確率モデルに基づく日本語情報フィルタリングにおけるフィードバックによる検索条件展開および検索精度評価

酒井哲也[†] Gareth J.F. Jones[††]
梶浦正浩[†] 住田一男[†]

本論文では，確率モデルに基づく日本語情報フィルタリングにおけるレレバンスフィードバックおよびローカルフィードバックによる検索条件展開実験について報告する．検索条件展開は，正解文書の本文あるいは見出しから抽出した語を，新たな検索語として初期検索条件に付加することにより実現する．レレバンスフィードバックの実験では，BMIR-J1および独自のテストコレクションを学習用・評価用に用い，平均適合率が18%まで向上することを示す．ローカルフィードバックの実験では，テストコレクションBMIR-J1およびBMIR-J2を用い，平均適合率が5%程度向上することを示す．さらに，テストコレクションで定義されている検索要求グループ別の評価により，レレバンスフィードバックとローカルフィードバックの結果の比較を行う．

Query Expansion through Feedback in Japanese Information Filtering Based on the Probabilistic Model

Tetsuya Sakai,[†] Gareth J.F. Jones,[††] Masahiro Kajiura[†] and Kazuo Sumita[†]

This paper reports experiments in query expansion through *relevance feedback* and *local feedback* for online Japanese news filtering with the probabilistic NEAT system. The expansion terms are extracted from either the full texts or the headings of the relevant documents. Using the standard BMIR-J1 test collection and a separate collection, relevance feedback is shown to produce an improvement of up to 18% in average performance. Local feedback is shown to give a smaller, but still significant, improvement for the BMIR-J1 and BMIR-J2 collections. In addition, using the query complexity groups defined for the collections, we compare our relevance and local feedback results.

1. はじめに

情報洪水の時代といわれる近年，膨大な量の情報の中から必要な情報を見つけ出す情報検索や情報フィルタリングの技術が注目を集めている．とくにテキスト情報は，WWWページ・電子メイル・ネットニュースなどに見られるように，マルチメディア情報の中で依然として主要な役割を果たしている．

情報検索の分野では過去数十年間にわたって，欧米の文書を検索対象とした多くの成果が得られてきた[1]．ところが最近になって，形態素解析やN-gramなどの前処理を施すことにより，これらの研究成果の多くが日本語や中国語などの単語間の区切りが明確でない言語に対しても適用可能であることが明らかになってきた[2]．

我々は文献3)において，見出しや段落といった文書構造情報および検索語の文書内頻度ベクトルを利用した情報フィルタリングシステムNEATのもとで，自然言語文から検索条件を自動生成する実験について報告した．そして，日本語情報検索システム評価用のテストコレクション[4]を用い，情報検索の分野で標準的に用いられている再現率・適合率をもとにした評価基準のもとでNEATが高い検索精度を実現することを示した．NEATは1996年以来，ユーザが興味を持ちそうな新聞記事を毎朝電子メイルやWWWで配信する情報フィルタリングサービスにおいて実運用されている．我々はその後，英文検索において有効性が示されている確率検索モデル（probabilistic retrieval model）[5]に基づく検索を行えるようにNEATの機能を拡張した[2]．

本論文では確率モデルに基づくNEATを用い，文

† 株式会社東芝研究開発センター
Toshiba R&D Center
†† Department of Computer Science, University of Exeter

献3)で扱ったような単純な初期検索条件をスタート地点とし，これに対してレレバンスフィードバック（relevance feedback，以下RF）およびローカルフィードバック（local feedback，以下LF）と呼ばれる技術を適用することにより検索条件を展開し，検索精度向上を図る．ここで，RFは，検索結果の各文書が正解であるか否かをユーザに評価させ，この評価情報を用いて初期検索条件を修正する技術である[1)☆1]．一方，LFは，ユーザの評価情報が得られない場合に，初期検索結果の上位の文書を正解であると仮定することによりRFと同様な修正を行うものである[6)]．

本論文におけるRFの実験では，日本語情報検索システム評価用テストコレクションBMIR-J1[4)☆2]およびこれとは記事の発行年が異なる独自のテストコレクションTCIR-N1[3),7)☆3]を学習用および評価用に用いることにより，均質な単一のテストコレクションを用いた従来の評価に比べてより情報フィルタリングの実際のタスクに即した厳しい条件での評価を行う．また，文書全体からのみではなく，文書の見出しのみから検索条件展開のための新たな検索語を抽出することも試みる．一方，LFの実験ではBMIR-J1およびBMIR-J2[8)☆4]により評価を行う．また，両フィードバックの実験において，テストコレクションで定義されている検索要求グループ[4)]別の分析も行う．

2章で本研究に関連する従来研究について，3章で情報フィルタリングシステムNEATについて，4章で本論文で用いたテストコレクションについて，5章でRFの実験について，6章でLFの実験について述べ，7章で結果および考察を，8章でまとめを述べる．

2. 従来研究

2.1 日本語検索におけるレレバンスフィードバック

英語を対象にしたRFについてこれまで多くの研究がなされてきたのに比べ，日本語検索におけるRFの歴史は浅い[2),7)]．文献9)では，コーパスに基づく新たな日本語のインデキシング手法が提案されており，初期検索における検索精度がテストコレクションBMIR-J1を用いて評価されているが，これまでにRFの実験についての報告はされていない．文献10)では，日本語のWWWページを検索対象としたRFの標準的手法を拡張した式が提案されているが，ここではテストコレクションを用いた客観的評価が行われていないため，その有効性が明らかになっていない．文献11)では，彼らの提案する多段情報フィルタリングにおいてRFの標準的手法の適用が検討されているが，ここではテストコレクションBMIR-J2の中の10件の検索要求を用いた評価しか行われておらず，統計的に信頼できるレベルの評価結果が得られているとはいい難い．

