文書検索におけるランキング検索技術

Ranking Retrieval in Document Retrieval

真野　博子[*]　　伊東　秀夫[*]　　小川　泰嗣[*]
Hiroko MANO　　Hideo ITOH　　Yasushi OGAWA

要　　旨

文書検索では，必要な文書を誰でも簡単に探せる技術が求められている．ランキング検索技術は，自然言語で表現された検索要求に対し，文書データベース中からその検索要求に適合する文書群を適合度の順にランキングして返す技術である．確率モデルに基づく文書のランク付け，自然言語解析系を用いた検索語抽出，および関連語による検索条件の拡張などの技術から成る高性能なランキング検索技術を開発した．本技術は，性能評価のため参加した国際的な検索システム評価会 NTCIR において高い検索有効性を示した．これら技術を組込んだ全文検索サーバ FTS を使用し公開特許公報ランキング検索サービスを開発した．

ABSTRACT

One of the goals of research in document retrieval is ease of use for end-users to find documents of their interest. Ranking retrieval is retrieval of documents from a document database in which retrieval requests are given in natural language and documents returned in response are ranked in terms of the degree of relevance to the request. A set of ranking retrieval technologies was developed using document ranking based on the probabilistic model, automatic extraction of query terms with a natural language analyzer and expansion of queries with additional related terms. The strengths of the technologies were shown in the international workshop NTCIR, where the system achieved high retrieval effectiveness. The technologies have been built in the full-text search server FTS and applied to the patent retrieval service.

[*] ソフトウェア研究開発本部　ユビキタスソリューション研究所
Ubiquitous Solution Lab, Software R&D Group

1. 背景と目的

文書電子化の進展に伴い，大量の文書群から必要な文書を素早く探し出す文書検索技術の重要性が高まっている．リコーのドキュメントハイウェイにおいても文書検索は基本的な機能の一つと位置付けられている．

文書検索のうち，文書の全文を検索対象とするものは全文検索とよばれる．従来，全文検索の代表的な方法として用いられてきたのはブーリアン検索とよばれる方法である．ブーリアン検索とは，検索したい文書の条件を，キーワードとAND，ORなどのブーリアン演算子を組み合わせたブーリアン演算式で表現するもので，例えば「リコーの環境保護への取り組み」についての文書を探したいとすれば，一例として以下のようなブーリアン演算式で表現する．

「リコー AND 環境保護 AND 取り組み」

ブーリアン検索では，検索したい文書を適切に表現する演算式を思いつけば，納得のいく検索結果を得られることも多い．だが，ブーリアン演算子に不慣れなユーザにとっては，どういった演算子を使用すればよいかわからないこと，また，ブーリアン演算では，ひとつひとつの文書について指定された演算式に当てはまるか否かのどちらかの判断しかされず，検索条件に適合する度合いを扱えないことなど，使いやすさなどの面で問題もあった．

ブーリアン検索のこういった問題点に対処するため情報検索の分野で研究されてきた技術に以下のものがある．

(1) 文書を検索条件への適合度によりランク付け（ランキング）する
(2) ブーリアン演算子を用いず自然言語で表現した文や文章をそのまま検索条件とする
(3) 検索条件にない関連語を加えて検索条件を拡張する

これらの技術（本稿では総称しランキング検索*とよぶ）は，必要な文書を誰でも簡単に探せるようにするための重要な技術であり，われわれもこれらを実現するため新たな技法を提案してきた．本稿では，ランキング検索技術について，要素技術ごとにわれわれの開発した技法を解説し，その検索性能，応用について説明する．

* ランキング検索は，関連文書検索，概念検索，類似検索などとよばれることもある．

2. 技術

2-1 文書のランク付け

前述のように，ランキング検索の第一の特徴は，検索結果において文書を検索条件に適合する度合いの順にランク付けすることである．文書のランク付け方法の代表的なものにはベクトル空間モデルと確率モデルがある[1]が，われわれはモデルの厳密性や検索有効性の高さから確率モデルを採用している．

2-1-1 確率モデルの基本

確率モデルの基本となる考え方は，検索条件中の検索語それぞれについて，その語が，検索条件に適合する文書に出現する確率と適合しない文書に出現する確率を基に，その語が適合文書を選り分けるのに貢献する度合いを示す「重み」を割り当てることである．つまり，適合文書に現れやすく適合しない文書に現れにくい検索語は，適合文書に特徴的な語とみなし重みを高くするのである．

