The 27th Symposium on Information Theory
and Its Applications (SITA2004)
Gero, Gifu, Japan, Dec. 14-17, 2004

複数のクエリベクトルを用いた適合性フィードバック手法
A Relevance Feedback Method using Several Query Vectors

林下 雄也* 八木 秀樹* 平澤 茂一*
Yuya HAYASHITA Hideki YAGI Shigeichi HIRASAWA

Abstract— Generally, user's interest is wide-ranged and unclear. In the information retrieval, it is difficult for a user to express his query with suitable keywords. In this paper, we propose a relevance feedback method using several query vectors. The proposed method generates clusters of the documents which the user has evaluated and creates several query vectors. By using several query vectors for retrieval, the proposed method recommends suitable documents when the user's interest is wide-ranged. Finally, we show that this method improves the retrieval accuracy compared with the conventional relevance feedback.

Keywords— information retrieval, VSM, cluster analysis, relevance feedback, several query vecters

1 はじめに

近年，ユーザが膨大な電子化された文書にアクセスできる環境が整いつつある．しかし，従来のキーワード検索では，ユーザが適切なキーワードを入力できず，ユーザの興味に合わない情報までも提示される．このため，本当に興味のある情報が埋もれてしまうという情報洪水の現象が起きている．

この問題を解決するために，ユーザに文書の判定をしてもらうことで検索結果の精度を向上させる適合性フィードバック手法の研究が盛んに行われている [1][2]．しかし，ベクトル空間を用いた従来のキーワード検索や適合性フィードバック手法はユーザの検索質問を1つのクエリベクトルでしか表現しないため，ユーザの欲しい情報がベクトル空間上で広範囲に散在している場合には，それらの情報がユーザに提示されにくいという問題点がある．

本研究では，適合性フィードバック手法において複数のクエリベクトルを用いることで，ユーザの欲しい情報が広範囲に及ぶ場合にも正しい情報が提示される手法を提案する．複数のクエリベクトルを作成する際には，クラスタ分析で用いられる凝集法を使ってユーザが適合と判定した文書をクラスタリングし，適切な数のクエリベクトルを生成する．また，この手法を情報検索のベンチマークデータ (BMIR-J2 [3]) に適用し，この手法が従来の適合性フィードバック手法に比べ検索精度を向上させることを示す．

2 準備

本論文では，検索モデルとして情報検索の分野でよく用いられるベクトル空間モデル (VSM) を用いる．以下で，VSM の説明を行う．

*〒 169-8555 東京都新宿区大久保 3-4-1 早稲田大学理工学部経営システム工学科 School of Science and Engineering, Waseda University Okubo 3-4-1, Shinjuku-ku, Tokyo, 169-8555 Japan. E-mail: hayashita@hirasa.mgmt.waseda.ac.jp

2.1 VSM [1]

VSM では，文書とユーザの検索質問をベクトルとして表現する．これによって，各文書がどれくらい検索質問に適しているかをベクトル間の類似度に帰着することができる．このベクトル空間は単語ごとに独立した次元を持ち，文書や検索質問が各次元で持つ値は，単語が文書や検索質問で持つ重みを表す．

文書ベクトルの要素の重みは TF·IDF 法 [4] を用いて計算する．TF·IDF 法では，文書データベース中の多くの文書に出現する語は重要でなく，特定の文書において多く出現する語は重要とすることで単語の重みを決定するものである．また，ベクトルどうしの類似度は両ベクトルの余弦によって計算する．

定義 1 (文書 d_j 中の単語 t_i の TF·IDF 値：w_{ij})

$$w_{ij} = \frac{tf(t_i, d_j)}{\sum_{t \in d_j} tf(t, d_j)} \cdot \left(\log_{10} \frac{M}{df(t_i)} + 1 \right) \tag{1}$$

$tf(t_i, d_j)$：文書 d_j 内の単語 w_i の出現頻度
M：文書データベース中の文書総数
$df(t_i)$：単語 t_i が出現する文書数

定義 2 (文書 d_j の文書ベクトル：d_j)

$$\boldsymbol{d}_j = (w_{1j}, w_{2j}, \ldots, w_{nj}) \tag{2}$$

n：システム中の全単語数

定義 3 (検索質問ベクトル (クエリベクトル)：q)

$$\boldsymbol{q} = (w_1^q, w_2^q, \ldots . w_n^q) \tag{3}$$

$$w_i^q = \begin{cases} 1, & \text{単語 } t_i \text{ が検索キーワードである;} \\ 0, & \text{単語 } t_i \text{ が検索キーワードでない.} \end{cases} \tag{4}$$

