データベースシステム 118－3
情 報 学 基 礎 54－3
（1999．5．17）

文書間の類似度における単語寄与度を利用した検索式拡張手法

帆足 啓一郎　松本 一則　井ノ上 直己　橋本 和夫

KDD 研究所

〒356-8502 上福岡市大原 2-1-15

E-mail: {hoashi,matsu,inoue,kh}@kddlabs.co.jp

テキスト情報検索システムへの入力文から生成される検索式の情報を拡張することによってより高い精度の検索を図る「検索式拡張」の有効性はさまざまな研究発表などによって実証されている．しかし，検索式拡張に使用される単語の選択時には，TF*IDF など，検索対象文書内での重要性のみが考慮される場合が多く，その単語が入力文と検索対象文書の間の類似度に与える影響は考慮されていない．そのため，検索式拡張において有効な単語が選択されていない可能性があると考えられる．本研究では類似度への単語の影響力を数値化した「単語寄与度」という概念を定義し，単語寄与度に基づいた新たな検索式拡張手法を提案する．また，Rocchio のアルゴリズムに基づく検索式拡張との比較実験を通し，提案手法の有効性を示す．

Query Expansion Method Based on Word Contribution to Query-Document Similarity

Keiichiro HOASHI　Kazunori MATSUMOTO　Naomi INOUE　Kazuo HASHIMOTO

KDD R&D Laboratories, Inc.

2-1-15 Ohara, Kamifukuoka, Saitama 356-8502 JAPAN

E-mail: {hoashi,matsu,inoue,kh}@kddlabs.co.jp

In this paper, we propose a novel query expansion method based on a measure called *word contribution*. Word contribution is a measure which expresses the influence a word has on the similarity between a query and a document. We presumed that such words with significant negative contribution to the similarity of documents are discriminative words of document relevance. Therefore, by extracting such words from documents relevant to the query, it is possible to make an effective query expansion.

We describe the experiments for the evaluation of our proposed query expansion method, which was made on TREC data. Through the comparison of our method to the Rocchio-weight based query expansion method, the effectiveness of our method was proved.

1 はじめに

情報検索システムから有効な検索結果を得るためには，効果的な入力文，あるいは検索式 (query) の作成が重要であることは言うまでもない．しかし，検索システムに対する一般的なユーザの入力はせいぜい数単語程度のみで構成されることが多いため，不要な情報が多く提示されるなど，充分な検索結果が得られないのが現状である．

こうした問題に対処するため，近年，この検索式の情報を自動的に拡大する「検索式拡張」(query expansion) の研究がさかんに行われており，TREC など数多くの会議でその有効性を示す研究が発表されている [1]-[4]．入力文に類似している文書から抽出された語であれば，これらの語を検索式に加えることは検索精度向上に貢献するとされていることから [5]，検索式拡張の導入は高精度な検索システム設計のためには非常に有効な手段であるといえる．

前述の説明から明らかなように，検索式拡張を行うためには入力文に対する類似文書を獲得する必要がある．この類似文書集合を獲得するために一般的に広く利用されている手法の一つとして「適合フィードバック」(relevance feedback) があげられる [6]．適合フィードバックでは，まず，拡張される前の検索式を用いた初期検索を行う．次に，初期検索の結果，上位にランクされた文書について類似性の判断を行い，その情報をシステムにフィードバックする．検索式拡張はこのフィードバックされた類似文書集合から単語を抽出し，元の検索式に加えることによって行われる．適合フィードバックの手法としては，ユーザが上位文書の類似性を判断し，その判断結果をシステムに返す手法 (manual feedback)[7] と，初期検索の結果，上位にランクされた文書を類似文書とみなし，その情報をシステムに返す手法 (pseudo feedback)[8] の 2 つの手法が提案されている．manual feedback は，初期検索の結果得られた文書に対し正確な類似性の評価が行われるため，検索式拡張がより有効になるという長所がある反面，類似性判断の負担がユーザにかけられてしまうという欠点がある．一方，pseudo feedback ではユーザへの負担は軽減されるものの，フィードバックされる類似性の判断が完全ではないぶん，検索式拡張後の検索精度が劣化するという短所があげられる．

本研究では，これまで説明した手法のうち，manual feedback をベースにした新たな検索式拡張手法を提案する．さらに，TREC データを使用した従来手法との比較実験を行い，その有効性を示す．