2.2 レレバンスフィードバックの評価方法

RFの評価を行う際，初期検索条件とフィードバック後の検索条件の検索精度を同一の検索対象を用いて単純に比較してしまうと正当な評価が行えないことが知られている．これは，RFにおいてはユーザからどの文書が正解であるかという情報が与えられるので，正解文書を検索結果の上位に持ってくるように検索条件を修正すれば初期検索結果よりも検索精度の高い検索結果を得ることは容易だからである．このため，RFの効果を正当に評価するための手法がいくつか提案されている[12)]．Rank freezing法およびresidual collection法は，単一のテストコレクションを検索対象として用いるが，フィードバック後の検索結果中に初期検索においてすでに検索された文書が含まれる場合にこれらを除外して評価しようという考え方に基づく手法である．一方，test and control法は，テストコレクションを二分割し，片方を学習用に，もう片方を評価用に用いるものである．ここで注意すべきは，これらのいずれの手法においても，フィードバックに利用する文書とフィードバック後の検索精度評価に用いる文書とが同一のテストコレクションに属するという意味で均質である点である．

固定的な情報を扱う従来の情報検索とは異なり，情報フィルタリングは日々発生する新しい情報の流れを扱うため，情報フィルタリングにおけるRFを考える場合には上記のような均質性の仮定は現実に即さない．ここでのタスクは，過去に到着した文書を利用してフィードバックを行い，新たに到着する文書に対する検索精度を向上させることである．文書内に出現する語彙は時間とともに移り変わると考えられるため，このタスクは均質で固定的な検索対象に対するRFよりも一般には難しいと考えられる．そこで我々は5章

☆1 一般には検索条件展開，すなわち検索条件に対する新たな検索語の付加や不要な検索語の削除のみでなく，既存の検索語の重みの調整（term reweighting）[1)]にも用いられるが，本論文では前者の，とくに新たな検索語の付加のみを扱う．

☆2 株式会社日本経済新聞社の協力によって，社団法人情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが，1993年9月1日から12月31日の日本経済新聞記事を基に構築した情報検索評価用データベース（テスト版）を利用．

☆3 日本経済新聞CD-ROM1995年版を利用．

☆4 社団法人情報処理学会・データベースシステム研究会が，新情報処理開発機構との共同作業により，毎日新聞CD-ROM'94データ版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BMIR-J2を利用．

において，検索要求を共有し，かつ検索対象の文書集合の発行年が異なる2つのテストコレクションを用いたRFの評価実験を行う．さらに，正解文書の本文全体を利用したRFに加え，見出し情報のみを利用したRFについても実験を行う．

2.3 ローカルフィードバック

RFはユーザが正確かつ矛盾のない十分な量の評価を初期検索結果に対して行ってくれることを前提としている．このような状況が期待できない場合は，ユーザの評価情報を使わずにLFを行うことにより検索精度が向上できる可能性がある．LFでは，正解であると仮定された上位の文書の中に実際には正解でない文書が混入している場合があり，これはノイズを含んだ学習に相当するため，一般にRFほどの効果が望めない．一方，初期検索結果の精度がある程度良い場合にはRFに近い効果が期待できる．実際，文献6)などではLFの有効性が示されている．ただし，日本語を対象としたLFを扱った研究はこれまでに報告されていない．

なお，RFとは異なり，LFはユーザの評価情報を利用しないため，評価用データと別の学習用データを必要とせず，前述のrank freezingやresidual collectionなどの検索精度評価手法を用いる必要もない．すなわち，フィードバック後の検索結果を初期検索結果と見なして評価することができる．我々は6章において，NEATがLFを施す前の段階においても文献9)などの従来研究に勝る検索精度を実現していることを示す．

3. 情報フィルタリングシステムNEAT

3.1 基本システム

NEATの基本システムは，文書構造などを利用した多様な条件に基づき個々の文書のスコアを算出し，文書のランキングを行う[3]．たとえば，検索語が**text**条件のもとに記述されると文書の本文全体がその検索語のマッチングの対象となり，**head**条件のもとに記述されると文書の見出しのみが対象となる．ここで，検索語のマッチングには，文字列マッチングと形態素マッチングがある☆．一般にNEATの検索条件は複数の条件から構成され，文書dのスコア$SCORE(d)$は各条件により算出されたスコアから以下のように合成される．

$$SCORE(d) = \frac{\sum_i w_c(i)score_c(i,d) + \sum_i w_m(i)score_m(i,d)}{\sum_i |w_c(i)| + \sum_i |w_m(i)|} \quad (1)$$

ここで，$score_c(i,d)$，$score_m(i,d)$は検索条件中に記述された第i番目の条件の文字列マッチングおよび形態素マッチングに基づくスコアであり，$w_c(i)$，$w_m(i)$は，これらの$SCORE(d)$に対する寄与度を決定する条件重みである．本論文では$w_c(i) = w_m(i) = w(i)$とする．

3.2 確率モデルに基づくシステム

現バージョンのNEATは，**text**および**head**条件のみについて，基本システムのスコア算出のかわりに確率検索モデルに基づくスコア算出を行うことができる．これまでに，少なくとも**text**条件のみを用いた場合には，基本システムよりも確率モデルに基づくシステムのほうが高い検索精度が達成できることが分かっている．そこで本論文では，確率モデルに基づくNEATにおける**text**および**head**条件のみを扱い，このもとで検索条件展開の実験を行う．

確率モデルに基づくNEATは，ある条件のもとに記述された検索語tおよび文書dに対する検索語重み$tw(t,d)$を以下のように算出する．

$$tw(t,d) = \frac{\log(|C|/df(t)) * tf(t,d) * (K+1)}{K * \left((1-b) + \frac{b * L(d) * |C|}{\sum_{d \in C} L(d)}\right) + tf(t,d)} \quad (2)$$

ここで，Cは検索対象となる文書集合，

$L(d)$ = 文書d，あるいはその見出しの文字数，

$tf(t,d)$ = 文書d中，あるいはその見出し中における検索語tの出現頻度，

$df(t)$ = 文書集合Cの中で検索語tを含む文書あるいは見出しの数，

$K = tf(t,d)$の影響を調整するために経験的に決定される定数$(0 \leq K)$，

$b = L(d)$の影響を調整するために経験的に決定される定数$(0 \leq b \leq 1)$．

確率モデルに基づくNEATでは，検索条件中の第i番目の条件により算出されるスコアは以下のようになる．

☆ 文字列マッチングだけでは，たとえば「プリン」という検索語に対して「プリンター」という語を含む文書がマッチしてしまう．これは文書の形態素解析結果に対して形態素マッチングを行うことにより回避できる．一方，形態素解析の精度は完全ではないので，形態素マッチングだけでは「スプリンター」という検索語にマッチする文書を見落とす可能性がある．文書中の「スプリンター」が「ス/プリンター」のように誤解析されているかもしれないからである．このように，両マッチングは相補的である．

表1 テストコレクション
Table 1 Test collections.

