Vol.2013-NL-214 No.3
2013/11/14

# 適合性フィードバックにおけるユーザ負荷軽減手法

金子弘明[†1]　梅澤猛[†1]　大澤範高[†1]

情報検索において初期検索結果に対するユーザ評価を基に有用な文献を収集・絞り込みを行う適合性フィードバック手法は，ユーザに特別な検索技術や知識を要さず再検索を容易にする．しかし，適合・不適合の判別精度がフィードバック数と相関を持つため，高い効果を得るにはユーザに検索結果を多くの文献を閲覧・評価する労力を要する．そこで本論文ではユーザの労力を軽減するために，少量のフィードバックから機械学習手法を用いて疑似的なフィードバックを得る手法を検討する．

# Reducing the Effort of the User on Relevance Feedback

HIROAKI KANEKO[†1]　TAKESHI UMEZAWA[†1]
NORITAKA OSAWA[†1]

Relevance feedback in Information retrieval can be used to refine search queries based on user feedback to results returned by initial query and collect relevance documents easier even if a user has no specialized knowledge. Since, precision of relevance feedback is correlated with the number of feedback, highly effective relevance feedback usually needs for the user to give much feedback and then to review many documents. This paper, investigates a method which get pseudo feedback from machine learning trained on a little feedback.

## 1. はじめに

web 検索において，あるキーワードに対して得られた結果がユーザの検索目的にそぐわず効率的な検索が出来ない場合や検索の結果として新たな興味が生じた場合，ユーザは再検索を行う．多くの場合，現在入力されているキーワードに新たな語を加えて絞り込みを行うが，適切に関連文献のみを抽出するキーワードを生成するのは容易ではない．ユーザの検索履歴や大規模な文献コーパスから追加する検索キーワードを推薦する手法も検討されているが，現在検索対象となっている分野に対するユーザの知識が不十分である場合にはキーワードによって得られる分野が推測できず，複数の推薦キーワードを試すために効率が改善されない．

本研究では，初めに得られた検索結果に対するユーザの挙動からユーザがキーワードの入力操作を行わずにユーザの意図に合致した検索結果に絞り込めるよう支援する事を目的とし，適合性フィードバック手法[1]を利用した解決を検討する．適合性フィードバックは，初期検索結果のうちユーザが閲覧した web ページに「適合・不適合」のラベル付を行い，適合文献・不適合文献それぞれに特徴的な情報を発掘し，未評価文献群から有用な文献を収集する手法である．ユーザは提示された文献の「適合・不適合」を判断するだけで良く，特別な知識に基づくアウトプットを必要としないという点でこの手法は優れている．反面，適合文献・不適合文献の特徴を抽出するためにユーザに多くの文献の閲覧・評価する労力を要するという問題がある．

国立情報学研究所が提供する，135 件の検索課題と検索候補である HTML 文書およびユーザ評価からなる情報検索システム評価用データセット NTCIR4-web[2]を用いて，適合性フィードバック手法を利用した再検索の精度とフィードバック数に関する事前実験を行ったところ，フィードバック数と検索精度の向上には正の相関がある事が認められた．フィードバック数とユーザの閲覧・評価に要する時間は比例するため，適合性フィードバックによる検索精度の向上効果とユーザの労力はトレードオフの関係にあると言える．

そのため，より少数のフィードバックから高精度な推測を行う事が効率的な再検索を実現する．柘植[3]は判別能力や汎用性に優れると言われる SVM (Support Vector Machine) を利用して適合・不適合の推測を行い，適合性フィードバックの代表的な手法である ROCCHIO アルゴリズムと比較して高精度なランキングを作成した．本研究では SVM と同様に高い検索精度を持ち，欠損値やノイズに対して強い耐性を持つ Random Forest [4] を用いて，高精度のランキングを作成した．

Random Forest は，ランダムに選択した特徴素群を用いて作成した判別木を弱識別器として，判別森の多数決を最終的な出力とする機械学習手法である．大規模なデータに対しても比較的高速，高精度に動作する事が知られており，SVM と比較してノイズや欠損値に対しても頑健である事から，近年画像照合やテキスト分類分野に応用されている．

NTCIR4-web を用いて比較実験を行ったところ，Random Forest は SVM と比較して少量のデータからでも正確な判別を行う事ができ，より精度の良いランキングを作成する

†1 千葉大学大学院 融合科学研究科
Chiba University, Graduate School of Advanced Integration Science

事が出来た．

また，より少数のフィードバックから高精度な推測を行う手法として半教師学習が挙げられる．半教師学習はラベルが付与されたデータとラベルが付与されていないデータを共に教師データとして学習を行う手法であり，文書分類[5]や動画像処理のような事前にラベルなしのデータが大量に収集出来るという状況において有効に働く．適合性フィードバックにおいても，実際の検索でユーザが数百件ものページを評価する事は困難でありフィードバックは少数しか得られないが，ユーザが入力したキーワードに関連するページ候補は大量に収集可能である．これらのページ候補にラベルを適切に伝播させる事が出来れば，大量の学習データからより高精度な学習が期待できる．

