社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－NL－161 (7)
2004／5／14

適合フィードバックにおける単語選択法

藤野昭典 上田 修功
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

概要: 確率モデルによるテキスト情報検索での適合フィードバックにおいて，問い合わせに用いることで検索性能の向上が期待できる単語を選択する手法を検討した．提案手法では，フィードバックされた適合文書から leave-one-out に基づく交差法と偏差法の 2 つの手法で各単語が適合文書に安定的に含まれる可能性を評価し，単語の選択順位を決定する．交差法と偏差法のそれぞれの手法で得られる単語の選択順位を，3 つのテストコレクション (MED, CRAN, CISI) を用いて適合フィードバック後の検索性能を調べることで評価した．選択する単語数を変えて検索性能を調べた結果，両手法ともに選択する単語数が少ない場合においても高い検索性能が得られた．

Term Selection Method on Relevance Feedback

Akinori FUJINO Naonori UEDA
NTT Communication Science Laboratories, NTT Corporation

Abstract: In this report, we address term selection problem for relevance feedback. Two ranking criteria for term selection are proposed; leave-one-out cross validation based criterion and deviation-based criterion. The first criterion is computed by validating term frequencies for a relevant document with term probabilities for the other relevant documents. The second criterion is computed by using the mean and standard deviation of term probabilities for relevant document subsets. The experimental results using three test collections (MED, CRAN, CISI) showed that the two criteria were able to select useful terms from relevant documents to improve average precision.

1 はじめに

テキスト情報検索では，ユーザの問い合わせ (query) と検索対象の文書に含まれる単語との類似性に基づいて検索順位を決定し，ユーザに検索結果を提示する [7, 2]．しかし，自然言語には同一の概念をもつ単語が数多く存在するという同義性の問題がある．これは，同じ意味でありながら問い合わせと文書に含まれる単語が異なることで，検索すべき文書が検索されない問題を引き起こす．また，単語には複数の概念をもつものも多い．このため，異なる意味で用いられながら問い合わせと同じ単語が検索対象文書で用いられることにより，ユーザの要求とは異なる文書を検索する可能性がある．

このような自然言語の特徴から生じる情報検索の問題を解決する一手法として，適合フィードバック (relevance feedback)[6, 4, 1] がある．適合フィードバックは，ユーザの検索文書に対する適合・不適合の判定情報を用いることで検索性能を高める技術である．適合フィードバックでは，ユーザが適合と判定した文書に含まれる単語を用いて，問い合わせへの単語の追加や重みの更新を行う．例えば，適合と判定された文書に含まれる単語が問い合わせに追加される．適合フィードバックにより，ユーザによる問い合わせに含まれる単語と同義の異なる単語や，検索対象の他の適合文書に含まれる単語が問い合わせに追加されれば，ユーザによる問い合わせのみでは検索できない適合文書の検索が可能となる．また，単語の追加により，問い合わせがもつ複数の概念が特定され，ユーザが要求しない文書が上位に検索される比率を下げることが期待できる．

一方，適合文書に含まれる単語のすべてが，ユーザによる問い合わせの単語と関連があるとは限らない．例えば，ユーザが適合と判定した文書に，ユー

ザが要求する情報と関連の低い記述が含まれる場合，それらの記述に含まれる単語はユーザの問い合わせと関連性が低い．このような関連性の低い単語は，問い合わせに追加されることで逆に検索性能を悪化させる可能性がある．

本研究では，ユーザが適合と判定した文書から問い合わせに追加すべき単語の優先順位を与える手法を検討する．従来研究では，適合と判定された文書集合での単語の出現頻度や単語の出現文書数といった統計情報をもとに単語を選択する手法が用いられてきた [1, 3, 9]．しかし，ユーザからフィードバックされる文書数は一般的に少なく，必ずしもフィードバックされた適合文書の統計情報が検索対象の適合文書の統計情報を近似しているとは限らない．このため，検索に効果的な単語を選択するには，フィードバックされた適合文書から検索対象の適合文書にも含まれるような汎化性の高い単語を推定することが課題となる．そこで本研究では，フィードバックされた適合文書を leave-one-out で交差確認 (cross validation) することにより，文書を適合と判定するのに有効な単語を選択するアプローチをとる．本アプローチでは，どの適合文書にも安定的に高頻度で出現する単語を選択することが期待できる．また，このような単語は検索対象の適合文書にも多く含まれる可能性が高いと考えられる．

