情報検索における能動学習

小野田　崇[†]　村田博士[†]　山田誠二[††]

大規模な文書データベースでの検索における人間と計算機との対話的文書検索方法として，ユーザに検索文書のいくつかを提示し，評価してもらうことで，ユーザの検索意図を絞り込んでいく適合フィードバックがある．そこでは，ユーザからフィードバック毎にクエリベクトルを学習し，修正していくが，その学習方法として分類学習が適用可能である．本稿では，対話的文書検索の学習方法として，サポートベクターマシンによる能動学習を適用した結果について報告する．

Active Learning in the Information Retieval

Takashi Onoda,[†] Hiroshi Murata[†] and Seiji Yamada[††]

We investigate the following data mining problem from Information Retrieval: From a large data set of documents, we need to find those that bind to human interesting in as few iterations of human testing or checking as possible. In each iteration a comparatively small batch of documents is screened for binding the human interesting. We apply active learning techniques for selecting successive batches.

1. はじめに

近年のIT技術の発展に伴い，個人で扱えるテキストデータの量が急激に増加している．このような状況の中，膨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加し，情報検索に関する研究への関心が高まっている．この情報検索に関する研究は，米国におけるTREC(Text Retrieval Conference)[1]，日本におけるIREX(Information Retrieval and Extraction Exercise)[2]や，NTCIR(NII-NACSIS Test Collection for IR System)[3]のワークショップを中心に広く行われている．

情報検索の枠組みとして，検索対象文書とクエリを多次元ベクトルで表現する**ベクトル空間モデル** (vector space model)[4]が広く利用されている．このモデルを用いた情報検索システムは，クエリベクトルと文書ベクトル間の類似度をベクトル間の内積などの計算により求め，その値の高い文書を検索結果として提示する．

このベクトル空間モデルに基づく情報検索システムの検索精度をユーザと対話的に改善する手法として，**適合フィードバック** (relevance feedback)[5]がある．この手法は，提示された検索結果に対し，ユーザが適合，非適合の判定を行い，その判定結果をシステムにフィードバックする．具体的なフィードバック方法としては，ユーザによる適合/非適合の評価を基にクエリベクトルを修正する手法がよく用いられる．これに対し，ユーザによる評価を適合文書クラスの正例，負例としてとらえ，検索対象文書を適合，非適合の2つのクラスに分類する**分類学習**の適用が考えられる[6][7]．

この分類学習に，学習データより決定される分離平面を用い，データ集合を2クラスに分類する能力の高いSupport Vector Machine(SVM)[8][9]を用いる手法が提案されている[7]．文献[7]では，フィードバック情報に基づく，SVMの適用を1回としているが，ユーザにとって，より有用な情報を提示するためには，必ずしもフィードバック情報を利用したSVMの適用を1回にする必要はない．また，従来の適合フィードバックのように，適合性の高いものからランキングした文書のリストを，ユーザによる評価のために表示する方法よりも，SVMによる能動学習の考えを用いて，判別が難しい文書のリストをユーザに評価してもらうことにより，より高い性能の適合フィードバックが期待できる．

このような背景から本稿では，情報検索をSVMに基づく能動学習の枠組みで捉え，提示文書の生成，フィードバック情報の利用を一貫してSVMで行い，検索文書中の適合文書を特定する新たなフィードバック手法を提案する．

岡部，山田[6]は，関係学習による適合文書の分類ルールの学習を対話的文書検索に応用している．彼らの研究は，分類のための知識がルールという記号で表現されるため，ユーザに対する可読性が高く，直接修正することも容易であるという点で評価できるが，SVMなどの連続値を扱える分類学習アルゴリズムの方がより精度の高い分類を実現できる可能性がある．

以下，2章でSVMについて簡単に紹介し，3章において，本稿で提案するSVMを用いた能動学習について述べ

† (財) 電力中央研究所
Central Research Institute of Electric Power Industry
†† 国立情報学研究所
National Institute of Informatics

る．提案手法の有効性を示すため，4 章において，Losangels Times を用いた情報検索実験を行い，実験結果に対する考察を行う．最後に，5 章で本稿のまとめと今後の課題について述べる．