2 従来手法

現在，最も有効な検索式拡張手法の一つとされている手法の一つに Rocchio のアルゴリズムに基づいた手法があげられる [6]．Rocchio のアルゴリズムは 1960 年代半ばに提案されており，現在に至るまで SMART [9] など数多くの検索システムに採用されている手法である．

Rocchio の手法は，検索対象文書などをベクトルとして表現するベクトル空間モデルに基づいており，ある入力文に対する類似性が既知の場合，その入力文を表す最適なベクトルとは類似している文書との類似度を最大にし，かつ非類似文書との類似度を最小化するものである…という思想に基づいて提案された手法である [10]．そして，この最適なベクトルは類似文書を表すベクトルの重心と，非類似文書を表すベクトルの重心との差分ベクトルであるとしている．したがって，最適なベクトル $\vec{Q}_{opt}$ は以下の数式によって表される．

$$\vec{Q}_{opt} = \frac{1}{R}\sum_{D\in Rel}\vec{D} - \frac{1}{N}\sum_{D\notin Rel}\vec{D}$$

但し，R, N はそれぞれ検索対象文書中の類似文書，非類似文書の数を表し，Rel は類似文書を表すベクトルの集合とする．上記の計算の結果，値が負になったベクトルの要素はその値を 0 とする．

この最適ベクトルの算出は元の入力文のベクトルを類似文書を表すベクトルに近づけるとともに，非類似文書のベクトルから遠ざけてる手法といえる．しかし，この過程では元の入力文を表すベクトルの特徴が反映されていない．そこで，最適ベクトルを算出する際に元の入力文のベクトルの特徴を取り入れた手法も開発されている [11]．以下にその数式を示す．

$$\vec{Q}_{new} = \alpha \times \vec{Q}_{org} + \beta \times \frac{1}{R} \sum_{D \in Rel} \vec{D} - \gamma \times \frac{1}{N} \sum_{D \notin Rel} \vec{D}$$

なお，この式には，元のベクトル，類似文書のベクトル，および非類似文書のベクトルの影響を制御するための係数α, β, γが付与されている．前節で述べた検索システム SMART はこの式に基づく検索式拡張手法を使用しており，高い検索精度が確認されている．

このように，Rocchio のアルゴリズムに基づく検索式拡張手法の有効性は多くの研究発表などで実証されている．しかし，この手法では各語の文書内における重要性（ベクトルの各要素の重み）についてしか考慮していないが，ある文書の中での重要性が高い単語でも，その文書と入力文書との間の類似度において大きな影響を与えていない可能性があり，また，文書内での重要性が低い語でも，類似文書との類似度に対し高い影響を与えている単語がある可能性がある．したがって Rocchio の手法のように，検索式拡張の際に選択する単語の基準として文書内での重要性のみを考慮するのでは充分な検索精度が得られない可能性がある．

3 単語寄与度に基づく検索式拡張手法の提案

前節で述べた従来手法の問題点をふまえ，本研究では文書間の類似度における影響を数値化した「単語寄与度」という概念を定義し，この単語寄与度に基づいた新たな検索式拡張手法を提案する．本節では単語寄与度についての説明を行い，続いて提案手法について述べる．

3.1 単語寄与度の定義

単語寄与度とは，前述したように，文書間の類似度における各単語の影響を数値化した尺度である．ある入力文書 qと検索対象文書 d との間の類似度における単語 wの単語寄与度は以下の数式によって定義される [12].

$$Cont(w, q, d) = Sim(q, d) - Sim(q'(w), d'(w))$$

表 1: 入力文書と類似文書の類似度における単語寄与度の例

Word	Contribution
levitation	0.08039449
superconductivity	0.02394392
phenomenon	0.02052002
application	0.00886015
possible	0.00309428
use	0.00258170
government	0.00058935
company	-0.00003276
text	-0.00003481
BFN	-0.00003655
...	...
commercial	-0.00194345
narrative	-0.00195197
hanging	-0.00246844
permanent	-0.00307957
Kanagawa	-0.00312267
iron	-0.00496679
flywheel	-0.00514038
magnet	-0.01156134
superconductor	-0.01881981
Maglev	-0.07156282