Random Forest は弱判別器からの出力の多数決をとるが，そのクラスごとの得票数を尤度とみなせば，疑似的なフィードバックの拡充の基準を容易に求めることが出来る．これを用いて疑似的なフィードバックを拡充する事で高精度な学習を行えれば，再検索精度を向上させることが出来ると考えた．

これに関して，必ずしもラベルの伝播は正確には行われず誤りを含むおそれがあったため，誤ったラベルを持つデータを含む教師データ群から学習を行った場合の検索精度を検証した．その結果，適合文献を不適合と誤判別した疑似的フィードバックは，不適合文献を適合と誤判別したものと比較して最終的な推測結果を大きくは低減させない示唆も得られた．従って，通常は過半数であるクラス分類の閾値を適合推測に関してのみ高く設定する事で，最終的な出力の精度を向上できると考えた．検証のため，NTCIR4-web を用いて初期フィードバック数 20 件の学習データから推測したラベルを付与した 130 件のデータ合わせて 150 件を新たな学習データとして判別を行うという実験を行った結果, 拡充を行わない 20 件のみを学習データとした場合の MAP(Mean Average Precision)が 0.598 であったのに対して，疑似的フィードバックを含めた 150 件を用いた場合は平均して 0.700 と精度が向上した．しかし，適切なラベルが伝播された場合，150 件のフィードバックがあれば平均して 0.901 となることからまだ改善の余地があると言える．また，適合推測に閾値を設けて不適合文献を適合と誤判別する確率を低減させる手法に関して，閾値の増減は判別精度に影響しなかった．

## 2. 関連研究

柘植[3]は SVM を用いてユーザのフィードバックから適合・不適合の 2 クラス分類モデルを作成する適合性フィードバック手法を提案し，経済・工学に関連する新聞記事を収集した日本語テストコレクション BMIR[6]を用いて適合性フィードバックの代表的な手法である ROCCHIO アルゴリズムと比較実験を行った．結果，SVM を用いた手法は ROCCHIO アルゴリズムを用いた場合より高精度なランキングを作成した．

SVM は特徴量空間におけるクラスの最適分離平面を, 判別面と各クラスで最も判別面に近いデータの距離の最大となるように決定する手法である．非常に高精度な判別が可能な事に加えて，近年カーネル関数を用いて特徴空間をより高次元の空間へ写像する事で線形分離不可能な場合においても十分な効果を発揮する事が確認され，文書分類や推薦に応用されている[7][8]. 特に少量の教師データから効率的な学習を行うために，信頼性の低いデータを改めてユーザに提示し評価を得る能動学習手法などによる対話的な検索手法は高い精度で情報検索を可能としている[9].

高精度な判別を行う機械学習手法として Random Forest が同分野で利用されている．金[10]は 200 編の小説，110 編の作文，60 編の日記と異なるタイプの文章を用いて，Random Forest 法を用いた著者推定手法の有効性を検証した．その結果，SVM などの他分類手法と比較して Random Forest の正解率が高く，また標本サイズの減少による影響が小さい事を確かめた．また Random Forest は新聞記事や研究論文などと比較して複数の話題を含み，くだけた表現など多様な特徴を持つ web 文書の分類に関しても高い精度を示している[11].

本研究では，情報検索技術の評価のための研究用データセットとして NTCIR4-web を用いて Random Forest による適合性フィードバックの有効性を SVM と比較検証すると共に，少ない教師データからより効率的な学習を行うための半教師学習手法を提案した．

## 3. Random Forest による適合性フィードバック

本研究では，初期検索結果に対してユーザから得られた適合・不適合のラベルが付与された文献データを教師データとして, Random Forest を用いた学習を行ってユーザが未評価の文献の適合・不適合を推測し，適合推定文献を優先したランキングを再生成する事によって精度の高い再検索結果を提供する手法を検討している．

なお本研究ではベクトル空間モデル(VSM: Vector Space Model) に従って検索を行う．VSM は単語出現頻度などの特徴を要素とする特徴量空間上での距離で文書の類似度を定義するモデルであり，情報検索は検索クエリを表す点との近傍点である文献を探索する問題と定義される．適合性フィードバックの代表的な手法である ROCCHIO アルゴリズムは適合・不適合文献群の重心ベクトルを用いて検索クエリを修正する．

また Random Forest が適合性フィードバック手法として有効であることを確認するため SVM を用いた従来手法[3]との比較を行う．

本論文で提案する Random Forest を用いた適合性フィードバック手法の検討のための基礎的な実験環境と手順を以下に示す.