2. Support Vector Machines

学習サンプル $(\mathbf{z}_1, y_1), \ldots, (\mathbf{z}_\ell, y_\ell), z_i \in F, y_i \in \{\pm 1\}$ が与えられ，次式を満たす判別関数 $f_{\mathbf{w},b} = \mathrm{sgn}((\mathbf{w}\cdot\mathbf{z})+b)$ を推定する問題を考える．

$$f_{\mathbf{w},b}(\mathbf{z}_i) = y_i, \quad i = 1, \ldots, \ell. \tag{1}$$

この関数が存在する場合，以下のの制約を考える．

$$y_i \cdot ((\mathbf{z}_i \cdot \mathbf{w}) + b) \geq 1, \quad i = 1, \ldots, \ell. \tag{2}$$

$(\mathbf{w}, b), (-\mathbf{w}, -b)$ のように $\mathbf{w}$ と b の方向の違いにより，同じ超平面判別関数の式が 2 つ存在することとなる．しかし，式 (1) と式 (2) によって判別関数は一意に定めることができる．

汎化能力の高い判別関数は式 (2) で表現されるの制約条件の下，次式を最小化することで推定できる．

$$\tau(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2. \tag{3}$$

この凸最適化問題を解くため，式 (3) の Lagrangian を計算すると

$$\begin{aligned} &L(\mathbf{w}, b, \vec{\alpha}) \\ &= \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{\ell} \alpha_i (y_i((\mathbf{z}_i \cdot \mathbf{w}) + b) - 1), \end{aligned} \tag{4}$$

ここで，$\alpha_i \geq 0$ は Lagrange 乗数である．この Lagrangian を α_i について最大化し，$\mathbf{w}$ と b について最小化する．パラメータ $\mathbf{w}$ と b についての L の導関数は鞍点において次式のように 0 にならなければならないので，

$$\frac{\partial}{\partial b} L(\mathbf{w}, b, \vec{\alpha}) = 0, \quad \frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \vec{\alpha}) = 0. \tag{5}$$

式 (5) から次式が成立する．

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0, \tag{6}$$

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \mathbf{z}_i. \tag{7}$$

結局，$\mathbf{w}$ は学習サンプルの展開式となる．$\mathbf{w}$ の解はただ一つに決まるが，係数 α_i はその必要がない．

Karush-Kuhn-Tucker 条件により，鞍点において Lagrange 乗数 α_i は，式 (2) を正確に表現し直した次式の制約条件に対して非ゼロでなくてはならない．

$$\alpha_i \cdot [y_i((\mathbf{z}_i \cdot \mathbf{w}) + b) - 1] = 0, \quad i = 1, \ldots, \ell. \tag{8}$$

$\alpha_i > 0$ を有するパターン $\mathbf{z}_i$ を *Support Vectors* と呼ぶ．式 (8) より，*Support Vectors* は margin 上に存在することとなる．*Support Vectors* 以外の学習サンプルは凸最適化問題の解法には関係のないものとなる．つまり，*Support Vectors* 以外の学習サンプルは式 (2) の制約条件を自動的に満たし，式 (7) の展開項の部分には現れないのである．

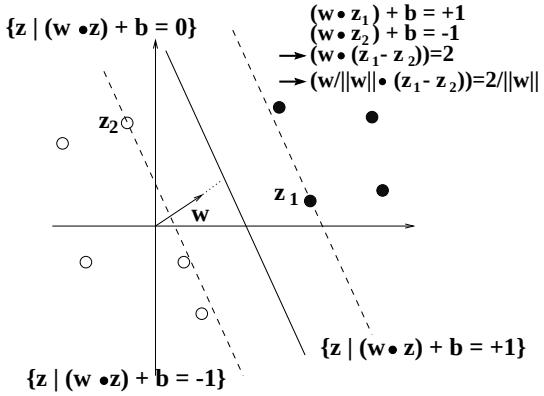

図 1 SVM の例

この凸最適化問題を解いて得られる超平面判別関数の汎化能力については，以下の命題が成立する[10]．

命題 1 サンプル数 ℓ の学習サンプルから得られる *Support Vectors* 数の期待値を $\ell - 1$ で割った値は，未学習サンプルに対する誤分類率の上限である．