ただし，$Sim(q, d)$ は q,d 間の類似度を表し，$q'(w)$ は q から単語 w が除かれた入力文書，$d'(w)$ は d から単語 w を除いた文書とする．すなわち，単語寄与度 $Cont(w, q, d)$ とは，qと dとの類似度と単語 wが存在しない場合の qと dとの類似度との差である．したがって，qと dに出現する全ての単語のうち，類似度を向上させる単語の寄与度は正であり，逆に類似度を下げる単語の寄与度は負である．表 1に，TREC データの入力文書 (Topic 313) およびこの入力文書に類似している文書 (FBIS3-30043) との類似度における単語寄与度の例を示す．

3.2 単語寄与度の分析

図 1は，表 1と同じ入力文書および検索対象文書中に出現する全ての単語の寄与度を降順に左から並べたものである．

この図より，出現単語のうち類似度に有意な影響を与えている単語は少なく，大多数の単語は類似度にほとんど無関係であることがわかる．類似度に関係のある単語のうち単語寄与度が正である単語は，単語寄与度の定義より入力文と検索対象文書に共起している単

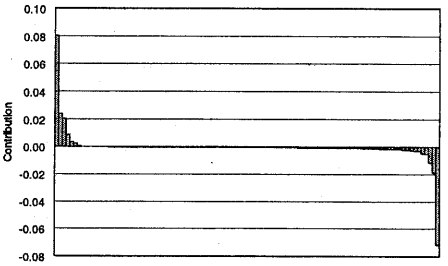

図 1: 入力文と類似文書の類似度における単語寄与度

語である．これに対し，単語寄与度が負の単語は入力文と検索対象文書に共起していない単語であり，かつ類似度に大きな影響を与えている単語であることから，検索対象文書の特徴を顕著に表している．

検索式拡張の最終的な目的は，検索対象文書集合の中から類似文書を抽出するのに効果的な単語を元の検索式に加えることである．したがって，単語寄与度が負の単語を検索式拡張に使用することにより，効果的な検索式拡張が実現できると考えられる．さらに，単語寄与度の絶対値がその単語の影響力の大きさを表していると仮定すると，単語寄与度の値をそのまま検索式拡張時に利用することも有効であるといえる．

3.3 提案手法

前節までに説明した仮定をもとに，単語寄与度を用いた新たな検索式拡張手法を提案する．

まず，入力文 q と類似している文書群 $D_{rel}(q) = \{d_1, \cdots, d_{Num}\}$ 中の各文書に出現する全ての単語の寄与度を求め，各類似文書から単語寄与度の低い単語を N 個抽出する．次に抽出された各単語の寄与度の総和に重み wgt をかけ，これを単語 w に対するスコアとする．単語 w の入力文 q と文書 d の類似度に対する寄与度を $Cont(w,q,d)$ とすると，単語 w のスコア $Score(w)$ は以下の数式によって表される．

$$Score(w) = wgt \times \sum_{d \in D_{rel}(q)} Cont(w,q,d)$$

最後に，抽出された単語のうち元の検索式に含まれていない単語とそのスコアを検索式に加える．なお，抽出された単語の寄与度の値は負であるため，wgt も負の値に設定する．また，スコアが負の単語は検索式拡張には使用しないものとする．

4 評価実験

提案手法の有効性を示すため，評価実験を行った．以下，この評価実験について述べる．

4.1 データ概要

本実験では TREC-6 のデータを使用した [4]．すなわち，入力文書は Topic 301-350 の 50 文書，検索対象文書は TREC CD-ROM Vol 4, 5 から *Congressional Records* を除いたおよそ 53 万個の文書である．これら全ての文書に対し，形態素解析によって名詞，固有名詞および未定義語を抽出し，各文書を表すベクトルの要素とした．また，適合フィードバックや検索結果野評価では TREC から提供されている各入力文書に対する類似文書一覧を使用した．

4.2 検索方法

ここでは本実験での検索方法について詳しい説明を行う．

4.2.1 類似度計算方法

本実験では，前述のように，ベクトル空間モデルに基づいた検索を行っている．各文書を表すベクトルの要素は TF*IDF を求めることによって計算する．以下に本実験で使用した TF および IDF の計算式を示す．

- TF factor
$$1 + \log(1 + tf)$$
- IDF factor
$$\log\left(\frac{M}{df}\right)$$

但し，tf は文書内の単語出現頻度，df は単語が出現する文書数，M は検索対象文書集合に含まれる文書数とする．TF の計算の際，tf に 1 を加えた値を使用しているが，これは単語寄与度による検索式拡張の際に tf が 1 未満になる（すなわち，$\log(tf)$ が負になる）単語に対処するためである．

類似度は入力文と検索対象文書のベクトルのコサイン値によって求める [13]. 以下にその数式を示す.

$$\cos(\vec{Q},\vec{D}) = \frac{\vec{Q}\cdot\vec{D}}{|\vec{Q}||\vec{D}|}$$

但し, $\vec{Q},\vec{D}$ はそれぞれ入力文書と検索対象文書を表すベクトルとし, $|\vec{D}|$ は $\vec{D}$ のユークリッド長とする.