	TCIR-N1	BMIR-J1	BMIR-J2
文書の種類	日本経済新聞	日本経済新聞	毎日新聞
発行年	1995	1993	1994
文書数	5,048	600	5,080
検索要求数	56	60	50
正解数/要求	15.1	10.1	33.6

$$score(i,d) = \sum_{t \in T(i)} tw(t,d) \quad (3)$$

ここで，$T(i)$ は第 i 番目の条件のもとに記述された検索語の集合である．

基本システムと同様に，式 (3) は文字列マッチングと形態素マッチングの両方の場合について算出され，最終的に式 (1) により文書のスコアが合成される．

4. テストコレクション

表1 に本論文で用いたテストコレクションの概要を示す．我々は，BMIR-J1 および J2 の全検索要求を用いて LF の実験を行った．これについては 6 章で説明する．TCIR-N1 は我々が独自に作成したテストコレクションであり，BMIR-J1 と 56 個の検索要求を共有している[3),7)]．我々は，これら 56 件を用いて，両テストコレクションの文書集合の片方を学習用に，もう一方を評価用にした RF の実験を行った．これについては 5 章で説明する．表 1 から分かるように BMIR-J1 および TCIR-N1 はともに日本経済新聞の記事であるが，前者は 1993 年の記事，後者は 1995 年の記事であるため，使われている語彙にいくらかの違いがあると考えられる．このように学習用・評価用データ間に時間的ギャップを与えている点で，我々の RF の評価方法は，従来の均質なテストコレクションを用いた評価に比べより情報フィルタリングの現実に即していると考えられる．

本論文では，テストコレクションを用いた検索精度の評価尺度として情報検索の分野で標準的とされている再現率・適合率曲線および 11 点平均適合率を主に用いる[1)]．また，7.3 節では，document cut-off value (dcv)[12)]に基づく検索結果の上位 15 件に着目した平均適合率[3)]による評価も補助的に行う．

5. レレバンスフィードバックの実験

5.1 初期検索条件の生成

まず，以下の手順で BMIR-J1 と TCIR-N1 に共通な 56 の検索要求に対応する初期検索条件を作成した．

(1) 検索要求の名詞句を形態素解析し，名詞・数字・未知語を抽出する．

(2) 抽出した語を **text** 条件のもとに羅列する．

(3) 各テストコレクションに対し式 (2) の定数 K および b を 11 点平均適合率の観点から最適化する．

(4) 検索条件に **head** 条件を追加し，条件重みを $w = 0.2, \ldots, 1.0$ と変化させる．羅列する語は **text** 条件のものと同じとする．

BMIR-J1 では $K = 1.0$, $b = 0.2$, TCIR-N1 では $K = 0.5$, $b = 0.6$ となった．BMIR-J1 では **head** 条件を $w = 0.2$ で付加することにより効果が見られたので，これを初期検索条件として **INITIAL** と名付けた．一方，TCIR-N1 では **head** 条件の効果が見られなかったため，**text** 条件のみからなる検索条件を初期検索条件として **INITIAL** と名付けた．

5.2 検索条件展開

RF による検索条件展開の実験では，まず TCIR-N1 を学習用に，BMIR-J1 を評価用に用い，次に BMIR-J1 を学習用に，TCIR-N1 を評価用に用いた．現存の日本語テストコレクションは比較的規模が小さく，統計的に有意な実験結果を示すことが難しいが，両方向の実験結果における一貫性を示すことにより結果の信頼性を高めることができると考えられる．

以下の手順で検索条件展開を行った．

(1) 学習用コレクションの正解文書の本文あるいは見出しから，$m = 5, 10, \ldots, 50$ 個の語を，後述する検索語選出基準に基づいて選出する．

(2) **text** あるいは **head** 条件のもとに上記展開語を羅列したものを検索条件 **INITIAL** に付加する．条件重みは $w = 0.2, \ldots, 1.0$ と変化させ，評価用コレクションに対して 11 点平均適合率の観点から最適化する．

(3) さらに，**text** と **head** 条件を併用した検索条件展開を試みる．

検索語選出基準としては，$rtf * idf$, $rntf * idf$, $rdf * rw$ の 3 種類を考慮した．これらは以下の統計量から算出される．

$$rtf(t) = \sum_{d \in R} tf(t,d) \quad (4)$$

$$rntf(t) = \sum_{d \subset R} tf(t,d)/L(d) \quad (5)$$

$$idf(t) = \log(|C|/df(t)) \quad (6)$$

$$rw(t) = \log \frac{\frac{rdf(t)+0.5}{|R|-rdf(t)+0.5}}{\frac{df(t)-rdf(t)+0.5}{|C|-df(t)-|R|+rdf(t)+0.5}} \quad (7)$$

ここで，R は正解文書集合，rdf はその検索語を含む正解文書の数，rtf は正解文書中の検索語頻度の和，

text :1, メーカー, 菓子;
text :0.2, 製菓, 菓, コメ, 食品, もち, 手当て, 不作, 越後, もち米, せんべい, くず米, 各社, 冷夏, 銘柄, 必要量, 中国産, 食糧, みそ, あられ, 社長, 原料, 不足, 大口, 米ドル, 佐藤, 割り当て, どれ, 直接, 加工, 製造;

図1 展開されたプロファイルの例
Fig. 1 An expanded profile.