式 (4) の Lagrangian に式 (6)，式 (7) の条件を代入すると，双対問題となる次の凸最適化問題を得ることができる．

$$\begin{aligned} \max_{\vec{\alpha}} \ & \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (\mathbf{z}_i \cdot \mathbf{z}_j) \\ \text{subject to } & \alpha_i \geq 0, \quad i = 1, \ldots, \ell, \\ & \sum_{i=1}^{\ell} \alpha_i y_i = 0. \end{aligned} \tag{9}$$

式 (7) の展開式を判別関数の式 (1) に代入することによって，式 (1) の判別関数を，分類されるパターンと *Support Vectors* との内積で評価される次式に書き換えることができる．

$$f(\mathbf{z}) = \mathrm{sgn}\left(\sum_{i=1}^{\ell} \alpha_i y_i (\mathbf{z} \cdot \mathbf{z}_i) + b\right). \tag{10}$$

以上より，式 (9) で表現される凸二次計画問題を解くことで，判別関数 $f_{\mathbf{w},b}(\mathbf{z}) = \mathrm{sgn}((\mathbf{w} \cdot \mathbf{z}) + b)$ を得ることができる．この例を図 1 に示す．図中，○と●は各々異なるラベルを有する学習データを表す．また，破線上の学習データは，Support Vectors と呼ばれる．

現実問題としては，学習サンプルを完全に分離できる超平面は存在しない場合が多い．そのような場合，次式で表現される緩和変数を導入して，式 (2) を満たさない学習サンプルが存在しても良いようにする[11]．

$$\xi_i \geq 0, \quad i = 1, \ldots, \ell. \tag{11}$$

この緩和変数を使って式 (2) の制約条件を次式のように緩和できる．

$$y_i((\mathbf{z}_i \cdot \mathbf{w}) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, \ell. \tag{12}$$

この緩和変数の導入によって，式 (3) と式 (2) で表現される凸最適化問題が次式のようになる．

$$\begin{aligned} \min_{\mathbf{w}, \vec{\xi}} \ & \tau(\mathbf{w}, \vec{\xi}) = \frac{1}{2}||\mathbf{w}||^2 + \gamma \sum_{i=1}^{\ell} \xi_i \\ \text{subject to } & y_i((\mathbf{z}_i \cdot \mathbf{w}) + b) \geq 1 - \xi_i, \\ & i = 1, \ldots, \ell. \end{aligned} \tag{13}$$

適切な正定数 γ を選択できるとすれば，式 (13) で表現される凸最適化問題は，任意の関数集合における，Vapnik の提唱する Structual Risk Minimization の概念を実践することとなる[12]．

学習サンプルが完全に分離できる場合の式 (7) と同様に，式 (13) の最適解において，$\mathbf{w}$ は次式のように，学習サンプルの展開式となる．

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \mathbf{z}_i. \tag{14}$$

ここで，係数 α_i が非ゼロとなるのは，学習サンプル $(\mathbf{z}_i, y_i)$ が制約条件式 (12) を満たす場合である．式 (13) で表現される最適化問題の双対問題となる以下の凸二次計画問題を解くことで，係数 α_i を求めることができる．

$$\begin{aligned} \max_{\vec{\alpha}} \ & \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (\mathbf{z}_i \cdot \mathbf{z}_j) \\ \text{subject to } & 0 \leq \alpha_i \leq \gamma, \quad i = 1, \ldots, \ell, \\ & \sum_{i=1}^{\ell} \alpha_i y_i = 0. \end{aligned} \tag{15}$$

Karush-Kuhn-Tucker 条件から，式 (15) で表現される凸二次計画問題の最適解は次の条件を満たす．

$$\begin{aligned} \alpha_i = 0 &\Rightarrow y_i f(\mathbf{z}_i) \geq 1 \\ 0 \leq \alpha_i \leq \gamma &\Rightarrow y_i f(\mathbf{z}_i) = 0 \\ \alpha_i = \gamma &\Rightarrow y_i f(\mathbf{z}_i) \leq 1 \end{aligned} \tag{16}$$