4.2.2 検索過程

本実験で最終的な検索結果を得るまでの過程は以下の通りである.

1. **初期検索**
   前述したように, まずは元の検索式を使用して初期検索を行う. 初期検索の結果, 類似度の上位 1000 件の文書を抽出する. 以下, この文書集合を「上位 1000 件文書集合」と呼ぶ.

2. **検索式拡張**
   上位 1000 件文書集合より, 類似文書を抽出し, 検索式拡張を行う. 本実験では 2 つの手法を用いて類似文書集合を抽出する.

   手法 A 上位 1000 件文書集合に含まれる類似文書のうち, 類似度の上位 Num 個の文書を類似文書集合として抽出する. また, Rocchio の手法で使用される非類似文書集合は, 非類似文書のうち類似度の上位 500 個の文書とする.

   手法 B 上位 1000 件文書集合のうち, 類似度の上位 20 件のみから類似文書集合ならびに非類似文書集合を抽出する. すなわち, 上位 20 件の文書のうち, 入力文に類似している文書を類似文書集合とし, 残りの文書を非類似文書集合とする.

   なお, Rocchio の手法で使用されている係数 α,β,γ については, TREC-7 にて発表された SMART システムの設定にならい, $\alpha=3,\beta=2,\gamma=2$ とする. また, 各々の元の検索式に加えられる単語も, TREC-7 の SMART 同様, Rocchio の手法で算出された重みの高かった単語の上位 20 個とする.

3. **最終検索**
   検索式拡張の結果, 作成された検索式をもとに検索を行い, 最終検索結果を得る.

5 評価結果

以下, 手法 A および手法 B による検索式拡張手法の実験結果について述べる.

5.1 手法 A の評価結果

手法 A に基づく提案手法には, 正解文書集合に含まれる文書の数 Num, 各類似文書から抽出される単語の数 N, 抽出された単語の寄与度に対する重み wgt の 3 つのパラメータがある. ここでは, $N=10$ と固定し, Num と wgt のみを調整して実験を行った.

$Num=10$ および $Num=20$ の場合の提案手法ならびに Rocchio の手法による検索の平均精度 (average precision) を表 2に示す. また, 比較のため初期検索の平均精度も示す. 表の中では, $Num=10$ の場合の提案手法を "WC10", Rocchio の手法を "Roc10", $Num=20$ の場合の提案手法を "WC20", Rocchio の手法を "Roc20", また初期検索を "Baseline" と表す.

表 2: 手法 A での各検索式拡張手法の平均精度

wgt	-100	-400	-1200	-2000	-3000
WC10	0.3696	0.3837	0.3844	0.3837	0.3854
WC20	0.4265	0.4475	0.4528	0.4530	0.4554
Roc10			0.3441		
Roc20			0.3789		
Baseline			0.1433		

表 2に示された結果から明らかなように, 提案手法による検索式拡張を行った結果, 初期検索と比較して $Num=10$ の場合は 157.9%～168.2%, $Num=20$ の場合は 197.6%～216.1% という, 高い検索精度向上が得られた. また, Rocchio の手法と比較しても高い検索精度が得られていることから, 提案手法による単語

抽出および抽出された単語に対する重み付けが有効であることが示された．

詳しい分析のため $Num = 10, 20$ のそれぞれについて提案手法 ($wgt = -1200$)，Rocchio の手法，ならびに初期検索の 3 つの検索手法の Recall-Precision 曲線を示す．図 2には $Num = 10$，図 3には $Num = 20$ の場合の曲線を示す．

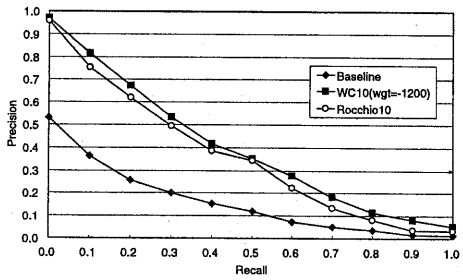

図 2: 各手法の Precision と Recall(手法 A,Num=10)