式（1）は，確率モデルにおける語の重みづけの基本となるものである[2]．ここで，w_t は語 t の重み，p_t は，適合文書に語 t が現れる確率，q_t は，適合しない文書に語 t が現れる確率である．これにより適合文書に現れる確率が高く適合しない文書に現れる確率の低い語に高い重みがつくようになる．

$$w_t = \log\frac{p_t}{1-p_t} - \log\frac{q_t}{1-q_t} = \log\frac{p_t(1-q_t)}{q_t(1-p_t)} \qquad (1)$$

適合文書に語 t が現れる確率 p_t および適合しない文書に語 t が現れる確率 q_t は，どれが適合文書か判明していれば，式（2）により正確に求められる[2]．ここで，N は全文書数，n_t は語 t をふくむ文書数，R は適合文書数，r_t はそのうちの語 t をふくむ文書数である．

$$p_t = \frac{r_t}{R}$$
$$q_t = \frac{n_t - r_t}{N - R} \qquad (2)$$

2-1-2　語の重み付け

このように，確率モデルでは，どれが検索条件に適合する文書であるかが判明しているという前提で，その情報を基に語に重みを付与する，という考え方だが，実際には適合文書は検索を実行した後に初めて得られるもので検索実行時にはわからない．そこで，適合文書が不明の場合には，全文書数や t をふくむ文書数から p_t と q_t を推定する．代表的な方法として，以下のものがある[3]．ここで p_0 は p_t の取りうる推定最小値である．

$$p_t = \frac{p_0}{p_0 + (1-p_0)\dfrac{N-n_t}{N}}$$

$$q_t = \frac{n_t}{N}$$

これを式（1）にあてはめると，以下のようになる．（定数となる部分は k_4 とする．）

$$w_t = \log\frac{p_0}{1-p_0} + \log\frac{N}{n_t} = k_4 + \log\frac{N}{n_t}$$

この値は，基本的には，語の重要度の尺度として広く知られている Inverse Document Frequency[4] と同じものになる．ただ，推定値 p_0 の値の選択を誤ると，Fig.1左のように p_t より q_t が大きくなる可能性があるという問題点も指摘されている．p_t より q_t が大きくなれば，重みが負の値になるため検索の有効性が大きく低下する．

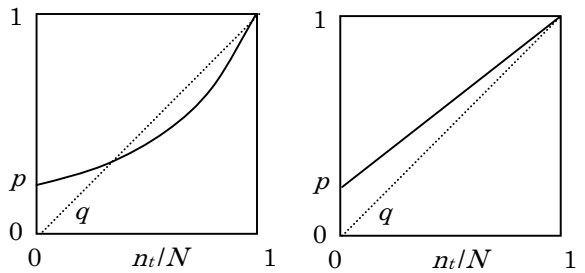

Fig.1　p and q assumptions

そこで，われわれは，p_t の求め方を以下のように改良した[5]．

$$p_t = p_0 + (1-p_0)\frac{n_t}{N}$$

この改良により，Fig.1右のように，p_t は必ず q_t より大きくなることが保証でき，安定した検索結果が得られるようになる．改良方式によれば，重みは，

$$\begin{aligned} w_t &= \log\left(\frac{p_0}{1-p_0}\cdot\frac{N}{N-n_t} + \frac{n_t}{N-n_t}\right) - \log\frac{n_t}{N-n_t} \\ &= \log\left(k'_4\cdot\frac{N}{n_t} + 1\right) \end{aligned} \qquad (3)$$

となる．（定数となる部分は k'_4 とする．）

このように，語の重みは，その語をふくむ文書の数が少ないほど値は大きくなるようになっている．つまり，より少ない文書にしか現れない語ほど検索語としての価値が高いとみなされることになる．

2-1-3　文書スコアリング

文書を検索条件への適合の度合いでランク付けするには，検索対象の各文書について検索条件への適合度を求める必要がある．重みは，検索条件中の検索語について，その語が適合文書を識別するのにどの程度有用かを示すものだが，文書の適合度を判断するには，こうした識別性の観点に加えて代表性の観点，すなわち，その語が，文書をどの程度代表しているかも反映されるようにする．

具体的には，検索語 t をふくむ検索条件 q への文書 d の適合度 $s_{d,q}$ は，以下の式[5]により算出する．

$$s_{d,q} = \sum_{t\in q}\frac{f_{t,d}}{K + f_{t,d}}\cdot\frac{w_t}{k'_4\cdot N + 1}$$

$$K = k_1\left((1-b) + b\frac{l_d}{l_{ave}}\right)$$

ここで，$f_{t,d}$ は文書内頻度，すなわち一文書内での出現回数であり，その文書を代表する語であれば高くなる．ただし文書内頻度は文書全体の長さとの関連で見る必要があるため，文書長 l_d や平均文書長 l_{ave} も考慮し，文書内頻度が同じなら，より短い文書の場合に適合度が高くなるようになっている．b，k_1 によってこの度合いを調整する．