定義 4 (d_j と q の類似度：$Sim(d_j, q)$)

$$Sim(d_j, q) = \frac{d_j \cdot q}{|d_j||q|} \tag{5}$$

$d_j \cdot q$：ベクトル d_j とベクトル q の内積
$|d_j|$：ベクトル d_j のノルム

3 適合性フィードバック手法 [2]

本節では，検索システムの検索精度を向上させる手法の1つである適合性フィードバック手法について示す．

3.1 適合性フィードバック手法の概要

検索結果の精度を向上させるために，ユーザに検索結果を提示し，ユーザがその結果を見てシステムの挙動を変化させるようにシステムのパラメータを調整する技術を一般に適合性フィードバック (relevance feedback) という．

適合性フィードバック手法の中で最も代表的な手法としては，システムが提示する検索結果の上位の文書をユーザに適合か不適合かの判定をさせ，その情報をもとに検索質問を修正するものである．

3.2 Rocchio フィードバック

VSM の適合性フィードバック手法においてクエリベクトルの要素の重みを修正する手法はいくつか提案されている．その中でも最も幅広く利用されているのが Rocchio フィードバックである [2]．

q_i を検索 i 回目のクエリベクトルとする．Rocchio フィードバックでは，以下の式によってクエリベクトルを修正する．

$$q_{i+1} = q_i + \frac{1}{||D^+||}\sum_{d_j \in D^+} d_j - \frac{1}{||D^-||}\sum_{d_j \in D^-} d_j \quad (6)$$

D^+：適合文書集合
D^-：不適合文書集合
$||A||$：集合 A に含まれる要素数

この式は，適合文書に含まれる単語の重みは大きく，不適合文書に含まれる単語の重みは小さくなるようにクエリベクトルの要素の重みを調整している．

以下に，Rocchio フィードバック手法のシステムの手順を示す．

[Rocchio フィードバック手法]

s1) ユーザが検索質問をシステムへ入力する．
s2) システムは，検索質問からクエリベクトルを作成し，各文書との類似度を計算して文書のランキングをユーザに提示する．
s3) ユーザはいくつかの文書に関して適合・不適合の判定を行い，その結果をシステムに返す．
s4) システムは，フィードバック情報をもとに式 (6) を用いてクエリベクトルを修正する．
s5) システムは，修正したクエリベクトルを用いて文書との類似度を再計算し，新しい文書のランキングをユーザに提示する． □

4 凝集法

本節では提案手法で用いるクラスタ分析の代表的な手法の 1 つである凝集法について示す．

4.1 クラスタ分析の概要

クラスタ分析とは，データ解析の諸技法の中で，外的基準なしに自動的に分類を行う方法，いいかえれば，データ以外にあらかじめ基準を設定することなく，データの集まりをいくつかのグループ (クラスタ) に分ける方法のことである．

クラスタ分析には階層的クラスタリングと非階層的クラスタリングがある．非階層的クラスタリングの代表的な手法に K-means 法があり，階層的クラスタリングの代表的手法に凝集法がある [5]．

4.2 凝集法のアルゴリズム [5]

提案手法では，クラスタ数を事前に定めずにクラスタリングを行い，クラスタリングを行った後にクラスタ数を決めるため，凝集法を用いる．以下に，凝集法のアルゴリズムを示す．

[凝集法アルゴリズム]

s1) N 個の個体 (オブジェクト) について 1 個のオブジェクトを 1 個のクラスタとして，クラスタ間の距離を計算し，類似度行列を作成する．
s2) 類似度行列の中で類似度が最大の 2 つのクラスタを併合し，1 つのクラスタを作る．
s3) 併合後のクラスタと他のクラスタとの類似度を計算し，類似度行列を更新する．
s4) step2 と step3 を繰り返し，クラスタ数が 1 になれば終了． □

4.3 距離の定義と併合ルール

凝集法では，オブジェクト間の距離の計算方法と併合ルールを適切に選択する必要がある．距離の計算方法としてはユークリッド距離，マハラノビス距離，マンハッタン距離，ベクトル間の余弦 (式 (5)) などがある．