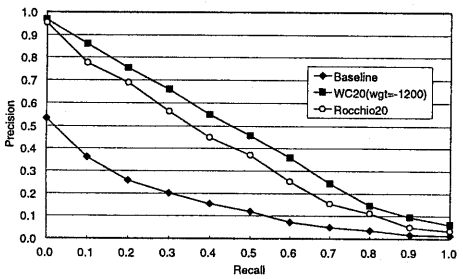

図 3: 各手法の Precision と Recall(手法 A,Num=20)

これらの Precision-Recall 曲線より，どの再現率においても 提案手法の精度が Rocchio の手法の精度を上回っていることがわかる．以上の結果より，提案手法の有効性が確認された．

5.2 手法 B の評価結果

手法 B では $N = 10$, $Num = 20$ と固定し， wgt のみを調整して実験を行った．

表 3に提案手法ならびに Rocchio の手法による検索の平均精度を示す．各検索手法の表記方法は表 2と同様とする．

表 3: 手法 B での各検索式拡張手法の平均精度

wgt	-100	-400	-1200	-2000	-3000
WC20	0.2540	0.2407	0.2335	0.2320	0.2307
Roc20			0.2310		
Baseline			0.1433		

ここでも提案手法による検索式拡張の結果，初期検索に対して 61.9%～77.5%の検索精度向上が得られていることがわかった．また，手法 A と同様，Rocchio の手法を上回る検索精度を得ることが出来た．

次に，提案手法 ($wgt = -100$)，Rocchio の手法，ならびに初期検索の Precision-Recall 曲線を図 4に示す．

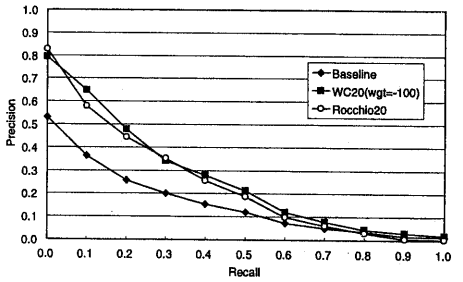

図 4: 各手法の Precision と Recall(手法 B,Num=20)

この Precision-Recall 曲線からは，手法 A の曲線ほど提案手法と Rocchio の手法との間に明確な差は見出すことができない．全体的には提案手法の方が精度が高いものの，Recall=0.0 の時点では Rocchio の手法の方が精度が高いことがわかる．

ここで，$wgt = -1200$ とした場合の提案手法と Rocchio の手法の Precision-Recall 曲線を図 5に示す．

この図より，wgt の絶対値を上げることにより，平均精度は下がるものの，Recall=0.0 での提案手法の精度を Rocchio の手法並みに上げられることが確認された．また，この wgt 調整を行いつつも，Rocchio の手法より高い平均精度が保たれていることがわかる．以上の結果より，本実験でも提案手法の有効性が示された．

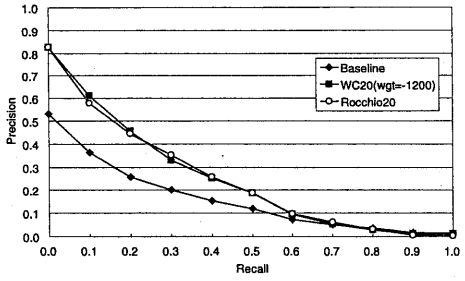

図 5: 各手法の Precision と Recall(手法 B,Num=20)

6 結論

本研究では 2 つの類似文書集合抽出手法に基づき，評価実験を行った．手法 A は検索式拡張に充分な適合フィードバックを与える手法であり，手法 B は類似文書集合が抽出される文書群を制限しているため，実際にフィードバックを行うユーザがいる状態をシミュレートした手法である．

手法 A に基づいた検索式拡張の実験の結果，提案手法では従来手法を大きく越える検索精度が得られた．したがって，本実験の結果から，適合フィードバックが充分な条件下では，従来手法に対する優位性が実証された．一方，手法 B に基づいた実験の結果，手法 A ほどの精度向上は得られなかったものの，Rocchio の手法と比較して高い検索精度は実現されており，提案手法の有効性が確認された．