2-2　検索語抽出

ランキング検索の第二の特徴は，検索条件にブーリアン演算子を使用しない，つまり，検索したい文書の内容を自然言語で表現した文や文章（自然文）をそのまま検索条件とすることができることである．そのためには，入力された自然文から検索語を取り出す自然言語解析技術が必要となる．こ

れには，大きく二つの段階があり，形態素解析により入力された自然文を形態素（意味を持つ最小単位）に分割する処理と，得られた全ての形態素から検索に有用な語のみを検索語として選出する処理から成る．

2-2-1　形態素解析

文字列を形態素に分解し品詞を付与する技術が形態素解析技術[6]である．以下，簡単にソフトウェア研究開発本部で開発した技法[7][8]の特徴について触れておく．

形態素解析の基本は，与えられた文字列の可能な形態素分解結果の中から品詞の組み合わせ規則と照らし合わせてどの規則がもっとも当てはまるかを判断することである．品詞の組み合わせ規則を定めるには，適切な品詞体系の設定が重要となる．正確な解析のためには，品詞体系は細かい方が望ましいが，単純に品詞を細かくすると，その分，規則数が増えてしまう．そこで，階層的品詞体系と，品詞よりも微小な文法属性を記述する素性を用い，多様なレベルでの文法記述を可能にすることにより，規則数を抑えつつ解析精度を高めている．

また，検索のための解析では，表記の揺れへの対応も不可欠である．たとえば，検索条件で「インターネット」と指定した場合，文書中にあるのが「インタネット」であると検索条件に合致しない．これを防ぐには，一般的には語の正規化を行い，内部的に一定の表記に統一する．ただ，正規化処理だけでは，過剰統一や過小統一に陥りやすいため，われわれの方法では，検索時に逆方向の処理である展開処理も加えることで，正規化の漏れに起因する検索漏れを回避しつつ，過剰な統一による検索ノイズの発生を抑えている．

実装面では，トライ構造を用いた辞書引き，動的計画法を用いたパス探索などの採用により，高速な形態素解析を実現している．

2-2-2　検索語の選出

形態素解析により検索文字列から形態素が抽出されるが，抽出した形態素がすべて検索に有用なわけではない．たとえば，「リコーの環境保護への取り組み」の「の」「への」は，検索のためには，あまり利用価値がない．よって，検索に有用な語のみを検索に使用するためには，検索語を選出するための処理が必要である．

(1)　品詞パターン辞書による絞り込み

検索語として使用すべき品詞のパターンを定めた辞書を用いて，検索語とする形態素のみを抽出する．これにより，たとえば，助詞の「の」は検索に用いないよう設定できる．パターン辞書の記述には正規表現を用いることが可能で複雑なパターンでも容易に記述できるようになっている．

(2)　品詞パターン辞書による隣接語の抽出

入力文から抜き出した形態素を，それぞれ独立の検索語としてのみ使用すると，たとえば，「環境保護」は「環境」と「保護」に分けられてしまい，「個人情報保護のための環境整備」といった内容でも合致してしまう可能性がある．これを防ぐために，隣接する形態素を連結したものも，品詞パターン辞書で有用性を判断した上で検索語として追加できるようにしている．これにより従来技術に比べきめ細かい指定ができ検索の有効性向上につながる．

(3)　禁止辞書によるふるい分け

どの文書中にも出現する頻出語は，たとえ品詞の上では有用であっても検索語には適さない．たとえば，特許公報のデータベースにおける「請求項」などである．このように，特定の語が多くの文書に共通に現れることがあらかじめわかっている場合は，そういった語を列挙した禁止辞書を用意しておき，そこにある語は検索に用いないようにする．

(4)　出現回数による優先付け

入力が長い文章であるような場合，ここまでの処理を実行しても，なお，多くの検索語が残り，これらをそのまま検索に利用すると検索時間が長くなる．そこで，残った検索語に優先度をつけ優先度の高いものから適当な上限数になるまで検索語とする．優先度は，通常は，入力文の中での出現回数の多いものが高くなるよう設定する．

(5)　蒸留処理による優先付け

前項のような入力文中の使用回数のみによる優先度では，入力文と検索対象の性質が異なる場合に適切に評価できないという問題がある．たとえば，手元にある新聞記事を元に，そこで紹介されている技術についての特許を特許公報データベースから検索するとする．このように検索条件として入力する文章（新聞記事）が検索対象としている文書（特許公報）と異なった種類の文書群に属する異種データ横断検索では，入力されるデータと検索されるデータの間で使われている語彙などが異なるために，前述のように語の検索対象デー