この条件より，分類結果 $\mathrm{sgn}(f(\mathbf{z}_i))$ が y_i と一致していて，margin 値 $y_i f(\mathbf{z}_i)$ が 1 より大きいサンプルに対応する α_i は 0 になることがわかる．

3. 情報検索における SVM による能動学習

ここでは，能動学習の考えに基づき，SVM による適合フィードバック手法を用いた情報検索について述べる．

2 章で述べたように，SVM は学習データにより決定される最適分離平面を用い，データ集合全体を 2 分割することができる．SVM により分離されるデータ集合が文書集合であるとすると，文書集合がある基準に対し，適合，非適合に分離可能であると考えることができる．文書集合が，適合，非適合に分離可能ならば，情報検索にとって有益な手法として用いることができるはずである．情報検索システムに一般的に使用されているベクトル空間モデルは，文書を多次元ベクトルで表現する．SVM で最適分離平面を決定するための学習データ，および最適分離平面で分離されるデータ集合全体にこの多次元ベクトルを用いることにより，情報検索システムに SVM の導入を図ることが可能となる．

適合フィードバック手法は，情報検索システムがクエリに対する検索を行った結果に対し，ユーザが適合，非適合の判定を行い，その判定結果をシステムにフードバックすることにより，さらに適合性の高い文書を検索する．この適合フィードバックにおけるフィードバック文書 (ユーザが判別を行った文書) を SVM の学習データとして用いれば，検索対象文書全体を適合，非適合に分類することが可能である．適合に分類された文書集合に対し，再度検索を行い，検索を行った結果に対し，ユーザが適合，非適合の判定を行って，その判定結果をシステムにフードバックすることにより，より高精度な情報検索が可能となると期待できる．このフィードバック手法を **SVM による能動学習に基づく適合フィードバック手法**として，本稿で提案する．

なお，SVM を対話的文書検索におうようすることの利点は以下のように考えられる．

- 一般に，文書ベクトルの次元数（属性数）は大きい（10,000 以上）が，SVM は大きい属性数に対応できる．
- 対話的文書検索では，ユーザの評価できる文書数は少ない（数十程度）が，SVM は，少ない訓練例からの学習に適している．

SVM による能動学習に基づく適合フィードバック手法は，以下に示す手続きでフィードバック，検索を行う．なお，Step 4 において，ランダムに文書を選択してユーザの評価を受けるのではなく，最も適合しているであろう文書を優先的に評価してもらうという訓練例選択のバイアスをかけている点で，本手続きは能動学習になっている．

Step 1 初期検索
ベクトル空間モデルを用い，ユーザが要求した質問に対し，検索を行い，類似度の高い上位 N 文書をユーザに提示する．

Step 2 ユーザによる判定
Step 1 で提示された文書に対し，ユーザは適合，非適合の判断を行う．適合と判断された文書には，ラベル “1”，非適合と判断された文書には，ラベル “-1” をつける．

Step 3 最適分離平面の決定 (SVM の学習)
ユーザが判定した文書を用い SVM の学習を行い，検索文書全体を適合，非適合に分類する最適分離平面を決定する．

Step 4 検索
フィードバック回数が M より小さい場合，決定された最適分離平面により，適合と分類された文書に対し，再度ベクトル空間モデルと最適分離平面からの距離を用いて検索を行い，類似度の高い (最適分離

平面からの距離が遠い) 上位 N 文書をユーザに提示し，**Step 2** へ．フィードバック回数が M 以上である場合，**Step 5** へ．

Step 5 検索結果出力

決定された最適分離平面により，適合と分類された文書に対し，再度ベクトル空間モデルと最適分離平面からの距離を用いて検索を行い，類似度の高い (最適分離平面からの距離が遠い)L 文書をシステムから検索結果として出力する．

但し，**Step 5** において，SVM により適合と判断された文書数が L 文書に至らない場合，適合と判断された文書を全て検索結果としてユーザに提示する．

4. 検索実験

4.1 実験条件

§3 で提案した SVM による能動学習に基づく適合フィードバック手法の有効性を検討するための実験を行った．実験用データには，文書検索に関する国際会議 TREC[1] で広く使用されているデータの中の英字新聞記事（The Los Angels Times， 約 13 万記事，平均単語数 526 語）を使用した．このデータには検索要求文とその要求に適合する文書集合が提供されており，本研究でもこれをクエリとして用いている．