### 3.1 実験手順

#### 3.1.1 データセットの作成

実験に使用するデータセットは NTCIR4-web のうち，包括的な検索を目的とする状況を想定した検索課題 35 件を使用した．検索課題はそれぞれ検索者の目的，適合判断基準, 検索者の情報検索技術, 検索キーワードなどの設定と，その検索候補となる HTML 文書数千件およびその評価から構成されている．今回は検索課題をそれぞれ 1 ケースの検索例として実験を適応し，最後に平均をとって実験結果とした．なお，検索候補文献は全ては使用せず，各ケースごとに無作為に 500 件抽出した.

#### 3.1.2 前処理

使用する特徴素は比較手法に従って，形態素解析を行って名詞（代名詞，数詞，形式名詞除く）と判断され，かつ 2 文献以上に出現した単語の出現頻度を用いた．単語の重みづけとして，以下に示す対数エントロピー法 [12] を用いた.

$$t_{ij} = L_{ij} * G_i$$
$$L_{ij} = \log(1 + f_{ij})$$
$$G_i = 1 + \log\left(\sum_j \frac{p_{ij} \cdot \log p_{ij}}{\log n}\right)$$

ここで，$d_{ij}$ は索引語-文書行列の要素であり，i, j はそれぞれ索引語番号と文書番号を表す．また，$f_{ij}$ は文書 j における単語 i の出現頻度を表し，$p_{ij} = f_{ij} / \sum_j f_{ij}$ である．$L_{ij}$ をローカル重み，$G_i$ をグローバル重みと呼ぶ.

#### 3.1.3 初期検索結果

ランキングは VSM に基づき，初期検索結果は全文献の重心ベクトルを検索クエリとして，クエリとのコサイン距離を昇順にソートしたものを，再検索結果では適合推測文献の重心ベクトルを検索クエリとして，クエリとのコサイン距離に適合・不適合の推測結果を加味したものを昇順にソートしたものを生成した．その評価にはランキング順位を考慮した情報検索の指標である MAP (Mean Average Precision)[13] を使用した.

MAP は検索課題ごとに求めた平均精度の平均であり, ある検索課題$r_i$ $(i = 1,2 \dots n)$ の平均精度が$p_i$である時,

$$\text{MAP} = \frac{1}{n}\sum_{i=1}^{n} p_i$$

で求められる．また，平均精度p は各再現率に対する適合率の平均であり，出力文書総数 N のランキングにおいて，第 i 位の文献が適合・不適合に関するラベル$x_i = \{1,0\}$（適合ならば$x_i = 1$，不適合ならば$x_i = 0$）を持つとすると

$$\text{p} = \frac{1}{\sum_{i=1}^{N} x_i}\sum_{i=1}^{N}\left(\frac{x_i}{i}\sum_{k=1}^{i} x_k\right)$$

と定義される．これは同じく適合率と再現率から算出する F 値とはランキング順位を考慮するという点で異なる．平均精度は値が高ければ高いほどランキング上位に適合文献が集中している事を表し，全ての適合文献が最上位から隙間なく並ぶ最高のランキングで値 1 をとる.

#### 3.1.4 機械学習手法

機械学習手法には Random Forest を利用した.

Random Forest は Breiman により 2001 年に提案されたブースティング手法[4]であり，1 本ごとに重複を許してランダムに選択した特徴量をもとに生成した判別森の多数決，または判別木ごとに求めたその出力ノードに到達する頻度確率の平均（または総積）によって判別を行う．大規模なデータベースにも比較的高速・高精度に学習・判別を行える事から，近年画像上の人物同定やテキスト分類など様々な分野に利用されている.

各判別木のノードでの分岐に注目する事で変数の判別への寄与を容易に検証できることや，頻度確率に基づく各判別木の尤度から出力データの判別結果の尤度を設定できることから，拡張時のパラメータ設定を容易に，かつ自動的に行えると考え，本研究で用いた.

以下に Random Forest の学習・判別手順を示す.

入力された教師データ $\text{I} = \{T_1, T_2 \dots T_n\}$ から k 本の判別木を作成する．ここで $T_i$ は web ページを表し，検索課題に対する適合・不適合のラベル $\text{l} = \{+1, -1\}$ と特徴ベクトル $\text{f} = \{f_1, f_2 \dots f_m\}$ を持つ.

(1) ブートストラップ・サンプリング

入力データ I から重複を許してランダムに N 個の文献を抽出してサブセット $\text{I} = \{I_1, I_2 \dots I_k\}$ を生成する．このサブセットをそれぞれ用いて判別木を作成する.

(2) 各ノードにおける判別関数の選択

各木で m 個の特徴量のうち，ランダムに M 個の特徴量を選択する．M が大きすぎる場合，計算量が大きくなると共に過学習が発生する可能性が高まり，M が小さくなると判別の精度が落ちる．経験的にクラスタリングの場合は $\text{M} = \sqrt{m}$ が適当と言われている.

また同時に，各特徴量に対応した分岐関数を選択する．分岐関数は予め用意した分岐関数集合からランダムに抽出したものから最も精度の良いものを選択する．ここでの精度の基準には, 情報利得・GINI 係数などが提案されている.

