

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[†]

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事例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、適合文書集合における平均的な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個々の適合文書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[†]
and KAZUO HASHIMOTO[†]

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数単語程度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、十分な検索結果が得られないのが現状である。

このような問題に対処するため、近年、検索式の情報を拡大する「検索式拡張」(query expansion) の研究がさかに行われており、TREC^{(1)~(4)} など数多くの会議でその有効性を示す研究事例が報告されている。検索式拡張は、一般的に入力文を表す検索式に新たに

語を加えるという処理によって実現される。この際、入力文に適合している文書 (以下、適合文書) から抽出された語を検索式に加えることにより検索精度が向上すると報告されており⁵⁾、高精度な検索システムを実現するためには検索式拡張の導入が有効な手段であると考えられている。

検索式拡張を行うためには入力文に対する適合文書集合を獲得する必要がある。この適合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に「適合フィードバック」(relevance feedback) がある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適合性の判断を行い、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適合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

[†] 株式会社 KDD 研究所
KDD R&D Laboratories, Inc.

書の適合性を判断し、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適合文書とみなし、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適合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利点がある反面、適合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適合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 manual feedback による検索式拡張に比べ劣化するという欠点がある。

本論文では、これまで説明した手法のうち、manual feedback をベースにした検索式拡張手法について検討を行う。まず、入力文と適合文書との類似度に対し、それぞれの文書に出現する単語の影響を表す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。次に、TREC データを使用した従来手法との比較実験を行い、提案手法の有効性を示す。

2. 従来手法

現在、最も有効な検索式拡張手法の一つに Rocchio のアルゴリズムに基づいた手法がある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 SMART⁹⁾など数多くの検索システムに採用されている。

Rocchio のアルゴリズムは、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いる。「ある入力文に対する適合性が既知の場合、その入力文を表す最適ベクトルとはその入力文に対する適合文書との類似度を最大にし、かつ非適合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り¹⁰⁾、この最適ベクトルは適合文書集合中の文書を表すベクトル群の重心と、非適合文書を表すベクトル群の重心との差分ベクトルであるとしている。Rocchio によると、ここでの最適ベクトル $\vec{Q}_{opt}$ は式 (1) によって表される。

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

但し、 R , N はそれぞれ検索対象文書集合中の適合文書、非適合文書の数を表し、 Rel は適合文書を表すベクトルの集合とする。式 (1) 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。

式 (1) の最適ベクトルの算出は、元の入力文のベク

トルを適合文書を表すベクトルに近づけるとともに、非適合文書のベクトルから遠ざけるという効果を持つ。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⁾。修正された最適ベクトルの定義を式 (2) に示す。

$$\begin{aligned} \vec{Q}_{opt} = & \\ & \alpha \times \vec{Q}_{org} + \beta \times \frac{1}{R} \sum_{D \in Rel} \vec{D} \\ & - \gamma \times \frac{1}{N} \sum_{D \notin Rel} \vec{D} \end{aligned} \quad (2)$$

式 (2) には、元の入力文、適合文書のベクトル、および非適合文書のベクトルの影響を制御するためのパラメータ α, β, γ が付与されている。SMART はこの式に基づく検索式拡張手法を使用しており、この手法により高い検索精度を得ている。

式 (2) より明らかなように、Rocchio のアルゴリズムに基づく検索式拡張では、入力文、適合文書集合および非適合文書集合に含まれる文書を表すベクトルを元に検索式拡張を行う。これらのベクトルの要素は、各文書に出現する単語を表しており、各要素に格納されている値はその要素が表す単語の文書内での重要度を表している。Rocchio の手法では、式 (2) の第 2 項に見られるように、この重要度を適合文書集合に対して平均して取り扱っている。したがって平均的な適合文書における単語の重要度に基づいて検索式拡張を行っていることになる。

しかし、ある適合文書内で重要度の高い単語が、別の適合文書内でも重要度が高いとは限らない。また、入力文に対する適合性が高い文書に含まれる重要語をより重視すべき場合も考えられる。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適合文書集合での平均的重要度のみを考慮するのではなく、入力文と個々の適合文書との類似度における各単語の影響を考慮したうえで検索式拡張を行えば、検索性能の一層の向上が期待できる。