文書ベクトルの算出には，文献[13] を参考に一般的な TFIDF[14] を改良した用い，具体的には以下の計算式を使った．

$$w_t^d = L * t * u$$

$$L = \frac{1+\log(tf(t,d))}{1+log(\text{average of } tf(t,d)ind)} \quad (tf)$$

$$t = \log(\frac{N+1}{df(t)}) \quad (idf)$$

$$u = \frac{1}{0.8+0.2\frac{uniq(d)}{\text{average of } uniq(d)}} \quad (normalization)$$

- w_t^d：文書 d における単語 t の重み．
- $tf(t,d)$：文書 d における単語 t の出現頻度
- N：データ集合内の文書総数
- $df(t)$：単語 t を含む文書数
- $uniq(d)$：文書 d における単語の異なり数（種類）

§3 の **Step 1** で述べたフィードバックに用いる文書数 N は，20 とした．また，フィードバックの回数 M は，1，2，3，4 とした．複数のフィードバック回数を用いるのは，フィードバックの回数が検索精度に与える影響を調べるためである．

SVM による最適分離平面は，線形分離により学習データ集合を 2 分割する手法により求めた．2 で述べたように，学習データ集合を分離できない場合には，緩和変数を導入することにより，最適分離平面を決定できるが，本稿で扱うベクトル空間モデルの場合，その空間が多次元であり，学習データ集合を分離できないことがないので，緩和変数を用いていない．また，SVM で最適分離平面を決定する手法として，カーネルトリックを用いる手法[8][9] が一般的に使われているが，本稿で取り扱う文書のベクトル空間モデルは，すでに多次元表現されており，学習データ集合を高次元空間表現する必要がないため，元の文書ベクトル空間での線形分離により最適分離平面を決定した．SVM の学習，クラス分類には，SVMLight[15] を用いて実験を行った．

提案手法の有効性を示すため，フィードバックを行わないベクトル空間モデルに基づく検索システムを基本手法として採用した．また，フィードバック手法との比較を行うため，基本的な適合フィードバックとして広く利用されている Rocchio-based フィードバック手法[5] を従来のフィードバック手法として用いた．

Rocchio-based フィードバック手法は，クエリベクトル (Q_i) を下式により更新し，検索精度を向上させる手法である．

$$Q_{i+1} = Q_i + \alpha \sum_{x \in R_r} x - \beta \sum_{x \in R_n} x, \qquad (17)$$

ここで，R_r は i 回目において検索され，適合と判定された文書集合，R_n は i 回目において検索され，非適合関連がないと判定された文書集合である．また，α，β は定数であり，それぞれ関連文書，関連のない文書をどの程度重要視するかを調整する．本稿では，文献[7] で使用されているのと同様に，経験的によいとされる $\alpha = 1.0$, $\beta = 0.5$ を採用して実験を行った．

広く知られているように，フィードバック手法は繰り返す毎に検索精度が好くなる．そこで，Rocchio-based フィードバック手法，および SVM による能動学習に基づく適合フィードバック手法のフィードバック回数を 1，2，3，4 として精度の比較を行った．

検索システムの精度の評価には，一般的に使用されている適合率 (Precision) と再現率 (Recall) を採用した[16][17]．再現率と適合率は，各々個別に用いてシステム評価を行うことができるが，本稿では，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-適合率曲線を用い，システムの評価を行った．この曲線は，各クエリに対し，一つの曲線が生成されるが，本稿では，全クエリの平均再現率- 適合曲線を用いた．また，これらの平均から得られる平均適合率についても評価を行った．適合率，再現率の計算は，以下の式に従って行った．

$$\text{適合率} = \frac{\text{適合文書数の内，検索できた文書数}}{\text{検索文書数}}$$

$$\text{再現率} = \frac{\text{適合文書数の内，検索できた文書数}}{\text{全適合文書数}}$$

4.2 検索結果

4.2.1 SVM に基づく能動学習によるフィードバックの有効性

SVM の学習に用いる文書数を 20 文書，すなわちユーザが判断を行う文書数を 20 とした検索実験を行った．実験結果を図 2 に再現率-適合率曲線で示す．図 2 は，SVM