$rntf$ は各正解文書中の検索語頻度を文書の文字数で正規化した同様の値である．一方，idf は inverse document frequency と呼ばれ，特定の文書に偏在する検索語を重視するために一般的に用いられる[1]．また，rw は Robertson/Sparck Jones relevance weight と呼ばれ，確率検索モデルにおいて導出される統計量である[5]．ここで，正解文書に関する情報が得られない場合，すなわち $|R| = rdf = 0$ である場合，式 (6) および式 (7) より，rw は idf と似た挙動を示すことが分かる．

我々は，文献 13) における NEAT の基本システムを用いた RF の実験で，rtf, $rntf$, idf を個別に検索語選択基準として用いた場合よりも $rtf * idf$ あるいは $rntf * idf$ のほうが有効であることを示した．さらに，文献 2) における確率モデルに基づく NEAT を用いた RF の実験で，rdf, rtf, rw を個別に用いた場合よりも $rdf * rw$ のほうが有効であることを示した．よって，本論文で扱う $rtf * idf$, $rntf * idf$, $rdf * rw$ は，我々の経験上最も有効な検索語選出基準である．

図1 に，「菓子メーカー」という検索要求に対して RF により展開された検索条件の実例を示す．この例では，検索条件展開により $m = 30$ 個の展開語を含む第 2 の **text** 条件が $w = 0.2$ で付加されている．

以上の方法で展開された検索条件のうち，評価用テストコレクションに対する 11 点平均適合率の観点から最適であった検索条件を **R-BEST** と名付けた．

6. ローカルフィードバックの実験

6.1 初期検索条件の生成

BMIR-J1 および J2 の初期検索条件をそれぞれ 5 章と同様な方法で生成し，**INITIAL** と名付けた．この結果，BMIR-J1 については 5 章と同様に $K = 1.0$, $b = 0.2$ となった．ここで注意すべきは，5 章では 56 件の検索要求のみが扱われたのに対し，本章では BMIR-J1 の全検索要求 60 件を扱っている点である．BMIR-J2 についても標準セット[8]全 50 件を扱っており，本章での検索精度は他システムと直接比較可能である．実際，BMIR-J1 の初期検索では **text** 条件を用いただけで 11 点平均適合率 0.525 が得られたが，これはすでに文献 9) の最適値 0.513 を上回っている．

なお，BMIR-J2 については $K = 0.5$, $b = 0.4$ となり，また **head** 条件の効果は見られなかったため **text** 条件のみからなる検索条件を **INITIAL** と名付けた．

6.2 検索条件展開

LF による検索条件展開は，BMIR-J1 および J2 に対して個別に，以下の手順で行った．

(1) 検索条件 **INITIAL** により初期検索を行う．
(2) 初期検索結果の上位 $n = 5, 10, \ldots, 20$ 文書を擬似正解文書とする．
(3) 5 章と同様に検索条件展開を行う．

以上の方法で展開された検索条件のうち，テストコレクションに対する 11 点平均適合率の観点から最適であった検索条件を **L-BEST** と名付けた．

7. 結果および考察

7.1 レレバンスフィードバック

表 2 および **表 3** に，それぞれ BMIR-J1 および TCIR-N1 を評価用に用いた場合の RF の主な結果を示す．各種検索条件は 11 点平均適合率の降順に並べてあり，「選出基準」と「m」の欄が空白になっている行はフィードバック前の初期検索条件を表している．たとえば，表 2 からは，条件重み $w = 1.0$ の **text** 条件と条件重み $w = 0.2$ の **head** 条件からなる初期検索条件が **INITIAL** と名付けられ，これに対して $rntf * idf$ により選出された 20 個の展開語を条件重み $w = 0.2$ の **text** 条件のもとに並べて付加したものが **R-BEST** と名付けられていることが読み取れる．「向上率」の欄は，**INITIAL** を基準として 11 点平均適合率が何%向上したかを示している．また，**図 2** は TCIR-N1 における **INITIAL** と **R-BEST** の再現率・適合率曲線である．以下，両表に沿って結果をまとめる．

- BMIR-J1 と TCIR-N1 の記事の発行年が 2 年違うにもかかわらず，**R-BEST** の **INITIAL** に対する向上率は大きく，BMIR-J1 に対して 11%，TCIR-N1 に対して 18%であった．TCIR-N1 を評価用に用いた場合の向上率のほうが大きいのは，TCIR-N1 では **INITIAL** の検索精度が BMIR-J1 のそれに比べて低かったためであると考えられる☆．

☆ 検索精度はテストコレクションの規模・均質さ・正解の含有率などに左右される．また，BMIR-J1 の正解文書は全数調査により決定されているのに比べ，TCIR-N1 の正解文書は網羅的でないため，検索精度の評価値は実際よりも低い値になると考えられる．これは BMIR-J2 についても同様である．

表2 RF の結果 (BMIR-J1: $K = 1.0, b = 0.2$)
Table 2 RF results (BMIR-J1: $K = 1.0, b = 0.2$).

条件 (w)	選出基準	m	11pt	向上率
INITIAL	$rntf * idf$	20	**0.570**	**11%**
+**text**(0.2)				
(**R-BEST**)				
INITIAL	$rdf * rw$	20	0.569	11%
+**text**(0.2)				
INITIAL	$rdf * rw$	45	0.537	5%
+**head**(0.2)				
INITIAL	$rntf * idf$	10	0.532	4%
+**head**(0.2)				
text(1.0)	–		**0.513**	**0%**
+**head**(0.2)				
(**INITIAL**)				
text(1.0)	–		0.512	–

表3 RF の結果 (TCIR-N1: $K = 0, 5, b = 0.6$)
Table 3 RF results (TCIR-N1: $K = 0, 5, b = 0.6$).