単語の選択順位の決定方法として，交差法を検索モデルに応用する方法 (交差法) と，出現頻度の偏差に基づく方法 (偏差法) を考案した．交差法では，ある適合文書を除去して適合フィードバックを行ったときに，その適合文書が上位に検索されるように問い合わせに追加する単語を選択する．偏差法では，適合文書に安定的に出現する確率を単語出現確率の平均と標準偏差をもとに定義し，その確率が上位の単語を選択する．本稿では，これら 2 つの方法を，フィードバックされた適合文書全体での出現確率をもとに単語を選択する方法と比較することで，その有効性を検証する．

2 検索モデル

本研究では，適合フィードバックに基づいて文書の検索順位を決定するための尺度として，確率モデル (Probabilitstic Model)[5] に基づく尺度を用いる．確率モデルに基づくテキスト情報検索では，文書の検索順位を決定する尺度のひとつとして，確率比 (Probability Ratio)[2] が提案されている．ある文書 n の確率比 $PR(n)$ は，適合クラスと文書集合全体での単語の出現確率 θ_{ri}, θ_{Gi} をもとに，以下の式で定義される．

$$PR(n) = \sum_{i=1}^{V} tf(n,i) \log \frac{\theta_{ri}}{\theta_{Gi}} \qquad (1)$$

式 (1) 中の $tf(n,i)$ は文書 n での単語 i の出現頻度，V は文書集合群での単語の種類の総数を表す．文書の確率比は文書集合全体より適合クラスでの出現確率が高い単語を多く含む場合に大きくなる．確率比の大きさが各文書のユーザの検索要求に適合する度合を表すとして，確率比の大きな文書を上位に検索する．適合フィードバックによる文書検索では，適合クラスの単語の出現確率はフィードバックされた適合文書に含まれる単語の出現頻度をもとに見積もる．単語の出現確率は，文書集合全体 G と適合 r の各クラス $c = G, r$ に対して，Naive Bayes モデルによって以下のように求める．式 (2) 中の ξ_i は平滑化パラメータを表す．

$$\theta_{ci} = \frac{\sum_{n \in c} tf(n,i) + \xi_i}{\sum_{i=1}^{V} \left\{ \sum_{n \in c} tf(n,i) + \xi_i \right\}} \qquad (2)$$

しかし，確率比 $PR(n)$ は長い文書であるほど，取り得る値の範囲が大きくなる．このため，長さの異なる文書の順位付けを正確に行えない可能性がある．そこで，文書長による影響を避けることを目的として，式 (1) に正規化項 $Z(n)$ を加えた以下の尺度 $NPR(n)$ を新たな尺度として定義する．

$$NPR(n) = \frac{1}{Z(n)} \sum_{i=1}^{V} tf(n,i) \log \frac{\theta_{ri}}{\theta_{Gi}} \qquad (3)$$

$$Z(n) = \sqrt{\sum_{i=1}^{V} tf(n,i)^2} \sqrt{\sum_{i=1}^{V} \left(\log \frac{\theta_{ri}}{\theta_{Gi}} \right)^2}$$

$Z(n)$ の第 2 の正規化項は $NPR(n)$ の絶対値を 1 以下にする効果があるが，文書 n に依存しないため検索順位の決定には関与しない．

この尺度では，文書集合全体で出現する単語すべてを用いて文書が順位付けされる．これは，すべての単語をフィードバック情報をもとに重み付けを行った上で問い合わせとして用いることに相当する．し