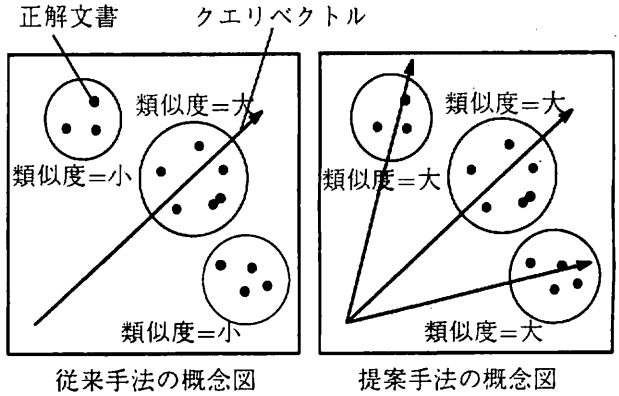
併合ルールとは，1 度いくつかのオブジェクトが一緒に併合された後，新しいクラスター間の距離をどのように定義するかというルールである．これには，2 つのクラスター間の距離をそれぞれのクラスター内の最も近いオブジェクト間の距離とする最近隣法や，2 つのクラスター内のすべてのオブジェクトのペアの距離の平均として定義する群平均法，クラスタ間の距離を各クラスタの重心同士の距離とする重心法などがある．

5 提案手法

5.1 従来手法の問題点

VSM における従来のキーワード検索や適合性フィードバック手法では，ユーザの検索質問を 1 つのクエリベクトルで表現し，それを用いて各文書との類似度を計算して文書のランク付けを行う．

しかし，文書の特徴をベクトル空間で表現した場合，ユーザが正解であるとする正解文書はある 1 つの場所にまとまって存在しているのではなく，いくつかの散在した場所に固まって存在していることが多いと考えられる．このような場合に 1 つのクエリベクトルを用いて文書との類似度を計算すると，ある場所に存在する正解文書に対しては類似度は高くなるが，そのほかの場所に散在している正解文書の類似度は低くなってしまう (図 1)．



従来手法の概念図　　提案手法の概念図

図 1: 従来手法と提案手法の概念図

5.2 提案手法の概要

前節の問題を解決するための方法として，正解文書が存在するいくつかの場所にそれぞれクエリベクトルを作成し，それらのクエリベクトルを用いて各文書との類似度を計算することで，散在している文書の類似度を大きくすることが考えられる (図 1)．

そこで本研究では，ユーザが判定した適合文書からの情報をもとに適切な複数のクエリベクトルを作成し，それらのクエリベクトルを用いて検索を行う手法を提案する．

その結果，正解文書が存在するであろう様々な位置にクエリベクトルを作成し，各クエリベクトルに近い文書の類似度を大きくすることで，散在している正解文書のランキングを上げることを実現する．そこで本研究では，クエリベクトルの集合 Q に対する文書の類似度を以下のように定義する．

定義 5 (d_j と Q の類似度：$Sim'(d_j, Q)$)

$$Sim'(d_j, Q) = \max_{q'_i \in Q} Sim(d_j, q'_i) \tag{7}$$

q'_i：Q 中の i 番目のクエリベクトル

類似度を式 (7) で求めると，複数のクエリベクトルの中でその文書と最も近いクエリベクトルとの類似度が文書の類似度となる．

5.3 複数のクエリベクトルの作成とクエリベクトル数の決定

VSM において正解文書の分布を判断することは多次元空間のため難しい問題である．また，検索課題によって正解文書の分布は異なってくる．そこで本手法では，ユーザが適合であると判定した文書をもとに異なる数のクエリベクトルを含むクエリベクトル集合をいくつか作成し，その中で最も適当であると考えられるクエリベクトルの集合を用いて検索を行う．ここで，いくつかのクエリベクトルの集合を作成するため，凝集法を用いて適合文書をクラスタリングする．以下に，提案手法のクエリベクトル集合の作成とクエリベクトル数を決定するアルゴリズムを示す．

[クエリベクトル生成アルゴリズム]

s1) 各適合文書を 1 つのクラスタとする．
s2) 各クラスタの重心ベクトルの集合をクエリベクトルの集合とおく．
s3) クエリベクトルの集合を用いて不適合文書との類似度を計算し，類似度の和を求める．
s4) 最も距離の近い 2 つのクラスタを 1 つのクラスタとする．
s5) step2 から step4 をクラスタ数が 2 つになるまで繰り返し，クラスタ数が 2 つになれば step5 に移る．
s6) 不適合文書との類似度の和が最小のときのクエリベクトルの集合を次の検索のクエリベクトルの集合とする． □