タベースでの出現状況だけを根拠に重み付けする通常の方法では，不適切な検索語が重要視されてしまう可能性がある．たとえば，新聞記事に多く出現する「社長」という語は，新聞記事の中では特に重要な語ではないにもかかわらず，特許公報では出現する文書数が小さいため重要とみなされ非常に大きな重みがついてしまう．

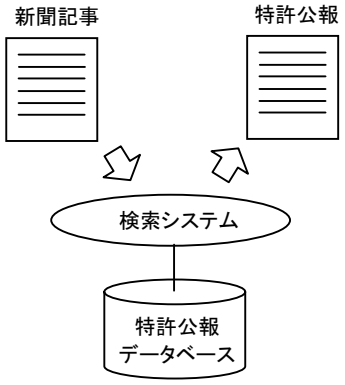

Fig.2　Cross-DB retrieval.

この問題に対処するためにわれわれが開発したのが検索語蒸留という，検索語群を「蒸留」して有用なものだけを取り出す処理方法である[9]．検索語蒸留では，入力する文と同種（検索対象とは異種）の文書群を収めたデータベースを参照し検索語の優先付けをおこない優先度の高い検索語のみを使用することで，先に述べたような検索対象データベースのみに基づいてランキングすることによる弊害を防ぐ．

具体的には，検索語 t の優先度 TDV_t を求めるには，検索対象データベースの全文書数 N，検索対象データベースで t の出現する文書数 n_t，異種データベースの全文書数 M，異種データベースで t の出現する文書数 m_t を用い，その語が，検索対象データベースに出現する確率 a_t および異種データベースに出現する確率 b_t を以下のように求め，

$$a_t = \frac{n_t}{N}$$

$$b_t = \frac{m_t}{M}$$

これら a_t，b_tを用いて，以下のように算出する．これは，実験で最も効果のあった，前述の式（1）を応用した計算式である．

$$TDV_t = f_t \cdot \log \frac{a_t(1-b_t)}{b_t(1-a_t)}$$

ここで f_t は，検索条件内での t の出現回数である．

2-3　関連語による検索条件の拡張

文書を探すのに，思いついた検索条件で検索してみて望ましい検索結果が得られない場合，考えられる原因のひとつは，検索条件中に使用した語が，探そうとしている文書で用いられている語と必ずしも一致しないことである．たとえば，環境保護についての文書を探そうとする場合，まず「環境」や「保護」といった検索条件で検索するのが普通だが，実際の文書では，環境保護について「環境」や「保護」という言葉でなく「ゴミゼロ運動」や「リサイクル推進」といった言葉で説明されていることもある．このような語の不一致は検索漏れにつながり，特定の内容に関する文書を網羅的に収集したい場合などには特に大きな問題となる．

こういった検索漏れを防ぐための手法のひとつが検索条件の拡張である．つまり「環境」や「保護」といった語が検索条件として入力された時，自動的にそれらに関連する「ゴミゼロ」や「リサイクル」などの語を補って検索できるようにする技術である．

具体的には，まずユーザの入力した検索条件中の語を検索語として利用して検索し，その結果得られた適合文書から関連語を抽出しそれを検索語に追加する．ここでの課題は，検索結果から適合文書を見つけること，および，適合文書中の語から関連語を選出することである．

2-3-1　適合文書判定

(1)　適合性フィードバック

適合文書を知るもっとも確実な方法は，ユーザに適合文書がどれかをシステムに指示してもらう方法である．この情報は適合性フィードバックとよばれ，確率モデルの考え方の基礎となっている[2]．

(2)　擬似適合性フィードバック

ユーザに適合文書であるかどうかの判断をしてもらうのは確実ではあるが，ユーザにとっては余計な作業が要求されることになるため，必ず適合文書が得られるとは限らない．このように，ユーザから充分な適合性フィードバックが得られない，あるいは，ユーザとのインタラクションなしに関連語を抽出したい場合には，かわりに擬似適合性フィードバックを利用する．これは，ユーザから与えられた検索条件でとりあえず検索した結果から上位一定数を適合文書であるとみ

なして処理するのである．

検索条件を拡張するには，一般には，適合性フィードバックもしくは擬似適合性フィードバックのどちらか一方を利用する．われわれは，適合性フィードバックと擬似適合性フィードバックを併用する方法も提案し，適合性フィードバックが充分でないときそれを補完することで検索性能を向上させている．

2-3-2　関連語の選出

検索条件への適合文書が判明したとして，そこから関連語を選出するには，まず，前述の検索語選出に述べた方法で，適合文書中のすべての語から関連語として利用できそうなものだけを抽出する．次に，抽出した関連語候補のそれぞれについて関連語としての有用度を計算し，有用度の高い順に上位一定数を関連語とする．