(3) 木の成長

ある分岐においてその先にはクラスがただ 1 つ存在する場合, その先には判別結果を出力する葉ノードを追加する．それ以外の場合, 先には再び分岐させるノードを追加する.

この再帰を，これ以上再帰を必要としない状態になる，既定の深さまで成長する，誤判別のリスクが一定以下にな

る，などの条件を満たすまで行う．

(4) 判別

判別したいデータを全ての判別木にかけ，出力された値を加算する．そのうえで，最も高い値を示したクラスを判別結果とする．

文献の適合・不適合の判別の学習には WEKA[14]の Random Forest を利用した．WEKA (Waikato Environment for Knowledge Analysis) はニュージーランドのワイカト大学で開発された機械学習ソフトウェアで SVM，決定木など多くの機械学習手法を実装している．Random Forest もそのうちの 1 つとして含まれている．

機械学習の各パラメータを表 1 に示す．

表 1 Random Forest パラメータ設定

| パラメータ | パラメータ名 | 値 |
|---|---|---|
| 警告・備考 | debug | TRUE |
| 木の深さ | maxDepth | unlimited |
| 1 本あたりの特徴量数 | numFeature | $\lfloor\sqrt{\text{特徴量数}}\rfloor$ |
| 木の本数 | numTrees | 100 |
| 乱数シード | seed | 課題番号 |

#### 3.1.5 再検索

推測結果に基づいて新しいランキングを作成し，MAP を用いて精度を計測する．なお，新しいランキングは ROCCHIO アルゴリズムを参考に，初期フィードバックで適合と判断された文献および機械学習により適合と推測された文献の重心ベクトルを新検索クエリとして距離を改めて算出して作成する．またその際，適合と推測された文献が不適合と推測されたものより上位となるよう，推測されたラベルに基づいて距離を修正する．具体的には，本研究で採用したコサイン距離は[0,1]の範囲内の実数であるから，不適合文献の場合に 1 を加算した．

### 3.2 フィードバック数と検索精度

#### 3.2.1 実験目的

一般に機械学習における教師データ数と判別精度には正の相関があると言われる．前述した手法における教師データ数と判別精度を検証する．

情報検索はより少ない手間・時間で目的の情報を効率良く収集する事を目標としている．適合性フィードバック手法においては，より少ないフィードバック数から高精度な再検索結果を提供することにより，ユーザの閲覧・評価時間を削減し，情報検索を効率化させる．

#### 3.2.2 実験内容

3.1 に示した実験手順において，ユーザからのフィードバックとみなすランキングの上位 N 件の範囲を 10~450 の範囲で 10 刻みに増加させながら，各フィードバック数に対する検索精度を測定した．

機械学習手法には，比較として SVM，提案手法に用いるものとして Random Forest を選択した．なお検索精度に加えて，各フィードバック数に対する各手法の判別精度も測定した．SVM にはコーネル大学の Thorsten Joachims によって開発されたフリーウェア SVM$^{light}$[15]を用いた．SVM$^{light}$ には学習手法として線形 SVM，多項式関数・RBF(Radial Basis Function)・シグモイド関数をそれぞれ用いた非線形 SVM が実装されている．最適なモデルとパラメータの設定のため様々な組み合わせを試したが，モデルやパラメータによって最終的な出力が大きく変わる事はなかったため，処理時間を考慮して非線形 SVM で高速に動作した RBF ($f(a,b) = \exp(-0.5 \cdot ||a-b||^2)$) を使用した．

#### 3.2.3 実験結果

フィードバック数と検索精度に関する実験結果を図 1 に示す．横軸はフィードバック数，縦軸は MAP を示す．青実線が Random Forest によって，赤破線が SVM によって学習・推測を行って作成したランキングの MAP を表している．

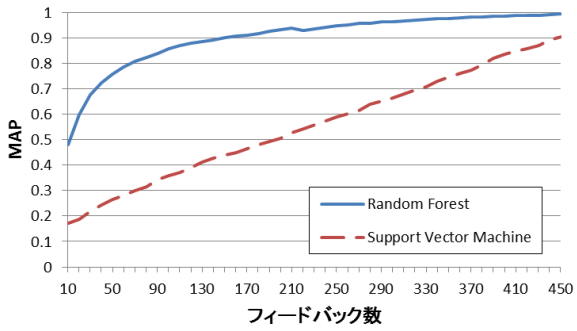


図 1 フィードバック数と検索精度

Random Forest は SVM と比較して高い再検索精度を示した．これは特に特徴量（平均 19787.3 単語）に対してデータ数が少ない状況で，SVM が 1 クラスのみに偏った出力をする事が多かった事に対して，Random Forest は比較的安定して分類を行えた事による．

結果はおおよそ右肩上がりであり，フィードバック数が増加するほど精度が増加する事を表す．これより，より良い再検索結果を得るには多くのフィードバックを得れば良いとわかる．しかし，実際に数百件もの文献を閲覧・評価する労力をユーザに求めることは現実的ではない．