かし，フィードバックされた適合文書に文書集合全体で出現する単語がすべて含まれているとは限らない．このような適合文書に含まれない単語は，検索順位を決定する上で有効でない可能性が高いと考えられる．適合文書に含まれない単語は，式 (1), (3) の文書の確率比を負にするように働き，それらの単語を多く含む文書ほど上位に検索されない．しかし，フィードバックされた適合文書に含まれない単語が，必ずしもユーザの問い合わせに内在する概念と関連がないとは限らず，そのような単語を含む文書がユーザが要求する文書である可能性がある．また，適合文書に含まれる単語のすべてが，新たに適合文書を検索するのに有効であるとは限らない．そこで，フィードバック情報をもとに生成する新たな検索順位の尺度として，問い合わせに用いる単語を限定した尺度を用いる．この尺度は，式 (3) に 0,1 いづれかの値をとる α_i を導入して定義する．

$$NPR'(n) = \frac{1}{Z(n)} \sum_{i=1}^{V} \alpha_i tf(n,i) \log \frac{\theta_{ri}}{\theta_{Gi}} \qquad (4)$$

ここで，α_i は文書検索に利用する単語を選択するパラメータであり，$\alpha_i = 1$ である単語の対数確率比が検索対象文書の順位付けに用いられる．$NPR'(n)$ を用いて文書検索を行うには，α_i を決定することが必要となる．次節では，$NPR'(n)$ の α_i を決定する手法を述べる．

3 単語選択法

フィードバック情報を利用する文書検索において，検索順位の尺度 $NPR'(n)$ に用いる単語の選択法として，交差法と偏差法の 2 つの手法を考案した．

3.1 交差法

leave-one-out 交差法は，例えば 2 クラスの文書の分類問題では，ある 1 つの文書を除いて分類器を学習したときに，除いた文書が正しく分類されるかどうかを確認することで，分類器の汎化能力を検証したり，汎化能力を向上させるためのパラメータを調整するのに用いられる．交差法を用いた単語の選択では，フィードバックされたある適合文書の確率比を，他の適合文書を用いて学習した検索順位の尺度で見積るときに，その確率比を大きくするのに寄与している単語を選択する．すなわち，フィードバックされた適合文書の一部の情報を用いて検索を行うときに，残りの適合文書を上位に検索する効果がある単語を見出すことを基本的な考え方としている．交差法による方法では，1 つの文書 n を除いて適合クラスの単語出現確率 $\theta_{ri}^{(-n)}$ を学習するとき，その出現確率を用いた尺度で見積られる文書 n の確率比を大きくする単語 i を求める．フィードバックされたすべての適合文書の確率比を一様に大きくする単語 i は，以下の式で定義する単語の寄与度 β_i に基づいて選択する．

$$\beta_i = \sum_{n \in r} \frac{1}{Z(n)} tf(n,i) \log \frac{\theta_{ri}^{(-n)}}{\theta_{Gi}} \qquad (5)$$

式 (5) より，β_i は単語ごとに適合文書における対数確率比を加算したものであり，β_i が大きな単語ほどそれぞれの適合文書の確率比を高くするのに寄与する．交差法では，選択する単語数が N_t のとき，尺度 $NPR'(n)$ のパラメータ α_i を寄与度 β_i が大きな N_t 個の単語で 1，それ以外の単語で 0 として決定する．

3.2 偏差法

偏差法では，交差法と同様にまずフィードバックされた適合文書から 1 つの文書 n を除いて適合クラスの単語出現確率 $\theta_{ri}^{(-n)}$ を推定する．すべての適合文書に関してこの単語出現確率を求めて式 (3) の θ_{ri} に適用すると，検索順位の尺度の候補をフィードバックされた適合文書の数 N だけ得ることができる．この N 個の尺度を用いて文書の確率比を見積るとき，どの尺度を用いても安定的に文書の確率比を大きくするのに寄与している単語を適合性の高い単語として選択する．単語 i の対数確率比 $ratio\_n(i) = \log \frac{\theta_{ri}^{(-n)}}{\theta_{Gi}}$ の n に対する分布が正規分布に従うと仮定すると，単語の対数確率比 $ratio\_n(i)$ が正になる確率は，

$$norm(i) = \frac{m(ratio\_n(i))}{\sigma(ratio\_n(i))} \qquad (6)$$

が大きくなるほど高くなる．ここで，$m(ratio\_n(i))$ は n に対する単語 i の対数確率比 $ratio\_n(i)$ の平均を，$\sigma(ratio\_n(i))$ は標準偏差を表す．対数確率比が正になる確率が高い単語は，その単語を含む文書の確率比を大きくするのに寄与する可能性が高いことを意味する．そこで偏差法では，選択する単語数が N_t のとき，尺度 $NPR'(n)$ のパラメータ α_i を $norm_i$ が大きな N_t 個の単語で 1，それ以外の単語で 0 として決定する．