(1)　有用度算出

有用度には，Robertson's Selection Value（RSV）を採用している[10]．RSVでは，適合文書に与えられる適合度の期待値と適合しない文書に与えられる適合度の期待値を比較しその差が大きくなる語ほど関連語として適しているとみなす．

すなわち語 t の有用度 RSV_t は

$$RSV_t = w2_t \cdot \left(\frac{r_t}{R} - \frac{n_t}{N} \right)$$

で求められる．ここで $w2_t$ は，以下で説明する適合性あるいは擬似適合性フィードバックを反映した重みである．こうして選択した関連語は元々の検索語に追加され新しい検索語となる．

(2)　再重み付け

関連語を追加した後の再検索では，単に検索語が追加されるだけでなく，元々の検索語についても，適合性・擬似適合性フィードバックを反映した重みに付け直される．これが，上記の $w2_t$ である．

前述のように，確率モデルでは，適合文書に語 t が現れる確率 p_t および適合しない文書に語 t が現れる確率 q_t に基づいて語に重みを付与する．ここで p_t ，q_t の値を得るのに，検索前の段階では適合文書がわからないため，適合文書数を用いない方法で推定した．しかし，一旦検索し適合性・擬似適合性フィードバックを得た後は，ある程度，適合文書が判明するため，より正確に p_t ，q_t の値を推定できる．

ここで仮に，適合性フィードバックにより，すべての適合文書が判明したと仮定する．この場合，p_t ，q_t は，式（2）で求められることは前述のとおりである．これを式（1）に当てはめると，重み w'_t は以下のようになる[2]．（0.5は，r_t が0であった場合でも対数が取れるようにするための調整値．）

$$w'_t = \log \frac{r_t}{R - r_t} - \log \frac{n_t - r_t}{N - R - (n_t - r_t)}$$

$$\approx \log \frac{\dfrac{r_t + 0.5}{R - r_t + 0.5}}{\dfrac{n_t - r_t + 0.5}{N - n_t - R + r_t + 0.5}}$$

実際には，適合性・擬似適合性フィードバックによって充分な数の適合文書が得られるとはかぎらず，その場合，少量の適合文書からp_t ，q_t の値を推測することになる．この，いわば証拠不足を補うため，フィードバック後の再重み付けでは，以下のように検索前の重みと上記の重みを線形結合により組み合わせて利用する[3]．ここで，α は組み合わせの比率を指定する定数である．

$$w2_t = \alpha \cdot w_t + (1 - \alpha) \cdot w'_t$$

このように拡張された検索条件により再検索を実行すれば，ユーザの求める文書が上位に出やすくなる．

3. 評価

検索システムの性能には，効率性（結果を出す速さ）と有効性（出た結果の妥当性）の二つの側面がある[11]．ランキング検索においては，効率性だけでなく，検索された各文書が検索条件に適合しているかという有効性が非常に重要である．以下，検索有効性の評価指標，評価のための評価会，およびそこでのわれわれの検索システムの評価結果を紹介する．

3-1　検索有効性評価指標

検索の有効性を測る指標として，もっとも基本的なものは再現率（recall）と精度（precision）である．再現率は，適合文書を漏れなく検索できる能力を示し，精度は，適合文書のみを検索できる能力を示す．具体的には，Fig.3のように，適合文書，検索された文書，検索された適合文書があるとき，

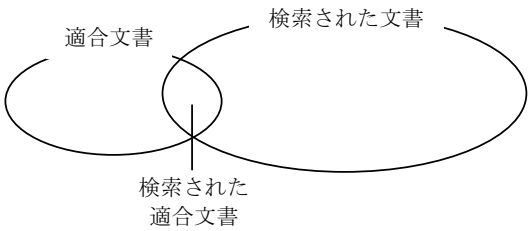

Fig.3　Retrieved and relevant documents.

以下のように計算する．

$$再現率 = \frac{検索された適合文書数}{適合文書数}$$

$$精度 = \frac{検索された適合文書数}{検索された文書数}$$

これらの指標は，検索された文書全体を評価するものであるので，個々の文書についてそれが検索された順位と適合性との関係については考慮されない．これを考慮した尺度のひとつが平均精度（average precision）である．平均精度は，検索された文書について，上位に順位付けられたものから順に適合性を判断していき，適合文書が検索されるごとに，それまでに検索された文書数および適合文書数から精度を求め，それらを平均したものである．そのため，より多くの適合文書がより高い順位で検索されるほど高い値になる．平均精度は評価会で検索システムの性能を比較するのに，もっともよく用いられる評価尺度のひとつである．