条件 (w)	選出基準	m	11pt	向上率
INITIAL	$rdf * rw$	30	**0.507**	**18%**
+**text**(0.2)				
(**R-BEST**)				
INITIAL	$rdf * rw$	15	0.462	7%
+**head**(0.2)				
INITIAL	$rntf * idf$	30	0.451	5%
+**head**(0.2)				
INITIAL	$rntf * idf$	5	0.449	4%
+**text**(0.4)				
text(1.0)	–		**0.431**	**0%**
(**INITIAL**)				
text(1.0)	–		0.410	–
+**head**(0.2)	–			

• 検索語選出基準としては $rdf * rw$ が最も安定して高精度であった☆. また, $rntf * idf$ と $rtf * idf$ を比較すると前者のほうが安定しており, 文書長による正規化がある程度有効であることが分かった. なお, 表中では $rtf * idf$ の結果[7]は省略している.

• 検索条件展開により付加した条件の重み w の最適値は, 検索語選出基準と語数 m のほぼすべての組合せについて 0.2 であった. 図 3 (上) は, 表 3 の **R-BEST** ($rdf * rw, m = 30, w = 0.2$) および ($rntf * idf, m = 5, w = 0.4$) に対応する検索条件の m を固定し, w を変動させた様子を表している. グラフがほぼ単調減少であることから, w の真の最適値は一般に 0.0 と 0.2 の間に位置すると考えられる☆☆. これを直感的に解釈すると, 展開語の重要度は初期検索語の重要度の 5 分の 1 以下であるということになる.

☆ 表2 では $rntf * idf$ が **R-BEST** に採用されているが, $rdf * rw$ との間に有意差はない.

☆☆ $w = 0.0$ は **INITIAL** と等価である.

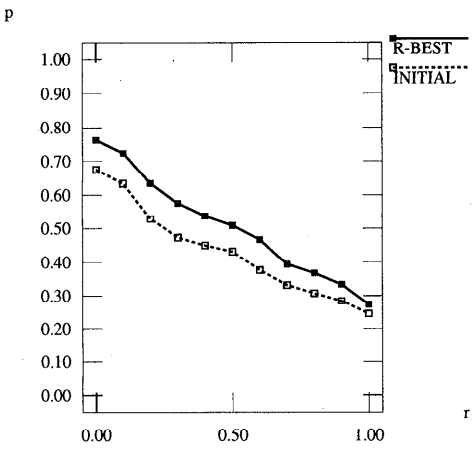


図2 RF (TCIR-N1) の再現率・適合率曲線
Fig. 2 R-P curve for RF (TCIR-N1).

• 展開語数 m の最適値は 20〜30 程度であった. 図 3 (下) は, 表 3 の **R-BEST** ($rdf * rw, m = 30, w = 0.2$) および ($rntf * idf, m = 5, w = 0.4$) に対応する検索条件の w を固定し, m を変動させた様子を表している. これより, 展開語数を無闇に増やしてもメリットがないことが推測される.

• 検索条件展開に用いる条件としては, **head** 条件は **text** 条件ほど有効ではなかった. これは, **head** 条件は見出しという制限された文書領域の情報を用いるものであることから, 予測されたことである.

• 今回の実験では, 検索条件展開における **text** と **head** 条件の併用の効果は見られなかった. 併用した場合の結果[7]は **R-BEST** に及ばなかったので表中では省略している.

text と **head** 条件の併用の効果が見られなかったのは, 本文全体から **text** 条件用の検索語を選出する際と見出しから **head** 条件用の検索語を選出する際に同じ検索語選出基準を用いたためであると考えられる. この方法では, **head** 条件用に選出された検索語はすでに **text** 条件用にも選出されているため新たな効果が得られにくい. たとえば, **text** 条件用には見出し以外の文書領域から検索語を選出するといった方法をとれば, 併用による精度向上が期待できる. なお, この結果から, 6 章の LF の実験では検索条件展開に **text** 条件のみを用いた.

時間的ギャップがある 2 つのテストコレクションにより RF を評価する場合, 検索要求がどれほど時代に依存している (chronologically sensitive である) かにより, その効果は大きく左右されると考えられる. たとえば,「菓子メーカー」という BMIR-J1 の検索要

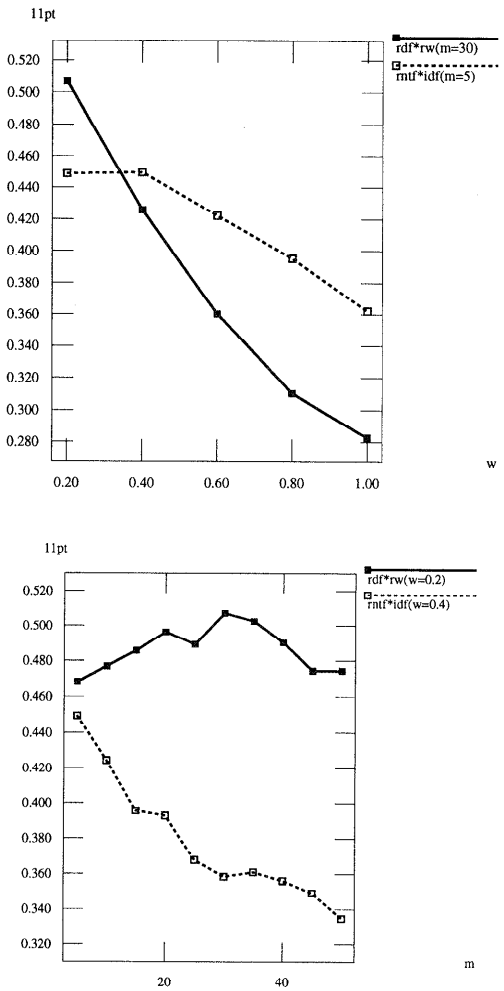


図3 RF（TCIR-N1）における w/m と 11pt の関係
Fig. 3 Relationship between w/m and 11pt in RF (TCIR-N1).