検索に対する個々の単語の影響を測る基準としては、Salton が提案した term discrimination value という概念があげられる⁷⁾。Term discrimination value とは、ある 1 つの単語が個々の文書を識別する能力を数値化したものである。文書識別能力の高い単語を重要

★ 本論文では Rocchio の手法での非適合文書集合の利用の効果に関する詳細な議論は行わないが、5.3.1 節に関連情報を記述した。

視することにより、個々の検索対象文書の特徴が明確になり、検索精度が向上するという考えである。

Term discrimination value は、その単語が検索対象文書を表す要素として加えられる前後の文書空間密度の差として計算することができる。文書識別能力の高い単語が文書表現の要素として加えられた場合、個々の文書の特徴が明確になるため、各文書間の距離は増加する。この結果、空間密度が減少することになり、term discrimination value が増加する。逆に、文書識別能力の低い単語が文書空間に加えられた場合は、文書間の距離が縮小するため、term discrimination value が減少する。単語 w が加えられる以前の空間密度を Q 、単語 w が加えられた後の空間密度を Q_w とすると、単語 w の term discrimination value dv_w は式 (3) によって定義される。

$$dv_w = Q - Q_w \quad (3)$$

空間密度 Q は空間内の全ての文書間の類似度の平均によって表すことができる。これを式 (4) に示す。

$$Q = \frac{1}{M(M-1)} \sum_{i=1}^M \sum_{k=1, i \neq k}^M Sim(D_i, D_k) \quad (4)$$

ここで、 $Sim(D_i, D_k)$ は文書 D_i と文書 D_k との間の類似度を表し、 M は空間内の全文書数を表すものとする。

また、検索対象文書集合中の全ての文書間の類似度を計算する代わりに、検索対象文書集合の重心 C との類似度を計算することにより空間密度計算を近似する手法も提案されている。これを式 (5) に示す。

$$Q = \frac{1}{M} \sum_{k=1}^M Sim(C, D_k) \quad (5)$$

しかし、式 (4) より明らかなように、term discrimination value を計算するのに必要な空間密度 Q を算出するためには類似度の計算を $M(M-1)$ 回行わなくてはならない。また、式 (5) のように空間密度計算を簡略化した場合でも M 回の類似度計算が必要となる。したがって検索対象文書数が多い一般的な情報検索の検索式拡張に term discrimination value を適用するためには膨大な計算量が必要である。

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解決するため、ここでは入力文と検索対象文書との類似度における個々の単語の影響を数値化した「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

表 1 入力文と適合文書の類似度における単語寄与度の例
Table 1 Data example of word contribution between query and relevant document

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
BFII	-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
Haglev	-0.07156282

明を行い、続いて提案する検索式拡張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 q と検索対象文書 d との間の類似度における単語 w の単語寄与度を式 (6) によって定義する¹²⁾。

$$Cont(w, q, d) = Sim(q, d) - Sim(q'(w), d'(w)) \quad (6)$$

ここで、 $Sim(q, d)$ 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、 $d'(w)$ 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効果がある単語の寄与度は正となり、逆に類似度を下げる効果のある単語の寄与度は負となる。

表 1 に、TREC データの入力文 (Topic 313) およびこの入力文に対する適合文書 (FBIS3-30043) との類似度における単語寄与度の例を示す。また、図 1 および図 2 にそれぞれ Topic 313 と FBIS3-30043 の一部を示す。

3.2 単語寄与度の分析

図 3 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。

```
(num) Number: 313
(title) Magnetic Levitation-Maglev

(desc) Description:
Commercial uses of Magnetic Levitation.

(narr) Narrative:
A relevant document must contain Magnetic Levitation or
Maglev. It should be concerned with possible commercial
applications of this phenomenon to include primarily mass
transit, but also other commercial applications such as
Maglev flywheels for cars. Discussions of superconductivity
when linked to Maglev and government support plans
when linked to Maglev are also relevant.
```