3-2　評価会NTCIR

NTCIR（NII/NACSIS Test Collection for Information Retrieval）は，国立情報学研究所の主催する，日本語で書かれた文書を主に対象とする評価会形式の国際的な情報検索ワークショップである[12]．1998年以降これまで3回開催されている．毎回のワークショップには，検索対象や検索方法などの異なる複数のタスクが設定されており，第3回では，特許公報を検索対象とした特許タスク，Web上の文書を検索対象としたWebタスクなど5タスクが設定された．NTCIRに参加する研究機関・企業は，共通の検索対象，検索課題を用い指定された検索実施要件に従って検索を実行し検索結果（複数可）を提出する．主催者側では提出された検索文書について適合判定し正解文書集合を定めた上で，それを基に各検索システムの性能を評価する．

われわれは，第2回から参加しており，昨年，実施された第3回（NTCIR-3）では，特許タスクとWebタスクに参加した．以下，それぞれのタスクについてわれわれが実施した検索実施要件と参加結果について述べる．

3-3　NTCIR-3

3-3-1　特許タスク

(1)　タスク概要

特許タスク[13]は，特許公報を検索対象とするものでNTCIR-3で初めて設定された．このタスクでは，新聞記事を元にその記事で取り上げられた技術に関する特許公報を探すという想定で検索課題が用意された．一例（抜粋）を挙げる．検索課題は全部で31課題あった．検索対象として，1998年から1999年にかけて公開された特許公報2年分の全文などが提供された．

```
<TITLE>
バーコードなどの符号を比較し優劣を判定する装置
</TITLE>
<ARTICLE>
<A-DOC>
<A-HEADLINE>エポック社の特許侵害訴訟、バンダイが敗訴－－東京地裁
</A-HEADLINE>
<A-TEXT>
カードゲームの特許を侵害されたとして、玩具（がんぐ）製造会社のエポック社がバンダイに2億6400万円の損害賠償を求めた訴訟で、東京地裁は30日、約1億1400万円の支払いを命じた。森義之裁判長は、バンダイが1992年7月～93年3月に製造・販売した小型ゲーム機「スーパーバーコードウォーズ」のキー操作などの機能について「エポック社が持つ特許の技術的範囲に属する」と指摘した。
</A-TEXT>
</A-DOC>
</ARTICLE>
<SUPPLEMENT>
バーコードなどを読み込み、これに基づく数値を比較して勝敗を決定していればよい。
</SUPPLEMENT>
```

Fig.4　Sample NTCIR-3 topic.

(2)　リコーの参加方針

このタスクは，第2章で述べた異種データ横断検索に相当し，検索語蒸留技術を重点的に適用した．その他，関連語により自動的に検索条件を拡張する方法など，これまで説明した技法を組み合わせて使用した．

(3)　評価結果

検索課題中の<ARTICLE>部と<SUPPLEMENT>部（Fig.4参照）を使用し検索語蒸留を用いないで検索した場合と用いて検索した場合の結果を比較するとTable 1のようになる．この結果から，検索語蒸留により，平均精度が約35%向上して

いることがわかる．

Table 1　Effect of term distillation.

	平均精度
検索語蒸留なし	0.1931
検索語蒸留あり	0.2637

他グループとの比較においても，われわれの検索語蒸留を用いた検索は優れた結果を示した．一例としてFig.5に，提出結果のうち，検索課題中の〈ARTICLE〉部と〈SUPPLEMENT〉部を用いて検索実行したもの上位10件分を示す．同じ実施要件で全8グループから18件の検索結果が提出された．リコーは前述のTDVの算出方法などを変えたものを4件提出した[9]．a，b，c，dがリコーの提出したものである．

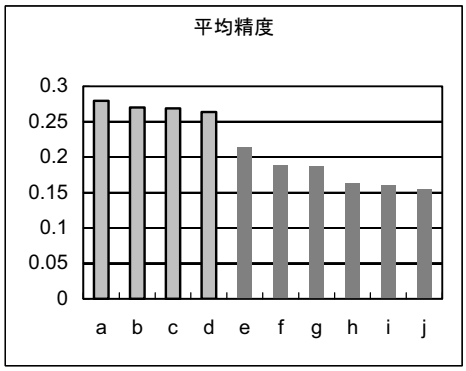

Fig.5　NTCIR-3 Patent task results.

