, No.4, pp.1113–1123 (2000).
16) Drucker, H., Wu, D. and Vapnik, V.: Support Vector Machines for Spam Categorization, *IEEE Trans. Nerural Networks*, Vol.10, No.5, pp.1048–1054 (1999).
17) Bazzi, I. and Katabi, D.: Using Support Vector Machines for Spoken Digit Recognition, *Proc. ICSLP* (2000).
18) Ganapathiraju, A., Hamaker, J. and Picone, J.: Hybrid SVM/HMM Architectures for Speech Recognition, *Proc. ICSLP* (2000).
19) 赤穂昭太郎，津田宏治：サポートベクターマシン基本的仕組みと最近の発展，数理科学，pp.52–59 (2000).
20) Kitani, T., et al.: Lessons from BMIR-J1: A Test Collection for Japanese IR Systems, *Proc. SIGIR*, pp.345–346 (1998).
21) 松本裕治，北内 啓，山下達雄，平野善隆，松田寛，浅原正幸：日本語形態素解析システム『茶筌』version 2.0使用説明書第二版，奈良先端科学技術大学院大学，Technical Report NAIST-IS-TR99012 (1999).
22) Chisholm, E. and Kolda, T.: New Term Weighting Formulas for the Vector Space Method in Information Retrieval, *Technical Memorandum ORNL-13756* (1999).
23) Joachims, T.: SVMLight: Support Vector Ma-

chine, Unversity of Dortmund (1999). http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/
24) Buckley, C., Salton, G. and Allan, J.: The Effect of Adding Relevance Information in a Relevance Feedback Environment, *Proc. SIGIR*, pp.292–298 (1994).
25) Lewis, D.: Evaluating Text Categorization, *Proc. Speech and Natural Language Workshop*, pp.312–318 (1991).
26) Witten, I., Moffat, A. and Bell, T.: *Managing Gigabytes: Compressing and Indexing Documents and Images*, Van Nostrand Reinhold, New York (1994).
27) 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).
28) Vijayakumar, S. and Wu, S.: Sequential Support Vector Classifiers and Regression, *Proc. SOCO'99*, pp.610–619 (1999).

(平成 14 年 2 月 6 日受付)
(平成 14 年 10 月 7 日採録)

柘植 覚（正会員）

平成 8 年徳島大学工学部知能情報工学科卒業．平成 10 年同大学大学院工学研究科博士前期課程知能情報工学専攻修了．平成 13 年同大学院工学研究科博士後期課程システム工学専攻修了．平成 12 年徳島大学工学部助手．博士（工学）．音声認識，情報検索等の研究に従事．日本音響学会会員．

獅々堀正幹（正会員）

平成 3 年徳島大学工学部情報工学科卒業．平成 5 年同大学大学院博士前期課程修了．平成 7 年同大学院博士後期課程退学．同年同大学工学部知能情報工学科助手．平成 9 年同大学工学部知能情報工学科講師．平成 13 年同大学工学部知能情報工学科助教授．現在同大学工学部知能情報工学科助教授．博士（工学）．マルチメディア情報検索，自然言語処理の研究に従事．著書『情報検索アルゴリズム』（共立出版），情報処理学会第 45 回全国大会奨励賞受賞．電子情報通信学会，言語処理学会会員．

黒岩 眞吾（正会員）

昭和 61 年電気通信大学電気通信学部通信工学科卒業．昭和 63 年同大学大学院修士課程修了．同年国際電信電話株式会社入社．昭和 63 年～平成 13 年同社研究所において電話音声認識システムの研究・開発に従事．平成 13 年徳島大学工学部助教授．博士（工学）．音声認識，話者照合，情報検索の研究に従事．電子情報通信学会平成 8 年度学術奨励賞，日本音響学会第 3 回および第 5 回技術開発賞受賞．日本音響学会，電子情報通信学会，人工知能学会各会員．

北 研二（正会員）

昭和 56 年早稲田大学理工学部数学科卒業．昭和 58 年沖電気工業（株）入社．昭和 62 年 ATR 自動翻訳電話研究所出向．平成 4 年徳島大学工学部講師．平成 5 年同助教授．平成 12 年同教授．平成 14 年同大学高度情報化基盤センター教授．工学博士．自然言語処理，情報検索等の研究に従事．平成 6 年日本音響学会技術開発賞受賞．著書『確率的言語モデル』（東京大学出版会），『情報検索アルゴリズム』（共著，共立出版）等．