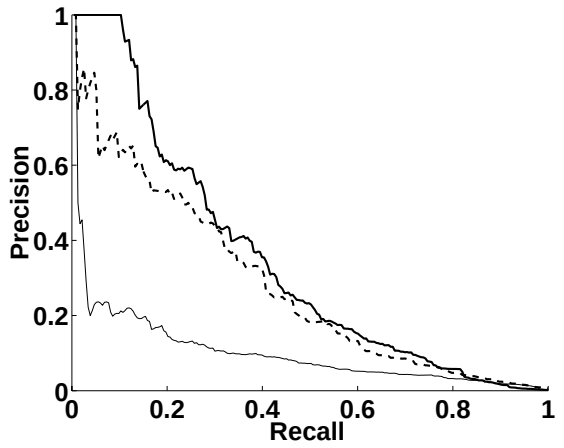

図 2 SVM による能動学習に基づくフィードバック手法の利点 (フィードバック文書数 20 文書，フィードバック回数 4 回の場合の再現率-適合率曲線):図中，太実線は提案フィードバック手法，点線は従来のフィードバック手法，および，細実線はフィードバック無しを表す.

表 1 SVM に基づく能動学習によるフィードバック回数と適合率の比較

フィードバック回数	平均適合率
1	0.2625
2	0.3500
3	0.6125
4	0.6375

による能動学習を 4 回行った後の結果を示している．提案手法である SVM による能動学習に基づく適合フィードバック手法との比較のため，図 2 には，フィードバックを行わない VSM，同条件 (フィードバックを 4 回行った) における従来のフィードバック手法 (Rocchio-based フィードバック手法) の検索結果も示した.

図 2 より，両フィードバック手法は，フィードバックを行わない VSM より検索精度が高いことがわかる．つまり，フィードバックを行い情報検索を行うことは，検索精度向上に有効な手法であることが確認できた.

また，図 2 より，提案した SVM に基づく能動学習によるフィードバック手法は，従来の Rocchio-based フィードバック手法と比較し，検索性能が高い．このため，SVM に基づく能動学習をフィードバックに用いることは，情報検索の精度向上に対し，有効な方法であると考えられる．さらに，以下で，実験結果の考察を行う.

4.2.2 検索精度とフィードバック回数との関連

ここでは，フィードバックの回数を 1，2，3，4 と変化させ，フィードバックの回数と検索精度の関連性について考察する．表 1 に，提案する手法に対するフィードバック回数とフィードバック検索結果を (平均適合率) を示す.

表 1 より，フィードバック回数の増加に伴い，検索結果が向上していることがわかる.

また，比較のため，従来フィードバック手法 (Rocchio-based フィードバック手法) を用いた場合のフィードバック回数と適合率を表 2 に示す。 従来の手法は，フィードバックの回数を増加させることによって，検索精度が向上することが知られている[7]．しかしながら，今回の実験では，フィードバックの回数の増加とともに，検索質問には顕著な変化が見られるものの，検索精度の向上に関しては，顕著な変化がみられなかった。これは，従来手法がフィードバックによって検索質問を拡張しているため，フィードバックの増加とともに，不適合の文書が増加し，検索質問の拡張が有効でなくなっているのではないかと考えられる。ただし，この原因に関しては，式 (17) の係数 α，β の値によっても，結果かが異なる場合が考えられ，今後さらなる考察が必要である.