表 1: テストコレクション

	MED	CRAN	CISI
文書数	1033	1400	1460
単語数	5462	4000	5041
問い合わせ数	30	219	74
平均適合文書数	23.2	8.3	42.1

表 2: 単語選択なしの場合の平均適合率

$N_{feedback}$	尺度	MED	CRAN	CISI
10	$NPR(n)$	0.30410	0.12843	0.09729
	$PR(n)$	0.11240	0.08346	0.06728
20	$NPR(n)$	0.31208	0.13470	0.08875
	$PR(n)$	0.13848	0.09613	0.06013
30	$NPR(n)$	0.30332	0.12492	0.09100
	$PR(n)$	0.14910	0.08719	0.06343

4 評価実験

交差法と偏差法の 2 つの単語選択法の評価実験には，テキスト情報検索の性能評価に用いられる MED, CRAN, CISI の 3 つのテストコレクションを用いた．これらのコレクションは，抄録からなる文書集合であり，評価用の問い合わせと，問い合わせとどの抄録が適合するかという適合情報が含まれている．評価実験では，各コレクションの抄録を単語の出現頻度ベクトルで表して検索に用いた．単語の出現頻度ベクトルを求める際，SMART システム [8] で適用された停止語リスト (stop word list) を用いて，冠詞など文書を特徴づける効果をもたない単語を取り除いた．また，各コレクションにおいて 1 つの抄録のみに出現する単語を低頻度語として取り除いた．各コレクションに含まれる文書数，単語数，問い合わせの数ならびに問い合わせに適合する平均文書数を表 1 に示す．

適合フードバックに用いる文書は，もとの問い合わせの各単語にコレクション全体での idf で重み付けを行ったものと文書の単語出現頻度ベクトルのコサイン類似度を求めて決定した．コサイン類似度が大きな上位 N 個の文書に対してユーザが適合・不適合の判定を行った結果をもとに検索順位の尺度を生成した．ユーザの判定は，コレクションに含まれる適合情報通りに行われるとした．フィードバックされる文書数 $N_{feedback}$ は 10,20,30 の 3 通りとした．実験には，2 つ以上の適合文書がフィードバックされ，かつ残りの検索対象文書に適合文書が含まれている問い合わせを用いた．フィードバック後，残りの検索対象文書から適合文書をどれだけ上位に検索できるかで評価した．性能評価には，情報検索の評価でよく用いられる 11 点平均適合率 [10] を，各コレクションの問い合わせで平均した値を用いた．11 点平均適合率は，0 から 1 までの 0.1 刻みの 11 の再現率レベルにおける適合率を加算平均したものである．平均適合率が大きいほど，適合文書を上位に検索する能力があることを示す．

表 2 は，式 (1), (3) で示した，各コレクションに含まれるすべての単語を用いた尺度 $PR(n)$, $NPR(n)$ を用いて検索順位を決定したときの平均適合率を示す．尺度に含まれる各単語の確率比は，式 (2) の平滑化パラメータ ξ_i を 0 として θ_{Gi} を，$\xi_i = \theta_{Gi}$ として θ_{ri} を推定することで求めた．θ_{ri} はフィードバックされた適合文書すべてとユーザの問い合わせの単語出現頻度を用いて推定した．表 2 より，正規化した尺度 $NPR(n)$ で検索性能が高い傾向があった．

つぎに，交差法と偏差法によって検索順位の尺度に用いる単語を選択した場合の結果を示す．これらの 2 つの方法は，フィードバックされた適合文書のすべてで推定した単語の確率比が大きいものから順に単純に単語を選択する方法と検索性能を比較することで有効性を調べた．性能評価に際して，検索には式 (4) の尺度 $NPR'(n)$ を用い，尺度内の θ_{ri} と θ_{Gi} は上述の $PR(n)$, $NPR(n)$ による実験と同様の方法で推定した．単語の選択を意味する α_i は各方法でそれぞれ決定した．