求に適合する文書に含まれる語彙は，時間がたってもそれほど変化するとは思われない．一方，「経営多角化の事例」のような検索要求の場合は，時代とともに様々な事例が発生し，関連する語彙は移り変わっていくと考えられる．しかし，今回の実験では，時代依存性と向上率の間には明らかな相関は見られなかった．実際，TCIR-N1 に対する評価において，検索要求 56 件のうち **INITIAL** にくらべ **R-BEST** の検索精度が向上したものは 35 件，低下したものは 16 件，変化がなかったものは 5 件であったが，大きな向上が見られたものの中にも，「管理部門の統廃合と営業部門の強化を行う会社」や「経営陣刷新」のように時代依存性が大きそうな検索要求があった．逆に，低下が見られたものの中には，「任天堂」や「農薬」など，比較的

表4 LF の結果（BMIR-J1: $K = 1.0, b = 0.2$）
Table 4 LF results (BMIR-J1: $K = 1.0, b = 0.2$).

条件 (w)	選出基準	n	m	11pt	向上率
INITIAL	$rdf*$	10	15	**0.553**	**5%**
+text(0.2)	rw				
(L-BEST)					
INITIAL	$rntf*$	5	15	0.530	0%
+text(0.2)	idf				
text(1.0)	–			**0.529**	**0%**
+head(0.2)					
(INITIAL)					
text(1.0)	–			0.525	–

表5 LF の結果（BMIR-J2: $K = 0.5, b = 0.4$）
Table 5 LF results (BMIR-J2: $K = 0.5, b = 0.4$).

条件 (w)	選出基準	n	m	11pt	向上率
INITIAL	$rdf*$	5	10	**0.482**	**5%**
+text(0.2)	rw				
(L-BEST)					
INITIAL	$rntf*$	5	5	0.477	4%
+text(0.2)	idf				
text(1.0)	–			**0.459**	**0%**
(INITIAL)					
text(1.0)	–			0.444	–
+head(0.2)					

語彙が安定していそうな検索要求があった．以上からはっきりとした結論を出すことは難しいが，我々は，たとえ検索要求の時代依存性が高い場合でも，正解文書から時代依存性の低い検索語をうまく抽出することにより，時間を隔てた RF もある程度機能すると考えている．このためには，いくつかの時代からサンプリングした文書を時系列に並べ，この文書集合に対して idf に似た検索語の散らばり具合を表す尺度を定義し，統計的に時代依存性を判定するというアプローチが考えられる．

7.2 ローカルフィードバック

表4 および **表5** に，それぞれ BMIR-J1 および J2 に対する LF の主な結果を示す．表の見方は擬似正解文書数 n のパラメータが加わったことを除けば表 2 および表 3 と同様である．また，**図4** は BMIR-J2 における **INITIAL** と **L-BEST** の再現率・適合率曲線である．以下，両表に沿って結果をまとめる．

- BMIR-J1, J2 の場合ともに **L-BEST** の **INITIAL** に対する向上率は 5%であり，やはり **R-BEST** の場合と比較すると効果は小さい．
- RF の場合と同様，検索語選出基準を有効な順に並べると $rdf*rw, rntf*idf, rtf*idf$ となる．表では $rtf*idf$ の結果[7]を省略しているが，かえって **INITIAL** よりも精度が低下するケースもあった．

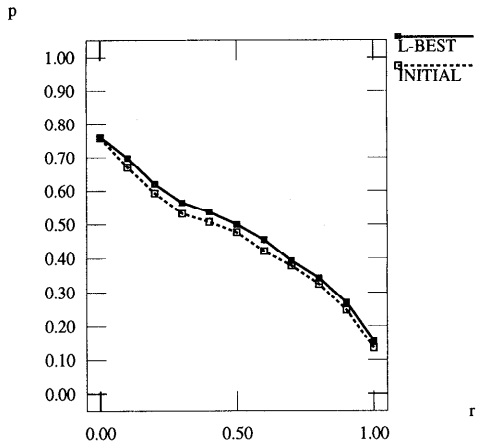
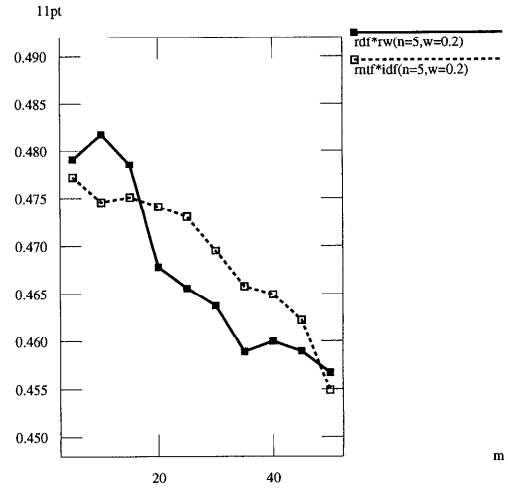


図4 LF（BMIR-J2）の再現率・適合率曲線
Fig. 4 R-P curve for LF (BMIR-J2).



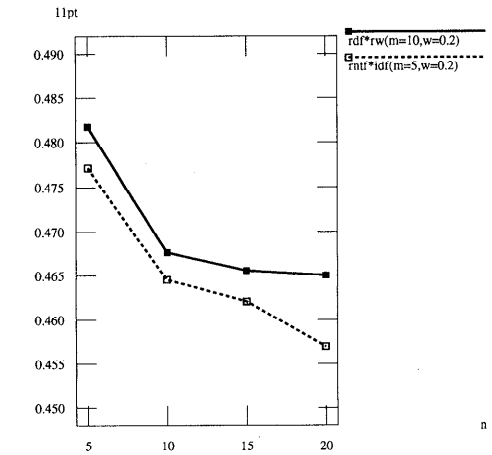


図5 LF（BMIR-J2）における m/n と 11pt の関係
Fig. 5 Relationship between m/n and 11pt in LF (BMIR-J2).