リコーの検索システムは，他のグループに比べ非常に高い検索性能を示した．検索語蒸留が有効に効いたことが性能差の大きな要因とおもわれる．他のグループでは，異種データ横断検索であることに対する特別の対策は採られていないようであった．

3-3-2　Webタスク

(1)　タスク概要

Webタスク[14]も今回から新しく設定されたタスクである．Web上の文書に対する検索であり，検索対象データベースが極めて大きいことが特徴のひとつである．また，文書中に他の文書へのハイパーリンクが含まれることも，他のタスクの検索対象には見られない特徴である．

このタスクには，さらにいくつかのサブタスクが設定され，たとえば，検索条件に適合する文書があらかじめ一つだけわかっている，という想定のサブタスク，類似文書検索サブタスクなどがあった．また，検索実施要件として，ハイパーリンク情報を利用するもの，しないものなどが設定された．

検索課題は，47課題であった．出題フォーマットは特許タスクに類似したものである．検索対象は，主催者側で収集した，おもにjpドメインにあるWebページ約100G分および約10G分である．

(2)　リコーの参加方針

前述のように，Webの検索では，ハイパーリンク情報を利用できるという，通常の文書検索とは異なる特徴があり，この技術を持つ東京大学生産技術研究所喜連川研究室と共同で参加した[15]．まず，われわれの技術を用いてハイパーリンク情報を利用しない検索を実行し，その検索結果を東京大学のハイパーリンク分析技術により補正するという方式を採った．

われわれの課題としては，特許タスク同様，擬似適合性フィードバックによる検索条件の自動拡張の適用，および類似文書検索サブタスクにおいて，第2章で述べた適合性フィードバックと擬似適合性フィードバックの併用などを試みた．

(3)　評価結果

100Gデータ対象に検索課題中の〈TITLE〉部を使用し擬似適合性フィードバックによる検索条件の自動拡張を用いた場合と用いない場合を比較するとTable 2の結果が得られた．擬似適合性フィードバックによる拡張は非常に効果があったことがわかる．

Table 2　Effect of query expansion.

	平均精度
擬似適合性フィードバックによる拡張なし	0.1211
擬似適合性フィードバックによる拡張あり	0.1506

他グループとの比較でも，リコーの方式は高い性能を示した．ここでは，100Gデータ対象のトピック検索とよばれるサブタスクでの例を挙げる．この実施要件では，全6グループ，25件の検索結果が提出され，リコーからは，ハイパーリンク情報を利用しないで（リコーの技術のみを用いて）検索したもの2件とハイパーリンク情報を利用して（リ

コーの技術に東京大学の技術を併用して）検索したもの2件を提出した[15]．Fig.6では，ハイパーリンク情報を利用しない検索について評価した結果の上位10件を挙げている．aとbがリコーの提出したものである．

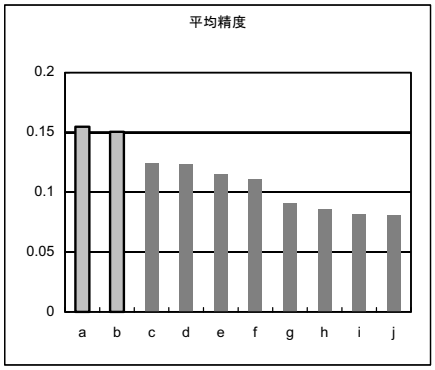

Fig.6　NTCIR-3 Web task results.

Web検索でも，リコーは，他のグループに比べ高い検索性能を示していることがわかる．

4. 応用

4-1　全文検索サーバーFTS*

われわれの開発したランキング検索技術は，ソフトウェア研究開発本部で開発した全文検索サーバーFTSに実装され，さまざまなアプリケーションで利用されている．

パッケージ製品では，オフィス文書管理システムRidoc Document Server Pro V2に搭載されている．また，（株）三愛・ギガネットワークスカンパニーのLモード公式サイトでは歌詞の断片から曲を検索するサービスに利用されている．他にも，リコーのWWWサイトR-FLASH内のページ分類・検索システムCrowwwでも使用している．

また，必要な文書を漏れなく効率よく検索できる検索性能の高さを活かし，特許検索への応用をすすめている．以下，簡単に紹介する．

4-2　公開特許公報ランキング検索サービス

FTSの特許検索へ応用として，イントラネット上で公開特許公報をランキング検索できるシステムを開発した．特に特許出願時の先行技術調査での利用を念頭に置いており，出願用の請求項をもとに同じ技術が既に特許出願されていないか探し出すといった使い方ができるようになっている．検索対象としては，93年以降の公開特許公報約360万件の要約と請求項などが登録されている．以下，典型的な使い方を示す．

Fig.7は，出願用の請求項を入力したところである．システムは入力された請求項から検索語として有効なものを取り出し検索を実行，検索された特許公報をランキング順にFig.8のように表示する．ユーザは，検索結果を確認し，探している内容に適合する公報が見つかったら，それをシステムに指示する．