図1 Topic 313
Fig. 1 Topic 313

```
(DOC)
(DOCNO) FBIS3-30043 (/DOCNO)
(HT) "dreas037_a94011" (/HT)

(HEADER)
(AU) FBIS-EAS-94-037-A (/AU)
Document Type:Daily Report
(DATE1) 24 February 1994 (/DATE1)

(/HEADER)

(...略...)

(TEXT)
Language: (F P=105) Japanese (/F)
Article Type:BfN

[Text] A piece of iron hovers in the air... The Kanagawa
Academy of Science and Technology (KAST), established
by Kanagawa Prefecture, has discovered a new magnetic
levitating phenomenon that uses high-temperature super-
conductors, and confirmed that practical uses are possible.
Previously, it was impossible to levitate something unless
a permanent magnet was used together with a supercon-
ductor.

(...略...)

(/TEXT)

(/DOC)
```

図2 FBIS3-30043 (一部)
Fig. 2 FBIS3-30043 (partial)

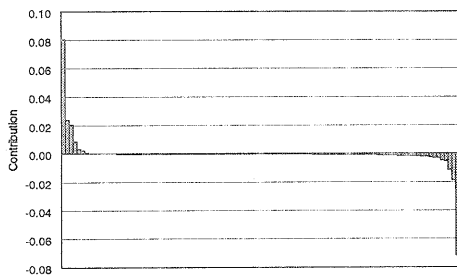


図3 入力文と類似文書の類似度における単語寄与度
Fig. 3 Word contribution to query-document similarity

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は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、単語寄与度の定義から、入力文と検索対象文書に共起している単語である。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。このような単語は単語寄与度の絶対値が小さい単語、すなわち入力文との類似度にはほとんど影響を与えていない単語とは異なり、その単語が出現する文書の特徴を顕著に表している単語と考えられる。

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適合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。この際、適合文書の特徴をより顕著に表す単語を抽出することが有効である。そこで、適合文書に出現する各単語の入力文との類似度に対する単語寄与度を計算し、これに基づいて検索式拡張に使用する単語を選択する手法を提案する。

前述の分析の結果、入力文と適合文書との類似度に対する単語寄与度を計算し、単語寄与度の絶対値が大きい単語を抽出することにより、適合文書に出現する全ての単語の中から入力文との類似度に優位な影響を与えている単語のみを抽出することが可能であることが明らかになった。しかし、単語寄与度の絶対値が大きい単語のうち単語寄与度の値が正の単語は入力文と適合文書に共起しているため、検索式拡張には使用できない。そこで、単語寄与度の値が大きく負の値を持っている単語のうち、適合文書にのみ出現する単語に着目する。このような単語は適合文書の特徴を顕著に表し、かつ入力文を表す検索式には出現しない単語である。したがって、このような単語を検索式拡張の際に使用することにより、有効な検索式拡張が実現できると考えられる。

単語寄与度は、注目単語の存在の有無による文書間の類似度の差を、その単語の影響度を測る基準としている点では、2章で説明した term discrimination value と同様である。しかし、個々の単語の影響を全ての検索対象文書間の類似度の差の平均に基づいて表す term discrimination value とは異なり、単語寄与度は入力文とそれに対する個々の適合文書との類似度の差によって表される。したがって、単語寄与度を基準として使用することにより、個々の適合文書と入力文との類似度を考慮しない term discrimination value に比べ、適合文書集合の中で特徴的な単語を効率的に抽出することができる。

また、2章で説明したように、検索対象文書数を M とすると単語寄与度と同様に個々の単語の影響を表す term discrimination value を計算するためには M 回以上の類似度計算が必要となる。一方、検索対象文書に出現する単語数を m とすると、検索対象文書に出現する全ての単語の寄与度の計算に必要な類似度計算は $(m+1)$ 回である。情報検索においては、多くの場合 $m \ll M$ であるため、単語寄与度を使用することにより、term discrimination value に比べ少ない計算量で各単語の影響度を計算することができる。

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 に対する適合文書集合 $D_{rel}(q) = \{d_1, \dots, d_{Num}\}$ 中の各文書に出現する全ての単語の寄与度を求め、各々の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 $Score(w)$ は式 (7) によって表される。

$$Score(w) = wgt \times \sum_{d \in D_{rel}(q)} Cont(w, q, d) \quad (7)$$

次に、抽出された単語のうち元の入力文に含まれていない単語を入力文を表すベクトルに加えることにより、検索式拡張を実現する。ある単語 w を入力文のベクトルに加える際には、式 (7) で計算されたスコア $Score(w)$ を単語 w が入力文に出現する頻度 (term frequency) とみなし、入力文のベクトル内で単語 w を表す要素の値を計算する。ベクトルの各要素が $TF \cdot IDF$ によって計算されている場合、 $Score(w)$ に単語 w の IDF をかけ、その結果得られた値を入力文のベクトルの単語 w の要素に入れることにより、検索式拡張を行う。