そこで，本研究では高々10 件,20 件のフィードバックから機械学習手法を用いて他文献の適合・不適合を推測し，疑似的に多数のフィードバックを得る事で，ユーザに求める労力を小さいまま再検索精度を向上させる事を目指す．

また必要な疑似的なフィードバック数に関して，実験結

果からフィードバック数が大きくなるに従って精度の増加の具合は緩やかになる事，および疑似的なフィードバック数が多くなるほど誤判別によるノイズが多くなる事から，本研究においては 150 件程度とした．

### 3.3 誤りフィードバック数の影響

#### 3.3.1 実験目的

擬似的フィードバック拡充において発生し得る，ラベルの誤伝搬の最終的な出力に対する影響を検証した．また，検索課題によっては適合文献と不適合文献の数に大きな偏りが存在すること，有用な文献は類似し特徴量空間上では近距離に位置するという VSM の仮定では適合文献が密集して存在することから，適合文献と不適合文献で誤った場合の影響や誤り確率が異なると考え，それぞれの最終的な出力に対する影響を検証した．

また，適合・不適合それぞれの誤りが最終的な出力に与える影響が異なるならば，Random Forest でのクラスを決定する得票数に閾値を設けて対応する事で最終的な出力を向上させることが出来ると考え，その閾値を検討した．

#### 3.3.2 実験内容

3.1 に示した実験手順において，初期フィードバックのラベルを確率的に誤るという条件を付与する．フィードバック数 20 件のデータを用いて疑似的にフィードバック数 150 件の学習データを作成した，という状況を想定し，初期ランキングの 1~20 位の適合・不適合のラベルはそのまま，21~150 位のラベルは確率的に誤るように付与したものを学習データとした．

誤りには，(1) 適合を不適合とのみ誤る，(2) 不適合を適合とのみ誤る，(3) 区別なく誤る，という 3 つの場合を考えた．またこの時，誤り確率を徐々に変化させながら最終的な出力への影響を検証した．

#### 3.3.3 実験結果

誤り確率と検索精度の関係を図 2 に示す．横軸は誤り確率，縦軸は判別精度の値を示す．青線は適合を不適合とのみ，赤線は不適合を適合とのみ，緑線は区別なく誤った場合の判別精度をそれぞれ表す．

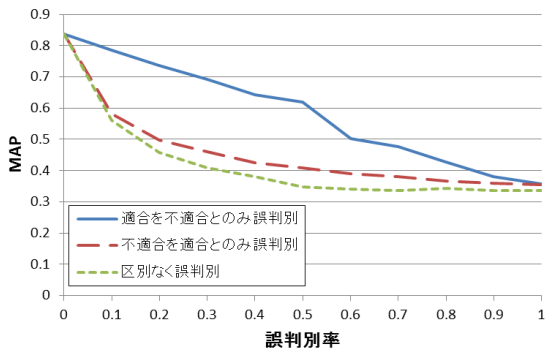


図 2　誤判別率と検索精度

3.2 より初期フィードバック数が 20 件の時の MAP は 0.59 であった．従って，たとえばラベルの誤りが適合を不適合とする誤りのみであり，その確率が 0.5 以下であれば疑似的なフィードバックを拡充する半教師学習により MAP が 0.6 以上となり精度が向上する．これより，初期フィードバックを用いた学習によって，ラベルの伝搬が高精度に行えれば，半教師学習は再検索精度向上に有効な手法だと言える．

なお，不適合文献を適合と誤る確率(偽陽性，FP：False Positive)が増加するに従って MAP が急激に減少するのに対して，適合文献を不適合と誤った確率(偽陰性，FN：False Negative)の増加に対して MAP は比較的緩やかに減少した．実際のラベルと推測したラベルの相違による最終的な再検索精度 MAP に与える影響を表 2 にまとめた．

これより特に不適合文献が適合と誤判別される割合 FP を低く抑える事が出来れば，最終的な検索精度を向上させる事が出来ると考えた．

表 2　伝播誤りによる再検索精度への影響

| | | 実際のラベル | |
|---|---|---|---|
| | | 適合 | 不適合 |
| 伝播したラベル | 適合 | 影響なし | 低下：大 |
| | 不適合 | 低下：小 | 影響なし |

そこで Random Forest の弱識別器の出力を統合するフェーズにおいて適合と推測する時の条件を厳しくする事を検討した．通常 Random Forest ではデータを各判別木にかけ，出力数が最も多いクラスに分類する．今回のような 2 クラス分類の場合には過半数の木が出力したクラスに分類される．つまり，適合と推測するには判別森から適合と出力される数が (木の本数 × 0.5) 以上である必要がある．

ここで適合と判別する際の弱識別器からの必要出力数の割合を通常の 0.5 より高く設定することにより，FP を低く抑えられると考えた．閾値を高く設定して適合と判断する確率を下げた場合，FN が増加するが，図 2 に示したように FN は FP と比較して最終的なランキングの精度を表す MAP に対する影響が小さいため，結果として MAP を向上させるはずである．