• RFの場合と同様，条件重み w の最適値は0.2であった．また，表5の **L-BEST** $(rdf*rw, n=5, m=10, w=0.2)$ および $(rntf*idf, n=5, m=5, w=0.2)$ に対応する検索条件の n および m を固定し，w を変動させると図3（上）同様の単調減少曲線が得られた．よって，やはり w の真の最適値は一般に0.2以下であると推測される．

• 展開語数 m の最適値は10〜15程度であり，予想されたとおりRFの場合よりも少ない．すなわち，LFでは正解文書にノイズが含まれるので，語数を増やすと不適切な語が選出されてしまう可能性が大きくなってしまうものと考えられる．**図5**（上）は表5の **L-BEST** $(rdf*rw, n=5, m=10, w=0.2)$ および $(rntf*idf, n=5, m=5, w=0.2)$ に対応する検索条件の n および w を固定し，m を変動させたものだが，RFの場合とは若干異なり，展開語数が増えるにつれ検索精度がほぼ単調に下がっていく様子が分かる．

• 擬似正解文書の件数 n の最適値は5程度であり，テストコレクションの平均正解文書数（表1参照）よりも若干小さな値になっている．図5（下）は，表5の **L-BEST** $(rdf*rw, n=5, m=10, w=0.2)$ および $(rntf*idf, n=5, m=5, w=0.2)$ に対応する検索条件の m および w を固定し，n を変動させた様子を表している．語数 m の場合と同様，n の増大はノイズの増大につながることが分かり，また n の真の最適値は一般に5以下であることが推測される．

なお，我々はBMIR-J1を用いて，LFを発展させた手法であるlocal context analysis[6]の考え方に基づく補助的な実験も行った．LFが上位文書の全文を検索語抽出対象とするのに対し，local context analysisは上位文書中で初期検索語がマッチした部分のテキスト（passage）のみを検索語抽出対象とする．この手法は，多様な内容を述べた長い文書のある特定の部分に初期検索語がマッチした際，その部分とは関係のない部分からも検索語が抽出されてしまうことを回避する効果があると考えられる．今回は，初期検索語がマッチした文あるいは段落をpassageと見なして実験を行ったが，LFの場合（すなわち全文を用いた場合）と，段落を用いた場合，文を用いた場合の間には有意な差が見られなかった．この主な原因は，もともとテストコレクションの新聞記事，少なくとも上位に来る記事には極端に長いものは少なく，単一の話題について述べているものが多いためであると考えられる．

表6 $dcv = 15$ における平均適合率および符号検定結果
Table 6 Paired sign tests for precision at $dcv = 15$.

	INITIAL	R-BEST	有意差
BMIR-J1	0.310	0.356	あり ($\alpha = 0.01$)
TCIR-N1	0.250	0.296	あり ($\alpha = 0.01$)
	INITIAL	**L-BEST**	有意差
BMIR-J1	0.301	0.319	あり ($\alpha = 0.01$)
BMIR-J2	0.461	0.496	あり ($\alpha = 0.05$)

7.3 統計的検定

RF および LF の効果の統計的有意性に関して考察を行う．情報検索の分野で検索精度の差の検定方法の1つとして推奨されている符号検定[12]を用いたところ，RF については，BMIR-J1，TCIR-N1 ともに **INITIAL** と **R-BEST** の 11 点平均適合率の間に有意差があった ($\alpha = 0.01$)．一方，LF については，BMIR-J1 については **INITIAL** と **L-BEST** の 11 点平均適合率の間に有意差があったが ($\alpha = 0.01$)，BMIR-J2 については有意差がなかった．

しかし，11 点平均適合率ではなく，**表 6** のように $dcv = 15$ における平均適合率を算出し，これに対して符号検定を行った結果，BMIR-J2 の **INITIAL** と **L-BEST** の間にも有意差があった ($\alpha = 0.05$)．$dcv = 15$ における平均検索精度は 11 点平均適合率よりもユーザの体感に近い数値であると考えられるので[3]☆，ユーザの観点から，本実験での LF は効果があると考えることができる．

7.4 検索要求グループ別の比較

表 7 および **表 8** に，BMIR-J1 に対するそれぞれ RF および LF の主な結果を検索要求グループ別に集計したものを示す☆☆．両表における各行は **INITIAL** の 11 点平均適合率の降順に示されており，また **all** の行が表 2 および表 4 の太字で示した 11 点平均適合率に対応する．これらから，グループ A～C よりもグループ D～F の検索精度が低くなっていることが分かる．このことをより明確にするために，グループ A～C およびグループ D～F に関する 11 点平均適合率の平均値をそれぞれ **ABC**，**DEF** の行に示している．両表における **ABC** および **DEF** の **INITIAL** に対する向上率を見てみると，**INITIAL** の検索精度が低かった **DEF** のほうが向上率は大きいことが分かる．さらに，**ABC** と **DEF** の向上率の差は，RF よりも LF の場合のほうが若干小さいように思われる．これは，LF においては，難しい検索要求であるグループ D～F の初期検索結果の上位には多くの不正解文書が含まれてしまい，これらを正解文書と見なして検索条件展開を行っても効果が小さいためであると考えられる．

☆ 11 点平均適合率は文書の件数でなく再現率を基準にしているが，上位記事のみを受信するユーザには，再現率すなわち検索洩れの度合は不明である．

☆☆ TCIR-N1 に対する RF および BMIR-J2 に対する LF のグループ別集計結果[7]はここでは省略する．

表7 検索要求グループ別の RF の精度（BMIR-J1）
Table 7 RF performance over different complexity groups (BMIR-J1).

グループ	検索要求数	INITIAL	R-BEST	向上率
B	5	0.760	0.769	1%
ABC	**20**	**0.704**	**0.752**	**7%**
A	10	0.691	0.778	13%
C	5	0.673	0.684	2%
all	**56**	**0.513**	**0.570**	**11%**
E	9	0.511	0.574	12%
D	10	0.430	0.473	10%
DEF	**36**	**0.407**	**0.469**	**15%**
F	17	0.337	0.412	22%

表8 検索要求グループ別の LF の精度（BMIR-J1）
Table 8 LF performance over different query groups (BMIR-J1).