4. 評価実験

提案手法の有効性を示すため、従来手法である Rocchio のアルゴリズムに基づく検索式拡張手法との比較実験を行う。本実験は manual feedback をベースにした検索式拡張手法の評価実験である。また、ここで提案手法との比較対象となる Rocchio の手法は式 (2) で表される手法である。以下、この評価実験について詳しく説明する。

4.1 データ概要

本実験では TREC-6 のデータを使用する⁴⁾。すなわち、入力文は 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。本実験では、Topic の全てのフィールド (Title, Description, Narrative) を入力文として使用する。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する。また、適合フィードバックや検索結果の評価は TREC から提供されている各入力文に対する適合文書データに基づいて行う。表 2 に TREC が公開しているデータ⁴⁾ に基づいた入力文および検索対象文書の種別毎の文書数ならびに平均単語数を示す。

表 2 TREC-6 データ概要
Table 2 TREC-6 data statistics

Doc type	# Docs	Avg # Words/Doc
"Financial Times"	210,158	412.7
"Federal Register"	55,630	644.7
"Foreign Broadcast Information Service"	130,471	543.6
"LA Times"	131,896	526.5
Topics 301-350	50	88.4

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、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 $TF \cdot IDF$ を求めることによって計算する。本実験で使用した TF および IDF の計算式を、式 (8), (9) に示す。

- TF factor

$$\log(1 + tf) \quad (8)$$

- IDF factor

$$\log \left(\frac{M}{df} \right) \quad (9)$$

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。

入力文と検索対象文書との類似度は、式 (10) で定義される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⁾。

$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| |\vec{d}|} \quad (10)$$

但し、 $\vec{q}$, $\vec{d}$ はそれぞれ入力文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

まず、元の入力文を表すベクトルを使用し、検索対象文書集合に対し、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、入力文に対する適合文書集合を抽出し、検索式拡張を行う。4.1節で述べたように、適合文書集合の抽出は各入力文に対する TREC の適合文書データに基づいて行われる。本実験では以下に述べる 2 つの手法（手法 A, B）を用いて適合文書集合を抽出する。手法 A は検索式拡張に十分な適合フィードバックを与える手法であり、手法 B は適合文書集合が抽出される文書群を制限することにより、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 上位 1000 件文書集合に含まれる入力文に対する適合文書のうち、式 (10) によって計算される入力文との類似度が高い文書 Num 個を適合文書集合として抽出する。また、Rocchio の手法で使用される非適合文書集合は、入力文に対する非適合文書のうち入力文との類似度の高い文書 500 個とする。

手法 B 上位 1000 件文書集合のうち、類似度の上位 20 件のみから適合文書集合ならびに非適合文書集合を抽出する。すなわち、上位 20 件の文書に含まれた適合文書を適合文書集合とし、残りの文書を非類似文書集合とする。

(3) 最終検索

検索式拡張の結果得られた検索式をもとに再度検索を行い、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、適合文書集合に含まれる文書の数 Num 、各適合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 は 10 および 20 の 2 通りに設定した。また、Rocchio の手法については、文献⁹⁾に示されているように、各入力文につき、式 (2) で計算された重みの大きい単語 20 個を検索式拡張に使用した。

まず、Rocchio の手法のパラメータ α, β, γ の最適化のため、文献⁹⁾および文献¹¹⁾に示された全てのパラメータ設定について $Num = 10$ および $Num = 20$ のそれぞれの条件下での Rocchio の手法による検索の平均精度 (average precision) を測定した。この結果を表 3 に示す。また、比較のため初期検索 (“Baseline”) の平均精度も示す。

表 3 Rocchio の手法による検索の平均精度 (手法 A)
Table 3 Average precision of Rocchio's method (Exp A, $Num=10$)