## 4. 提案手法

前述の実験により，ラベルの伝播精度を高める事で，初期検索に対するフィードバック数が少量でもラベルを伝播させた疑似的なフィードバックと併せて学習データとすることで再検索精度を向上させる示唆を得た．さらにラベル伝播精度に関して，表 2 に示したように FP は FN と比較して最終的な再検索精度に対する影響が大きく見られたた

め，適合と伝播されたものが適合である(真陽性, TP：True Positive)が高い事，言い換えれば適合でないものが適合と伝播されない確率が高い事が特に重要であると考えた．

そこで本研究では，少量のフィードバックから高精度な再検索を行うために，初期検索結果に対するユーザのフィードバックから Random Forest を用いて行った学習により，特に適合ラベル推測に制約をかけてラベルを伝播させ，拡充した疑似的なフィードバックを学習データとして用いることで推測・ランキング精度を向上させる半教師学習手法を提案する．

事前実験をふまえ，MAP が 0.9 を超えた 150 件をフィードバック数の拡充目標とした．また適合ラベル推測に関する制約に関しては，初期フィードバックから作成した判別森で適合ラベルの得票率が s (s>0.5)を超えた文献を擬似的な適合ラベル, それ以外を不適合ラベルとすることとした．

以下に手順を示す．

(1) 初期検索結果の提示・評価

ユーザが入力した問い合わせに対して初期検索結果を提示する. 初期検索結果は重要度からランキングされており，ユーザはそれを上位から順次閲覧し，有用である場合は適合，そうでない場合には不適合のラベルを付与する．

ユーザは任意数の web ページを閲覧後，情報が十分であるなら検索行動を終了する．情報が不十分かつランキング上位に有用な文献が少なく，効率的な検索が困難であると判断した場合に「再検索」メッセージをシステムに送信する．

(2) 教師データの作成

システムはユーザからのフィードバックを基に教師データを作成する．予め作成しておいた web ページと特徴量のテーブルを基に，対応する web ページの特徴とユーザ評価をラベルとしたものを統合したものを 1 つのサンプルとする．

(3) 教師データの拡充

学習データを入力として Random Forest の学習を行う．作成された判別森を用いて，初期ランキング N 位までででユーザが未閲覧の web ページの適合・不適合を推測する．この N 件は今回 150 件と設定した．

この時，判別森での適合ラベルの得票率が s (s>0.5)を超えたものを適合，それ以外を不適合とした．

(4) 再検索結果出力

各 web ページの重要度に推測されたラベルを加味し，ランキングし直したものを再検索結果として提示する．

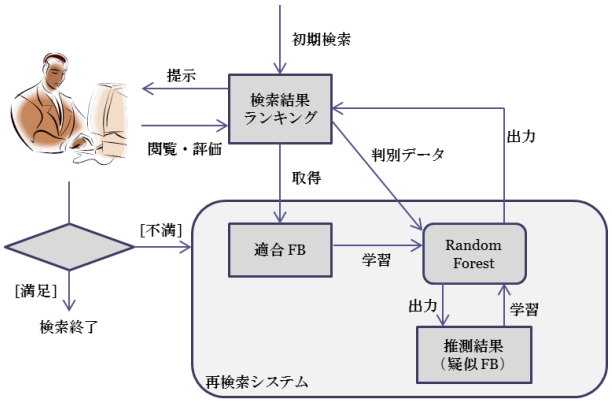

図 3　提案手法概要

以上より，ユーザは特別なキーワード入力・操作を要さず，簡易な操作から再検索が可能となる．

## 5. 実験

本論文で提案するフィードバック拡充手法の有効性を検証するために，情報検索エンジン評価のためのデータセット NTCIR4-web を用いた実験を行った．実験は初期フィードバック数 N 件から疑似的なフィードバックを加えて M 件の教師データを作成した場合のランキング精度 MAP を検証するものであり，正しくラベル付された N，M 件を学習データとした場合との比較を行った．

### 5.1 実験内容

3.1 に示した実験手順において，(5)再検索フェーズに疑似フィードバック拡張の段階を加えた．なお，初期検索結果に対するユーザのフィードバックはランキング上位 20 位までとした．

(5)以降の拡張部分を以下に示す．

(5)’ 疑似的フィードバックの拡充

(1)~(4)フェーズで得られた, フィードバック数 20 位までを教師データとして学習を行った推測結果を元に疑似的なフィードバックを拡充した．拡充のルールを以下に示す

- 1~20 位はユーザのフィードバックをそのまま利用
- 20~150 位は推測結果を基にラベルを付与
  - 不適合と推測された場合は不適合ラベルを付与
  - 適合と推測された場合は判別森が出力した適合ラベルの割合が閾値を超える場合は適合ラベル, 下回る場合は不適合ラベルを付与