4.1 交差法の評価

交差法では，寄与度 β_i の大きな N_t 個の単語の α_i を 1 とすることで，検索順位の尺度 $NPR'(n)$ を決定する．β_i は，式 (2) の平滑化パラメータ ξ_i を MED, CRAN, CISI の各コレクションで 0.1, 0.05, 0.05 として $\theta_{ri}^{(-n)}$ を推定し，式 (5) を用いて求めた．$\theta_{ri}^{(-n)}$ はフィードバックされた適合文書と問い合わせの単語出現頻度を用いて推定した．

評価実験では，交差法による寄与度 β_i を用いて α_i を決定する場合と，単純に式 (4) 中の単語の対数確率比 $ratio(i) = \log \frac{\theta_{ri}}{\theta_{Gi}}$ が大きな単語の α_i を 1 にする場合の検索性能を比較した．2 つの方法でそれぞれ同数の単語を選択し，その単語を用いて尺度を生成したときの検索性能を求めた．選択する単語の数 N_t は，$\beta_i > 0$ となる単語数 N_β と対数確率比

が $ratio(i) > 0$ となる単語数 N_r を基準として

$$N_t = (1-\gamma)N_\beta + \gamma N_r \quad (7)$$

から決定した．N_β と N_r には $N_\beta < N_r$ の関係があり，γ が大きいほど選択される単語数 N_t が多くなる．実験では，それぞれの問い合わせで，γ を-0.25 から 1.5 の範囲で 0.25 刻みで変えて尺度を生成し，検索結果に対する平均適合率を求めた．

表 3 は，各問い合わせに対する平均適合率を γ ごとに平均した結果を示す．$gamma$ が同じときの交差法と確率比による選択法との平均適合率の差は，同数の単語を尺度に用いたときの検索性能の差を示す．表より，どのコレクションに対しても，γ が小さいときに交差法でより検索性能が高く，γ が大きくなるにしたがって検索性能の差が縮まる傾向がみられる．これは，尺度に用いる単語が少なくても交差法では検索性能が高い，すなわち交差法では検索に有効な単語を優先的に選択できていることを示している．3 つのコレクションのうち，MED については，2 つの方法の性能差は小さいものの，CRAN と CISI ではより大きな性能差があった．また，この検索性能の差の傾向はフィードバックする文書数によらなかった．

4.2 偏差法の評価

偏差法では，式 (6) の $norm(i)$ が大きな N_t 個の単語の α_i を 1 とすることで，検索順位の尺度 $NPR'(n)$ を決定する．単語の選択順位を与える $norm(i)$ は，各単語の単語出現確率 $\theta_{ri}^{(-n)}$ を平滑化パラメータ $\xi_i = \theta_{Gi}$ の下でフィードバックされた適合文書と問い合わせの単語出現頻度から推定することで求めた．選択する単語の数 N_t は，$norm(i) > 2$ となる単語数 N_2 と，$norm(i) > 0$ となる単語数 N_0 を基準として，

$$N_t = (1-\zeta)N_2 + \zeta N_0 \quad (8)$$

から決定した．N_2 と N_0 には $N_2 > N_0$ の関係があり，ζ が大きいほど尺度に用いる単語数は多くなる．実験では，交差法と同様に ζ を-0.25 から 1.5 の範囲で 0.25 刻みで変えて平均適合率を求めた．

表 4 は，偏差法と確率比による選択法で同数の単語を選択したときの検索性能の差を示したものである．表より，ζ が小さいとき偏差法により高い検索性能があり，ζ が大きくなるに従ってその差が小さくなる傾向がみられる．この傾向は，交差法を用いた場合と同様であった．

5 考察

交差法と偏差法の 2 つの単語選択法を用いて，適合フィードバックによる文書検索を行った結果，確率比の大きさから単純に単語を選択する場合と比べて，選択する単語数が少ない場合でも高い性能があることが示された．これは，フィードバックされた適合文書から検索に有効な単語を抽出する能力が，交差法と偏差法で高いことを示している．これらの方法では，ユーザの問い合わせがもつ多義性と同義性の問題を減らすのに効果的な単語が選択されていることが示唆された．