グループ	検索要求数	INITIAL	L-BEST	向上率
B	5	0.760	0.749	−1%
ABC	**21**	**0.718**	**0.743**	**3%**
C	6	0.727	0.753	4%
A	10	0.691	0.734	6%
E	10	0.545	0.545	0%
all	**60**	**0.529**	**0.553**	**5%**
D	12	0.455	0.499	9%
DEF	**39**	**0.427**	**0.450**	**5%**
F	17	0.337	0.360	7%

8. ま と め

本論文では，日本語情報フィルタリングのための RF および LF による検索条件展開の実験について報告した．RF の実験では，発行年の異なる 2 つのテストコレクションを利用した実験によりその有効性を示した．実際の情報フィルタリングでは，学習用・評価用データ間の時間的ギャップはより小さいと考えられるので，ユーザから十分な正解文書情報が与えられさえすればより大きな検索精度向上が期待できる．ただし今後は，バッチ的にではなく徐々に検索条件の修正を行う手法である incremental relevance feedback[14]などを前提に，検索語の時代依存性の判定方法を検討する必要がある．また，LF の実験では，ユーザにより正解文書情報が与えられない場合でも検索精度をある程度向上できることを確認した．

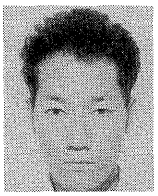
今後は，NEAT の基本システムがサポートしている **text** および **head** 条件以外についても確率モデルに対応できるようにし，とくに，文書構造を利用した

検索と確率モデルとの整合性を探っていく．また，現在NEATは日本語と英語に対する検索をサポートしているが，cross-language検索・多言語検索の方向にも研究を進めていく予定である[15]．

参考文献

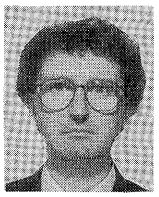
1) Salton, G., et al.: *Introduction to Modern Information Retrieval*, Computer Science Series, McGraw-Hill Book Company (1983).
2) Jones, G.J.F., et al.: Experiments in Japanese Text Retrieval and Routing using the NEAT System, *Proc. ACM SIGIR '98*, pp.197–205 (1998).
3) 酒井ほか：情報フィルタリングのためのブール式と文書構造を利用した検索条件生成と検索精度評価，情報処理学会論文誌，Vol.39, No.11, pp.3076–3083 (1998).
4) 木本ほか：日本語情報検索システム評価用テストコレクションの構築，情報学シンポジウム'98 (1998).
5) Robertson, S.E., et al.: *Simple, Proven Approaches to Text Retrieval*, Computer Laboratory, University of Cambridge (1994).
6) Ballesteros, L., et al.: Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval, *Proc. ACM SIGIR '97*, pp.84–91 (1997).
7) Sakai, T., et al.: Application of Query Expansion Techniques in Probabilistic Japanese News Filtering, *Proc. IRAL '98*, pp.46–55 (1998).
8) 酒井ほか：情報検索システム評価のためのテストコレクション，*Computer Today*, No.87, pp.31–35, サイエンス社 (1998).
9) Ogawa, Y., et al.: Overlapping Statistical Word Indexing: A New Indexing Method for Japanese Text, *Proc. ACM SIGIR '97*, pp.226–234 (1997).
10) Eguchi, K., et al.: Information Retrieval Considering Adaptation to User's Behaviors on the WWW, *Proc. IRAL '97*, pp.108–113 (1997).
11) 菅井ほか：多段情報フィルタリング方式とその評価，人工知能学会第12回全国大会論文集，pp.390–393 (1998).
12) Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments, *Proc. ACM SIGIR '93*, pp.329–338 (1993).
13) 酒井ほか：ユーザの要求に応じた情報フィルタリングシステムNEATのプロファイル生成，インタラクション'98論文集，pp.149–152 (1998).
14) Allan, J.: Incremental Relevance Feedback for Information Filtering, *Proc. ACM SIGIR '96*, pp.270–278 (1996).
15) 酒井ほか：Cross-language 情報検索のためのBMIR-J2を用いた一考察，情処学会自然言語処理研究会，NL-129-7, pp.41–48 (1999).

(平成10年9月16日受付)
(平成11年2月 8日採録)



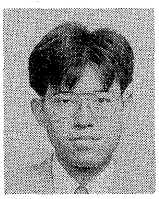
酒井 哲也（正会員）

昭和43年生．平成5年早稲田大学大学院理工学研究科工業経営学専門分野修士課程修了．同年（株）東芝入社．情報検索・情報フィルタリングの研究開発に従事．



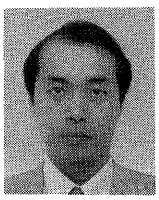
Gareth J.F. Jones

Gareth J.F. Jones received a B.Eng in 1989 and a PhD in 1994 from the University of Bristol, UK. From 1993 to 1996 he was a Research Associate at the University of Cambridge, UK. In 1996 he was appointed as a Lecturer at the University of Exeter, UK. From 1997 to 1998 he was a Toshiba Fellow at the Toshiba R&D Center. His research interests include: information retrieval, speech recognition, natural language engineering, and virtual technology. He is a member of the UK IEE and British Computer Society IRSG.



梶浦 正浩（正会員）

昭和41年生．平成3年慶應義塾大学大学院理工学研究科計算機科学専攻修士課程修了．平成6年同専攻後期博士課程単位取得退学．同年（株）東芝入社．情報検索・情報フィルタリング，ECの研究開発に従事．



住田 一男（正会員）

昭和32年生．昭和57年東京工業大学大学院総合理工学研究科物理情報工学専攻修士課程修了．同年（株）東芝入社．自然言語処理，情報検索・フィルタリングの研究開発に従事．