表 2 Rocchio-based フィードバック手法によるフィードバック回数と適合率の比較

フィードバック回数	平均適合率
1	0.2250
2	0.2500
3	0.2350
4	0.2250

表 3 フィードバック回数と上位 20 文書中の関連文書数

フィードバック回数	関連文書数	
	従来手法	提案手法
1	11	9
2	13	18
3	12	20
4	11	20

表 1，2 により，提案手法と従来手法との比較を行う．フィードバックを行わない場合の平均適合率は，0.15 であった．従来手法も提案手法もフィードバックを行うことにより，検索精度が向上していることがわかる．また，従来手法ではフィードバックを繰り返すことに顕著な検索精度の改善が起っていないが，提案手法では，1,2,3,4 回とフィードバックの回数を増やしていくに伴って，検索精度が向上している．この点を顕著に表している例を表 3 に示す． この例は，最初の検索質問に対し，システムが適合と判断した上位 20 文書の中に，最も多くの関連文書 (5 文書) が入っていた場合である．表 3 には，フィードバック回数とその際に，システムがユーザに提示するために適合と判断した上位 20 文書中の適合文書数を示した．表 3 より，従来手法では，あるフィードバック回数を境に，検索精度の改善が起っていないことがわかる．ただし，この結果は，式 (17) の係数 α，β の値によっても，異なる場合が考えられ，今後さらなる考察が必要である.

さらに，表 1，2 より，フィードバック回数に依らず，常に提案手法の検索精度は，従来手法の検索精度を上回っていることがわかる.

これらの結果より，本稿で提案した SVM に基づく能動学習によるフィードバック手法は，ユーザが判断を行う情報検索において有効な手法であるといえる.

5. おわりに

本稿では，情報検索の検索精度向上のための SVM に基づく能動学習による適合フィードバック手法を提案した．SVM は，高い分類能力を持つため，ユーザからのフィードバックを用いることにより，より関連性の高い文書を識別することが可能である．また，SVM は，最適分離平面を決定できるため，これを利用した追加学習データの生成，つまり，ユーザへ提示し，ユーザに判定してもらう文書の生成が可能である．

TREC の新聞記事のデータを用いた実験において，本稿提案手法の有効性を示した．フィードバック文書が 20 文書という少ない場合でも，本提案手法が従来のフィードバック手法より高い検索精度を示した．

本稿では，最適分離平面を線形識別関数とし，その分離平面からの距離で追加学習データ，つまり，ユーザに提示するフィードバック文書を決定した．しかし，その距離と検索精度との関連についての解析については，今後の課題とする．

参 考 文 献

1) TREC Web page: http://trec.nist.gov/.
2) IREX: http://cs.nyu.edu/cs/projects/ proteus/irex/.
3) NTCIR: http://www.rd.nacsis.ac.jp/˜ntcadm/.
4) Salton, G. and McGill, J.: *Introduction to modern information retrieval*, McGraw-Hill (1983).
5) Rocchio, J.: *Relevance feedback in information retrieval*, Englewood Cliffs, N.J.: Prentice Hall, pp. 313–323 (1971).
6) 岡部正幸, 山田誠二: 関係学習を用いた対話的文書検索, 人工知能学会誌, Vol. 16, No. 1, F (2001).
7) 柘植覚, 獅々堀正幹, 北研二: 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, 研究報告「自然言語処理」, No. 141-14 (2001).
8) 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).
9) 小野田崇: Large Margin Classifiers, 人工知能学会誌, Vol. 17, No. 1, pp. 21–30 (2002).
10) Vapnik, V.: *Statistical Learning Theory*, Wiley, New York (1998).
11) Cortes, C. and Vapnik, V.: Support Vector Networks, *Machine Learning*, Vol. 20, pp. 273 – 297 (1995).
12) Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer (1995).
13) Schapire, R., Singer, Y. and Singhal, A.: Boosting and Rocchio Applied to Text Filtering, *Proceedings of the Twenty-First Annual International ACM SIGIR*, pp. 215–223 (1998).
14) Yates, R.B. and Neto, B.R.: *Modern Information Retrieval*, Addison Wesley (1999).
15) Joachims, T.: Svmlight:Support vector machine, http://www.ai.cs.uni-dortmund.de/ SOFTWARE/SVMLIGHT/ (1999).
16) Lewis, D.: Evaluating text categorization, *Proceedings of Speech and Natural Language Workshop*, pp. 312–318 (1991).
17) Witten, I., Moffat, A. and Bell, T.: *Managing Gigabytes: Compressing and Indexing Documents and Images*, Van Nostrand Reinhold, New York (1994).