なお判別森のラベル出力の割合に関する閾値は $s = \{0.5, 0.6, 0.7, 0.8, 0.9\}$ として実験を行った. s=0.5 が多数決による通常の決定方法である．

(6)’ 機械学習

拡充したフィードバック 150 件を教師データとして機械学習を行い，150~500 位までの文献の適合・不適合を推測する．

(7)’ 出力（再検索結果）

VSM に基づき，適合と推測された文献群の重心ベクトルを新しい検索クエリとして各データとのコサイン距離を算出する．各文献の重要度を，コサイン距離とラベルの差として，昇順にソードしたものを再検索結果とする．

再検索結果の精度を MAP を用いて測定する．

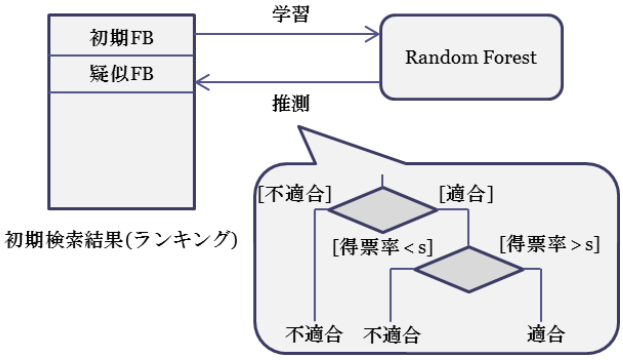


図 4 疑似的フィードバックの拡充

## 6. 結果

結果を表 3 に示す．比較として，拡充を行わない場合である初期フィードバック数 20 件を学習データとしたもの，拡充の際の推測が完全に正確だった場合である初期フィードバック数 150 件を学習データとしたものの再検索精度を示す．

表 3 教師データと再検索精度

| 教師データ | 再検索精度 |
|---|---|
| 初期フィードバック 20 件 | 0.598 |
| 初期フィードバック 150 件 | 0.901 |
| 提案手法 閾値 s =0.5 | 0.700 |
| 提案手法 閾値 s = 0.6 | 0.700 |
| 提案手法 閾値 s = 0.7 | 0.697 |
| 提案手法 閾値 s = 0.8 | 0.697 |
| 提案手法 閾値 s = 0.9 | 0.695 |

初期フィードバックを用いた推測結果を基に疑似的なフィードバックを加えて学習データ数を増やす事により，最終的な精度は向上した．しかし，拡充の基となる推測は誤りを含むため，同数の正確な教師データを用いた場合と比較して大幅に劣った．

また，期待していた適合ラベル推測に対する閾値設定による精度向上の効果は認められなかった．最終的な精度は閾値によって大きくは変化しなかった．

閾値の効果を確認するため，拡充フィードバックにおける誤判別率を検証した．フィードバック数 20 件を教師データとして実際に学習を行った時の誤判別率を表 4 に示す．ここで FN は適合を不適合と誤る確率，FP は不適合を適合と誤る確率，False は区別なく誤る確率である．また，閾値 s=0.5 は通常の判別の多数決による推測である．

表 4 実際の誤判別率

| | 閾値 | | | | |
|---|---|---|---|---|---|
| | s = 0.5 | s = 0.6 | s = 0.7 | s = 0.8 | s =0.9 |
| FN | 0.64 | 0.67 | 0.67 | 0.70 | 0.70 |
| FP | 0.059 | 0.029 | 0.029 | 0.0 | 0.0 |
| False | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |

確かに FP は減少していたが，元々が 0.06 と低い値であったために再検索精度に大きく影響しなかったと考えられる．これは使用したデータセットにおける適合文献と不適合文献の割合が平均して 0.13:0.87 と偏っており，不適合と推測される事が多かったためである．特に検索課題によっては適合文献の割合が 0.01 以下であり，20 件の初期フィードバック中に適合文献が 1 件しかないという状況があった．Random Forest はこのように偏った学習データを用いた場合，木の数が多いと過学習を起こして推測の精度が低下する傾向がある．今回の場合は全て不適合と判別するという問題が見られた．しかし，それでも最終的な精度がそこまで悪くなかったのは，今回の実験手順では推測結果は検索クエリの修正と距離の補正に使用されており，全て不適合の場合は距離の補正にデータによる差がなく，一般的な ROCCHIO アルゴリズムに類似したものとなるためだと考える．参考までに，適合文献数が特に少なかった 3 検索課題と多かった 3 検索課題の再検索精度を表 5 に示す．

提案手法による検索精度は基となるデータセットおよびそれから生成される初期学習データセットに大きく影響されるようである．なお，適合文献数が多い学習データセットを用いても閾値の影響が小さいのは，各推測の得票数が 1 またはそれに非常に近い値である事が多く，閾値の変化の幅の中になかったためである．これもまた，Random Forest において特徴量に対して教師データ数が少なく，クラスごとのデータ数が偏っている場合に見られる傾向である．