以上，2 点の評価結果により，単語寄与度に基づく検索式拡張手法の有効性が示された．

7 考察

ここでは評価実験の結果を詳細に分析し，本研究の今後の課題を示す．

今回行われた 2 つの評価実験のうち，手法 B では提案手法と従来手法との間に手法 A で確認されたほどの性能の差は見出されなかった．この理由の一つとして，初期検索の精度が低いことがあげられる．本実験での初期検索の上位 20 位までの平均精度は 0.2450 であった．すなわち，各々の入力文書に対する検索式拡張で使用される類似文書集合に含まれる文書の数は平均 4.9 個であり，手法 A と比べてかなり少ない情報を元に検索式拡張が行われていたことがわかる．

表 4に各入力文書に対する初期検索の結果，上位 20 位までに含まれていた類似文書数の分布を示す．

表 4: 各入力文書に対する上位 20 位までの類似文書数

類似文書数	入力文書数
16 以上	2
11-15	6
6-10	10
1-5	23
0	9
合計	50

この表から明らかなように，50 個の入力文書中，32 個の文書に対しては 5 個以下の類似文書から検索式拡張が行われており，そのうち 9 個の入力文書は上位 20 件の中に類似文書が含まれていないため，提案手法では検索式拡張を行うことが出来なかった．

この問題への対処法は 2 通り考えられる．1 つは，初期検索の精度を向上させることである．今回の実験では，検索対象文書に対する形態素解析の結果，名詞などを抽出して各文書をベクトル化するという，単純なシステムによって文書のインデクシングを行っている．そのため，同じアルゴリズムで検索を行っている SMART に関する論文などで発表されている初期検索の精度を下回る精度しか得られていない．今後は辞書データの改良や単語の共起情報を利用することなどにより，初期検索の精度向上を図る必要があると思われる．

もう 1 つの対処法として，Rocchio の手法のように非類似文書集合の情報を利用することが考えられる．非類似文書集合の情報の検索式拡張への具体的な利用については今後の課題の一つと考える．

また，全体的な課題として，本実験で採用したコサイン値による類似度計算だけでなく，さまざまな類似度計算手法に基づいた評価実験を行い，より一般的な有効性を示す必要がある．単語寄与度自体は類似度計算手法を問わず算出が可能な尺度であるため，他の手法に基づいたシステムへの適用が容易である．この利点をふまえ，ベクトル空間モデルに基づく他の検索

アルゴリズムだけでなく，確率モデルに基づいた検索アルゴリズムなどにも適用し，評価実験を行う予定である．

参考文献

[1] D Harman, "Overview of the Third Text REtrieval Conference", NIST SP 500-226, 1994.

[2] D Harman, "The Fourth Text REtrieval Conference ", NIST SP 500-236, 1995.

[3] E Voorhees and D Harman, "The Fifth Text REtrieval Conference", NIST SP 500-238, 1996.

[4] E Voorhees and D Harman, "The Sixth Text REtrieval Conference", NIST SP 500-240, 1997.

[5] C Buckley and G Salton, "Optimization of Relevance Feedback Weights", Proceedings of SIGIR'95, pp 351-357, 1995.

[6] J Rocchio: "Relevance Feedback in Information Retrieval", in "The SMART Retrieval System – Experiments in Automatic Document Processing", Prentice Hall Inc., pp 313-323, 1971.

[7] G Salton, "Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer", Addison-Wesley, 1988.

[8] S Robertson, S Walker, S Jones, M Hancock-Beaulieu, and M Gatford, "Okapi at TREC-3", Overview of the Third Text REtrieval Conference, pp 109-125, 1994.

[9] A Singhal, J Choi, D Hindle, D Lewis, and F Pereira: "AT&T at TREC-7", The Seventh Text REtrieval Conference, 1998. (to be published)

[10] A Singhal, M Mitra, and C Buckley, "Learning Routing Queries in a Query Zone", Proceedings of SIGIR'97, pp 25-32, 1997.

[11] G Salton and C Buckley, "Improving Retrieval Performance by Relevance Feedback", Journal of the American Society for Information Science, 41(4):288-297, 1990.

[12] 帆足，松本，青木，橋本: "テキストの絞り込み検索のための特徴抽出手法の検討"，情報処理学会第 56 回全国大会講演論文集，Vol.3, pp 124-125, 1998.

[13] I Witten, A Moffat, and T Bell: "Managing Gigabytes: Compressing and Indexing Documents and Images", Van Nostrand Reinhold, 1994.