Parameters	Average precision	
	$Num=10$	$Num=20$
$\{\alpha, \beta, \gamma\}$		
$\{2, 16, 2\}$	0.2983	0.3319
$\{2, 32, 2\}$	0.2985	0.3217
$\{2, 32, 4\}$	0.2935	0.3247
$\{2, 32, 8\}$	0.2926	0.3214
$\{2, 4, 1\}$	0.3210	0.3517
$\{2, 64, 2\}$	0.2916	0.3223
$\{3, 2, 2\}$	0.3410	0.3885
Baseline	0.1433	

表 3 より明らかなように、 $Num = 10$ および $Num = 20$ の両条件下ともに $\alpha=3, \beta=2, \gamma=2$ と設定したときに最大の平均精度が得られた。したがって、これ以降の提案手法との比較ではこのパラメータ設定での結果を使用する。

次に、 $Num = 10$ および $Num = 20$ の場合の提案手法の平均精度を表 4 に示す。また、比較のため初期検索の平均精度も示す。表中、 $Num = 10$ の場合の提案手法を “WC10”、 $Num = 20$ の場合の提案手法を “WC20” と表す。

表 4 手法 A での各検索式拡張手法の平均精度
Table 4 Average precision of each query expansion method in experiment A

wgt	WC10	WC20
-100	0.3696	0.4265
-400	0.3837	0.4475
-1200	0.3844	0.4528
-2000	0.3837	0.4530
-3000	0.3854	0.4554
-5000	0.3865	0.4565
-10000	0.3850	0.4559
-20000	0.3815	0.4539
Baseline	0.1433	

表4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、表3に示された Rocchio の手法の結果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 ($wgt = -5000$)、Rocchio の手法 ($\alpha=3, \beta=2, \gamma=2$)、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4には $Num = 10$ 、図5には $Num = 20$ の場合の曲線を示す。

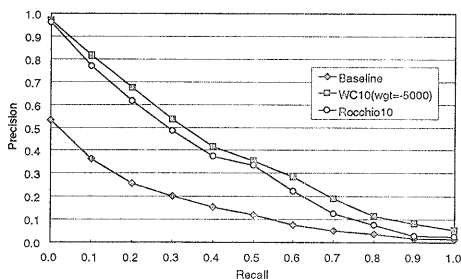


図4 各手法の Precision と Recall (手法 A, $Num=10$)

Fig. 4 Recall-Precision curve (Exp A, $Num=10$)

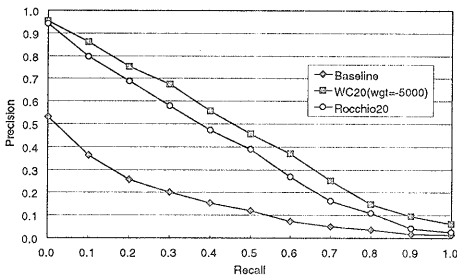


図5 各手法の Precision と Recall (手法 A, $Num=20$)

Fig. 5 Recall-Precision curve (Exp A, $Num=20$)

これらの Recall-Precision 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。

まず、Rocchio の手法のパラメータ最適化のため、

手法 A での評価実験と同じ α, β, γ のパラメータの組み合わせに基づく検索の平均精度を測定した。この結果を表5に示す。

表5 Rocchio の手法による検索の平均精度 (手法 B, $Num=20$)

Table 5 Average precision of Rocchio's method (Exp B, $Num=20$)

$\{\alpha, \beta, \gamma\}$	Avg prec
$\{2, 16, 2\}$	0.2248
$\{2, 32, 2\}$	0.2229
$\{2, 32, 4\}$	0.2220
$\{2, 32, 8\}$	0.2188
$\{2, 4, 1\}$	0.2329
$\{2, 64, 2\}$	0.2023
$\{3, 2, 2\}$	0.2330

表5より、手法 B でも手法 A 同様 $\alpha=3, \beta=2, \gamma=2$ と設定したときに平均精度が最大になることが明らかになった。

次に、提案手法による検索の平均精度を表6に示す。

表6 手法 B での提案手法の平均精度

Table 6 Average precision of word contribution based query expansion method in experiment B

wgt	WC20
-10	0.2127
-25	0.2488
-50	0.2584
-100	0.2540
-400	0.2407
-1200	0.2335
-2000	0.2320
-3000	0.2307
Baseline	0.1433

これより、手法 A 同様、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48.4%~80.3%の検索精度向上が得られたことが明らかになった。また、手法 A ほど明確な差は確認されなかったものの、 wgt の設定により Rocchio の手法を上回る検索精度が得られることがわかった。