ここで，本研究では複数のクエリベクトルを用いる手法を考えているため，クラスタ数が 2 の時点でクラスタリングを終了する．例えば適合文書が 5 文書だとすると step2 でクエリベクトル数が 5 個の集合が作成され，それから 4 個の集合，3 個の集合，2 個の集合が作成される．そして，それぞれのクエリベクトル集合と不適合文書との類似度の和を計算し，その和が最小のときのクエリベクトルの集合が次の検索に用いられる．

本手法では，適切なクエリベクトルの数を決定するのに不適合文書との類似度の和が最小という基準を用いた．これは，各クエリベクトル集合と適合文書の類似度を計算すると，step1 で求められるクエリベクトル集合が適合文書の特徴ベクトルであるため，各適合文書の類似度の最大値が 1 となり常にこのときの類似度の和が最大となってしまうからである．

上記の手順において，クエリベクトルの集合と不適合文書の類似度は式 (7) で求める．また，凝集法におけるオブジェクト間の距離尺度としてベクトルの余弦を用い，併合ルールとして重心法を用いる．

6 シミュレーションと考察

提案手法の有効性を示すために 2 つの実験を行う．

6.1 シミュレーション条件

文書集合には，毎日新聞 CD-ROM'94 データ版 [6] を基に構築した情報検索システム評価用テストコレクションである BMIR-J2 [3] を用いた (5080 文書，抽出された名詞約 25500 語)．検索課題は，BMIR-J2 が提供する検索課題のうち正解件数が偏らないよう考慮して 9 課題を選んだ．また各課題に対して，BMIR-J2 が提供する A，B 両レベルの正解を正解文書とした．

6.2 実験 1：検索性能の評価

従来手法である Rocchio フィードバックと提案手法の検索精度を比較する．検索精度を測る評価値として，検索結果の適合率と再現率を用いる．さらに，再現率が 0.0,0.1,0.2, …,1.0 のときの適合率である 11 点適合率を求め，その適合率の平均である 11 点平均適合率を算出する．適合率，再現率の計算は以下の式で行う．

$$\text{適合率} = \frac{\text{検索結果文書中の正解文書数}}{\text{検索結果文書数}} \tag{8}$$

$$\text{再現率} = \frac{\text{検索結果文書中の正解文書数}}{\text{正解文書数}} \tag{9}$$

6.3 実験 1 の結果と考察

フィードバック文書数を 30 文書，フィードバック回数を 1 回としたときの 11 点適合率を求め，その適合率・再現率グラフを図 2 に示す．

表 1: 11 点平均適合率の比較

feedback 文書数	Rocchio feedback 数			提案手法 feedback 数
	1 回	3 回	5 回	1 回
10 文書	0.537	0.610	0.652	0.734
20 文書	0.543	0.625	0.680	0.756
30 文書	0.541	0.626	0.681	0.777

また，フィードバック文書数が 10 文書，20 文書，30 文書のときのフィードバック回数による Rocchio フィードバックの 11 点平均適合率と，提案手法のフィードバック回数が 1 回のときの 11 点平均適合率との比較を表 1 に

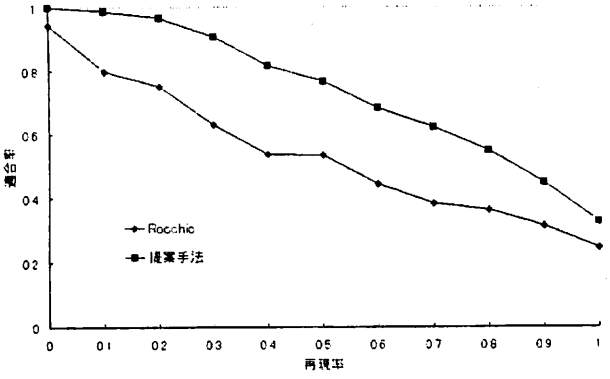


図 2: 適合率再現率グラフ (フィードバック文書数 30 文書)

示す．図 2 中の適合率と表 1 の 11 点平均適合率は全検索課題を平均したものである．図 2，表 1 いずれも Rocchio フィードバックよりも提案手法の方が検索精度がよくなっていることが分かる．特に表 1 では，フィードバック回数が 5 回のときの Rocchio フィードバックの 11 点平均適合率よりもフィードバック回数が 1 回のときの提案手法の 11 点平均適合率が優っている．このことから，提案手法は少ないフィードバック回数で精度の高い検索結果を出すことができることが分かる．これは，複数のクエリベクトルを用いることによって散在している正解文書の類似度を 1 回のフィードバックで大きくすることができた結果だと考えられる．