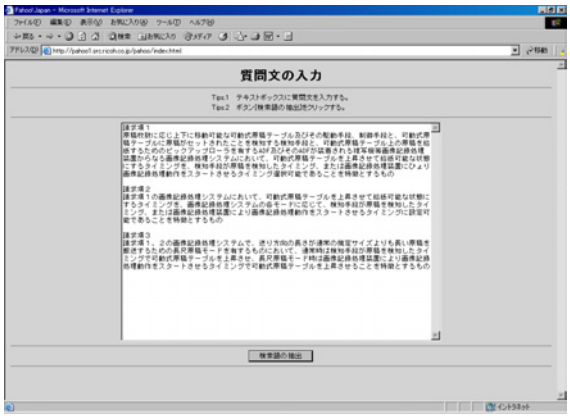

Fig.7　Sample input.

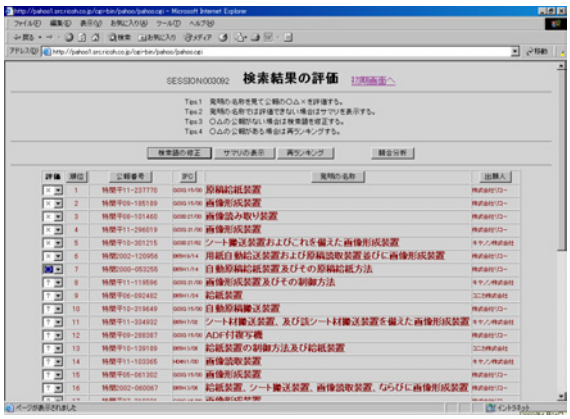

Fig.8　Sample output.

システムは，ユーザの指示した公報を適合性フィードバックとして用いて再検索を実行する．関連語が自動的に追加され，ユーザの指示した公報に近いものが上位にランクされるようになる．

* FTSはTRMeisterとして現在，商標登録出願中である．

プロトタイプを現在，ソフトウェア研究開発本部などで試用中である．

5. 今後の展開

リコーのランキング検索技術について各要素技術，性能評価，およびその応用について述べた．

今後は，検索ノイズを減らしたり検索文書にラベル付けをおこなえるよう検索文書の分類技術も研究していく予定である．また，社内のマニュアル・保守サービス情報検索への応用を検討している．

参考文献

1) W.B. Frakes and R. Baeza-Yates (ed.): Information Retrieval, Prentice Hall, (1992)

2) S.E. Robertson and K. Sparck Jones: Relevance weighting of search terms, Journal of American Society for Information Science, 27, (1976), pp. 129-146

3) S.E. Robertson and S. Walker: On relevance weights with little relevance information, Proceedings of the 20th Annual International ACM SIGIR Conference (SIGIR '97), (1997), pp. 16-24

4) K. Sparck Jones: A Statistical Interpretation of Term Specificity and Its Application in Retrieval, Journal of Documentation, 28, 1, (1972), pp. 11-21

5) Y. Ogawa, H. Mano, M. Narita and S. Honma: Structuring and expanding queries in the probabilistic model, Proceedings of The Eighth Text REtrieval Conference (TREC-8), (2000), pp. 541-548

6) 長尾真 (監修): 日本語情報処理, 2版, 電子情報通信学会, (1985)

7) 望主雅子, 亀田雅之: 品詞体系の変換を考慮した中間品詞体系と品詞変換表の設計, 言語処理学会第4回大会, (1998)

8) 望主雅子, 伊藤篤: 日本語形態素解析における素性の導入, 情報処理学会第43回大会, (1991)

9) H. Itoh, H. Mano and Y. Ogawa: Term Distillation for Cross-DB Retrieval, Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, (2003予定)

10) S.E. Robertson: On term selection for query expansion, Journal of Documentation, 46, 4, (1990), pp. 359-364

11) 徳永健伸: 情報検索と言語処理, 初版, 東京大学出版会 (1999)

12) 岸田和明, 岩山真, 江口浩二: 検索実験の方法と実際：NTCIRワークショップでの試み, NTCIR-3 Workshop 前講演資料, (2002)

13) M. Iwayama, A. Fujii, N. Kando and A. Takano: Overview of Patent Retrieval Task at NTCIR-3, Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, (2003予定)

14) K. Eguchi, K. Oyama, E. Ishida, N. Kando and K. Kuriyama: Overview of Web Retrieval Task at the Third NTCIR Workshop, Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, (2003予定)

15) M. Toyoda, M. Kitsuregawa, H. Mano, H. Itoh and Y. Ogawa: University of Tokyo/RICOH at NTCIR-3 Web Retrieval Task, Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, (2003予定)

注1) LモードはNTTの登録商標です．