表 5 適合文献数と検索精度

| 検索課題番号 | 適合文献数 (%) | 提案手法 閾値 s = 0.5 | 提案手法 閾値 s = 0.9 |
|---|---|---|---|
| 0091 | 0.40 | 0.380 | 0.380 |
| 0001 | 1.20 | 0.574 | 0.574 |
| 0004 | 1.20 | 0.636 | 0.636 |
| 0065 | 34.0 | 0.919 | 0.636 |
| 0006 | 48.0 | 0.953 | 0.952 |
| 0082 | 59.6 | 0.960 | 0.960 |

閾値の効果を含めて誤判別率に注目した拡充手法のさらなる検証には適合文献と不適合文献の数に偏りのないデータを用いた実験が必要である．しかし，情報検索における

実際の状況で初期検索結果が適合と不適合が同数出力されるとは限らない.少なくとも本研究が目的とする再検索は，適合文献数が少ない場合に必要とされる機能である．従って，偏りのある初期検索結果から偏りのない学習データを作成する手法[16][17]が必要であると考える．

## 7. おわりに

本研究では NTCIR4-web を用いて Random Forest による適合性フィードバックの有効性を検証した．また半教師学習手法を用いた精度向上を検討し，拡充における誤判別に関して，適合文献を不適合と誤った場合と比較して不適合文献を適合と誤った場合の影響が大きかったことから，適合と推測する条件を厳しくする事で不適合文献を適合と誤る場合を抑える事で最終的な再検索結果を向上させる手法を提案した．結果，Random Forest を用いた再検索手法は SVM を使用した場合と比較して高精度なランキングを作成した．また Random Forest の推測結果を用いて疑似的なフィードバックを拡充する事により精度をさらに向上させる示唆を得た．しかし，提案した適合ラベル伝播に対する制約による再検索精度の向上効果は見られなかった．

大きな原因として，使用したデータセットにおいて適合文献の割合が小さかったため Random Forest による推測が不適合に偏ってしまったことが挙げられる．しかし，情報検索は本来大規模なデータベース中から少量の有用な文献を探索する技術であり，適合文献と不適合文献数が同数という状況は考えづらい．従って，拡充した教師データにおける適合文献数と不適合文献数が同数となるように疑似的なフィードバックを選択するなどの工夫が再検索の精度を向上させると考える．

## 参考文献

[1] J. J. Rocchio, “Relevance feedback in information retrieval.”, The SMART Retreival System, Experiment in Automatic Document Processing, PrenticeHall, pp.313-323, 1971
[2] NTCIR4 Workshop Home Page, http://research.nii.ac.jp/ntcir/ntcir-ws4/ws-ja.html. 2003
[3] 柘植 覚, 獅子堀 正幹, 黒岩 眞吾, 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”, 情報処理学会論文誌 44-1, pp.59-67, 2003
[4] L. Breiman, “Random Forests”, Machine learning 45.1, pp.5-32, 2001
[5] 池田 大介, 高村 大地, 奥村 学, ”blog 分類のための半教師有り学習”, IPSJ SIG Technical report 2008(4), pp.59-66, 2008
[6] T. Kitani et al. Lessons from BMIR-J1: A test collection for Japanese IR systems, In proc, SIGIR, pp.345-346, 1998
[7] T. Joachims, “Text Categorization with Support Vector Machine: Learning with Many Relevant Features”, ECML ’98 Proceedings of the 10th European Conference on Machine Learning, pp.137-142, 1998
[8] 高村 大地, 松本 裕治, ”SVM を用いた文書分類と構成的機能学習法”, IPSJ SIG_3(TOD_17), pp.1-10, 2003
[9] N. Graham, 森 信介, ”能動学習による効率的な情報フィルタリング”, 言語処理学会年次大会発表論文 18 巻(CD-ROM), ROMBUNNO.A4-1, 2012
[10] 金 明哲, 村上 征勝, ”ランダムフォレスト法による文章の書き手の同定”, 統計数理, Vol.55, No.2, pp.255-268, 2007
[11] M. Klassen, N. Paturi, “Web Document Classifcation by Keyword Using Random Forest.”, Networked Digital Technologies Communications in Computer and Information Science, Vol. 88, pp.256-261, 2010
[12]C. Erica and T. G. Kolda, “New Term Weighting Formulas for the Vector Space Method in Information Retreival.”, Technical Memorandum ORNL-13756, 1999
[13] 石岡 恒憲, “情報検索における信頼性評価基準について”, IPSJ SIG2(TOD13) Vol.43, pp.11-26, 2002
[14] WEKA 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/
[15] SVM$^{light}$, http://svmlight.joachims.org/
[16]D. Yao, J. Yang and X. Zhan, “An Improved Random Forest Algorithm for Class-Imbalanced Data Classification and its Application in PAD Risk Factors Analysis.”, The Open Electrical & Electronic Engineering Journal, pp.62-70, 2013