次に，従来手法と提案手法の計算時間を比較する．提案手法は従来手法と比べて，適合文書のクラスタリングと，クエリベクトル集合と文書の類似度を計算する際の計算時間が増加する．しかし，クラスタリングは適合文書数 −2 回だけ行われ，クエリベクトル数も最大で適合文書数となり，適合性フィードバック手法において適合文書数は大きな数ではないため，提案手法の計算時間は実現可能な範囲であると考えられる．

6.4 実験 2：クエリベクトル数の評価

提案手法が適切なクエリベクトルの数を選択できているかどうかについて調べるために，以下の実験を行う．まず，クラスタリングの各段階で作成された各クエリベクトル集合を全文書に対する検索に用いて，そのときの 11 点平均適合率を求める．そして，提案手法で選択されたクエリベクトル集合の 11 点平均適合率が何位なのかを調べ，選択されたクエリベクトルの数と 11 点平均適合率が最大になるときのクエリベクトルの数を調べる．

6.5 実験 2 の結果と考察

フィードバック文書数が 10 文書，20 文書，30 文書のときのシミュレーション結果を表 2 に示す．表 2 において，平均適合率の順位とは提案手法で選択されたクエリベクトル集合の 11 点平均適合率の順位である．また，表 2 の値は，全て全検索課題の平均値である．

表 2 より，フィードバック文書数が増えるにしたがって適合率最大のときのクエリベクトル数が増えていることがわかる．これは，フィードバック文書数が増えるだけ適合文書の数も増え，検索課題にあった適切なクエリベクトルが生成されているからだと考えられる．

表 2: クエリベクトル集合に関する結果

	10 文書	20 文書	30 文書
平均適合率の順位	2.00	3.89	4.67
提案手法が選択したクエリベクトル数	2.56	2.89	2.33
適合率が最大の時のクエリベクトル数	2.67	4.67	6.67

一方，提案手法で選択したクエリベクトルの数はすべて 3 以下となっている．これは，クエリベクトル集合を選ぶ際の指標としてクエリベクトル集合と不適合文書との類似度の和を用いていることが原因であると考えられる．つまり，一般的に不適合文書は適合文書よりも様々な場所に分布しており，クエリベクトルの数が多くなると不適合文書との和が大きくなってしまうため，クエリベクトルの数が少ないときが選択されたと考えられる．そのため，フィードバック文書数が増えると，最良のクエリベクトル数と提案手法で選択されたクエリベクトル数に差が生じてしまい，11 点平均適合率の順位が下がっていると考えられる．

7 まとめと今後の課題

本研究では，適合性フィードバック手法の適合文書をクラスタリングし，適切な複数のクエリベクトルを作成することによって，散在した正解文書の類似度を大きくする手法を提案した．また，ベンチマークを適用したシミュレーションにより提案手法の有効性を示すことができた．

しかし，提案手法では検索精度が最良となるときのクエリベクトル集合を選択することができていないため，クエリベクトル集合を選択する際の指標について今後検討する必要がある．その際，提案手法では 1 度クラスタ数が 2 個になるまでクラスタリングを行い，その後にクエリベクトルの数を決定しているため，最良のクラスタ数が 2 個でない場合は無駄な計算をしているといえる．そこで，1 度評価値が下がったときにはそれ以上のクラスタリングを行わないでいいような評価方法が望ましい．また，適合文書をクラスタリングする際に重心法以外の結合ルールを用いた場合や，フィードバック回数を増やしたときの評価も検討する必要がある．

8 謝辞

著者の一人である林下は，本研究を行うにあたり，数多くのご助言，ご支援を賜りました早稲田大学平澤研究室の各氏に感謝いたします．

参考文献

[1] 徳永健伸，情報検索と言語処理，財団法人東京大学出版会，1999 年．
[2] Rocchio, J., "Relevance Feedback in Information Retrieval", *The SMART Retrieval System Experiments in Automatic Document Processing*, Prentice Hall Inc, 1971.
[3] (社) 情報処理学会データベースシステム研究会，BMIR-J2 テストコレクション．新情報処理開発機構
[4] Salton, G. and Buckley, C., "Term-Weighting Approaches in Automatic Text Retrieval", *Information Processing Management*, 24(5), pp513–523, 1998.
[5] 宮本定明，クラスター分析入門，森北出版株式会社，1999.
[6] 毎日新聞社，CD-毎日新聞'94 データ集．日外アソシエーツ．