交差法，偏差法と確率比による選択法との検索性能の差は，選択する単語数が増えるに従って小さくなった．選択する単語の候補は有限であるため，それぞれの方法で選択した場合に同じ単語が含まれる比率は多くの単語を選択するほど高くなると考えられる．それゆえ，選択する単語数が多い場合では，それぞれの方法で検索に用いる単語が似通ってしまい，検索性能の差が小さくなったと考えられる．

表 3, 4 と表 2 との比較から明らかなように，検索順位の尺度に用いる単語を選択する場合は，すべての単語を用いる場合より全体的に検索性能の向上がみられた．単語を選択する場合では，フィードバックされた適合文書に含まれない単語は検索順位の尺度に用いられない．例えば，偏差法では適合文書に含まれない単語では，$ratio\_{}_n(i)$ の平均値が適合文書に含まれる単語よりも小さく，標準偏差は 0 となる．このため，これらの単語の $norm(i)$ は適合文書に含まれるどの単語よりも小さくなる．よって，検索順位の尺度から除外される．一方，すべての単語を用いる場合では，適合文書に含まれない多くの単語が検索順位の尺度に用いられる．これらの単語は文書の検索順位を下げるのに寄与する．しかし，フィードバックされた適合文書に含まれていない単語すべてがユーザの問い合わせと関連性が低いとは限らない。すべての単語を用いる場合で検索性能が低いのは，フィードバックされた適合文書に含まれない単語を一律に不適合とみなす尺度に原因があると考えられる．

6 おわりに

確率モデルを用いた文書検索において，フィードバック情報から検索対象文書の順位決定に用いる単語の選択法として，leave-one-out に基づく交差法と

表 3: 平均適合率での交差法と確率比による方法との比較: **太字**は 10%以上適合率が高いことを示す.

	$N_{feedback}$	方法	γ = -0.25	0	0.25	0.5	0.75	1	1.25	1.5
MED	10	交差法	**0.53154**	0.54075	0.54050	0.57317	0.57131	0.56425	0.56412	0.56412
		確率比	0.47534	0.53163	0.54425	0.56213	0.56603	0.56677	0.57121	0.56767
	20	交差法	0.47905	0.49908	0.50850	0.49566	0.50907	0.54092	0.53020	0.53020
		確率比	0.47321	0.50936	0.50472	0.52082	0.52337	0.52684	0.52346	0.51191
	30	交差法	**0.48882**	0.49900	0.51150	0.51750	0.51878	0.52955	0.53516	0.53524
		確率比	0.43181	0.45401	0.52137	0.50523	0.52212	0.53808	0.53586	0.53035
CRAN	10	交差法	**0.29186**	**0.34350**	0.34927	0.35558	0.35808	0.35562	0.35158	0.35158
		確率比	0.21481	0.29295	0.31988	0.33010	0.35300	0.35625	0.35369	0.34847
	20	交差法	**0.28573**	**0.31891**	**0.33293**	0.33654	0.34113	0.33949	0.33430	0.33438
		確率比	0.21523	0.26900	0.30060	0.31646	0.33803	0.33879	0.33673	0.32979
	30	交差法	**0.27885**	**0.31049**	**0.32708**	**0.32337**	0.31667	0.32082	0.31464	0.31456
		確率比	0.18890	0.24497	0.28652	0.29015	0.31035	0.31370	0.31408	0.31145
CISI	10	交差法	**0.22875**	**0.24630**	**0.24615**	**0.23954**	0.24097	0.25350	0.25073	0.25073
		確率比	0.11865	0.18460	0.20224	0.21550	0.23077	0.24646	0.23512	0.23478
	20	交差法	**0.21488**	**0.21885**	**0.20808**	0.20600	0.20947	0.21269	0.21328	0.21328
		確率比	0.11391	0.14811	0.17584	0.19349	0.20424	0.21757	0.21459	0.21098
	30	交差法	**0.20766**	**0.21980**	**0.21055**	0.20473	0.20975	0.20920	0.20948	0.20948
		確率比	0.10455	0.14752	0.17754	0.19838	0.20970	0.21304	0.20994	0.20586

表 4: 平均適合率での偏差法と確率比による方法との比較: **太字**は 10%以上適合率が高いことを示す.

	$N_{feedback}$	方法	ζ = -0.25	0	0.25	0.5	0.75	1	1.25	1.5
MED	10	偏差法	**0.34923**	**0.54353**	0.55432	0.56898	0.56351	0.56643	0.56452	0.55779
		確率比	0.14793	0.45049	0.52087	0.53515	0.55129	0.55642	0.56516	0.56712
	20	偏差法	**0.40975**	**0.50575**	0.50956	0.51942	0.51956	0.52889	0.51966	0.50525
		確率比	0.25608	0.44866	0.48322	0.50922	0.52002	0.52339	0.53072	0.52018
	30	偏差法	**0.41339**	**0.49007**	**0.51153**	0.51600	0.51177	0.53609	0.53782	0.50860
		確率比	0.29658	0.43433	0.45374	0.49230	0.49818	0.51377	0.51623	0.50822
CRAN	10	偏差法	**0.25121**	**0.29839**	**0.31341**	**0.33409**	**0.33875**	0.34079	0.34728	0.35042
		確率比	0.14881	0.23589	0.27758	0.28719	0.30216	0.31618	0.32931	0.33183
	20	偏差法	**0.22719**	**0.28900**	**0.30457**	**0.32295**	0.32891	0.32116	0.32626	0.33078
		確率比	0.16419	0.24390	0.27050	0.28500	0.30420	0.30697	0.31483	0.30970
	30	偏差法	**0.20701**	**0.28452**	**0.29518**	**0.31561**	**0.31840**	0.30107	0.30391	0.31423
		確率比	0.14673	0.20959	0.23989	0.26231	0.26762	0.27680	0.29475	0.29386
CISI	10	偏差法	**0.14945**	**0.24893**	**0.25103**	**0.25069**	**0.23966**	0.24063	0.23640	0.22716
		確率比	0.11209	0.15875	0.17534	0.20781	0.21664	0.22690	0.22977	0.22745
	20	偏差法	**0.13719**	**0.21926**	**0.21530**	**0.21042**	0.21141	0.20971	0.20885	0.20109
		確率比	0.08911	0.13495	0.15348	0.18293	0.20031	0.20670	0.20632	0.21034
	30	偏差法	**0.13714**	**0.21292**	**0.21515**	**0.21219**	0.20754	0.20968	0.20360	0.19270
		確率比	0.09173	0.12700	0.15314	0.18017	0.20079	0.20563	0.20576	0.20257

偏差法の 2 つの方法を考案し，その有効性を検証した．評価実験の結果，2 つの方法では，検索性能を高める効果の高い単語を優先的に選択できることが示された．単語の選択数や平滑化パラメータの最適化が今後の課題である．

参考文献

[1] Harman, D.: Relevance feedback revisitied, *Proc. of 15th Ann Int'l SIGIR '92*, 1–10 (1992).

[2] Lavrenko, V. and Croft, W.B.: Relevance Models in informaiton retrieval, in *Language Modeling for information retrieval* (Kluwer academic publishers, Netherlands, 2003), pp. 11–56.

[3] Ng, K: A maximum likelifood ratio information retrieval model, *Proc. of the Eighth Text Retrieval Conference (TREC-8)*, pp. 483-492 (1999).

[4] Robertson, S.E. and Sparck Jones K.: Relevance weighting of search terms, *Journal of the Ameciran Society for Information Science*, **27**, 129–146 (1976).

[5] Robertson, S.E.: The probability ranking principle in IR, *Journal of Doumentation*, **33**, 294–304 (1977).

[6] Rocchio, J.J.: Relevance feedback in information retrieval, in *The SMART Retrieval System* (Prentice-Hall, Inc., N.J., 1971), pp. 313–323.

[7] Salton, G., Wang, A. and Yang, C.S.: A vector space model for information retrieval, *Journal of the American Society for Information Science*, **18**, 613–620 (1975).

[8] Salton, G. and McGill, M.J.: *Introduction to Modarn Information Retrieval* (NcGraw-Hill, New York, 1983).

[9] Zhai, C. and Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval, *Tenth International ACM Conference on Information and Knowledge Management (CIKM'01)* , (2001).

[10] 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).