次に、提案手法 ($wgt = -50$)、Rocchio の手法 ($\alpha=3, \beta=2, \gamma=2$)、ならびに初期検索の Recall-Precision 曲線を図6に示す。

この Recall-Precision 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。そのうえ、全体的には提案手法の方が精度が高いものの、 $Recall=0.0$ の時点では提案手法の精度が0.7760なのに対し、Rocchio の手法は0.8268と、Rocchio の手法の方が精度が高い。

ここで、 $wgt = -1200$ とした場合の提案手法と

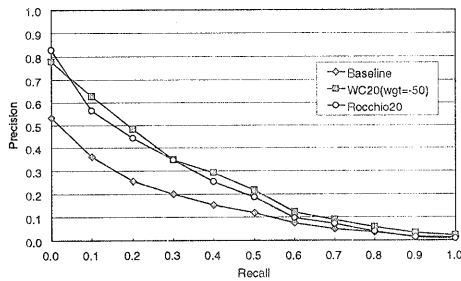


図6 各手法の Precision と Recall (手法 B, Num=20, wgt=-50)

Fig. 6 Recall-Precision curve (Exp B, Num=20, wgt=-50)

Rocchio の手法の Recall-Precision 曲線を図7に示す。

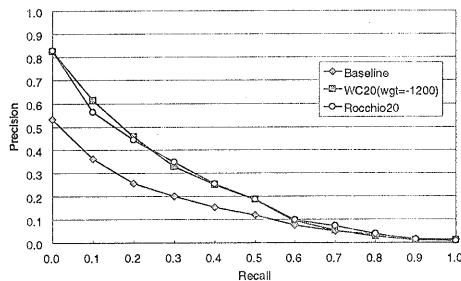


図7 各手法の Precision と Recall (手法 B, Num=20, wgt=-1200)

Fig. 7 Recall-Precision curve (Exp B, Num=20, wgt=-1200)

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 0.825 と, Rocchio と同程度にまで向上させることができる。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。

しかし, 手法 B での実験の結果, 手法 A で確認されたほどの性能の改善は見出されなかった。この大きな理由の一つとして, 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適合文書集合に含まれる文書数は平均 4.9 個であり, 手法 A と比べ非常に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7に, 初期検索の結果, 上位 20 件までに含まれていた適合文書数の分布に対する入力文の数を示す。

表7から明らかなように, 50 個の入力文中, 32 個の文書に対しては 5 個以下の適合文書から検索式拡張が行われており, そのうち 9 個の入力文は上位 20 件の中に適合文書が含まれていないため, 提案手法では

表7 上位 20 位までの適合文書数に対する入力文の分布
Table 7 Distribution of queries to number of relevant documents in top 20 documents

適合文書数	入力文数
16 以上	2
11-15	6
6-10	10
1-5	23
0	9
合計	50

検索式拡張を行うことが出来なかった。

ここで, 上位 20 件文書に適合文書が含まれなかった入力文を除いた 41 件の入力文に対し, 手法 B での提案手法 ($wgt=-50$) および Rocchio の手法 ($\alpha=3, \beta=2, \gamma=2$) で検索式拡張を行った際の各手法の平均精度を比較した。その結果を表8に示す。

表8 手法 B での各検索式拡張手法の平均精度比較
Table 8 Comparison of average precision of each query expansion method in experiment B

精度比較	入力文数	割合
WC < Roc	14	34.1%
WC > Roc	27	65.9%
Total	41	100.0%

この表より, 検索式拡張が行われた入力文の 65.9% では提案手法が Rocchio の手法を上回る精度を得ていることがわかる。この結果は, 検索式拡張の元となる情報が少ない場合でも, 提案手法による検索式拡張が有効であることを示している。

5.3 Rocchio の手法との厳密な比較

本実験で比較を行った提案手法と Rocchio の手法の条件に以下の相違点があげられる。

- (1) 非適合文書情報使用の有無
- (2) 検索式拡張時に使用される単語数

前述のように, 本実験での Rocchio の手法の各パラメータは文献⁹⁾などに述べられている条件を元に設定しているが, 以上の相違点が提案手法との検索精度比較に影響を及ぼしている可能性がある。そこで, ここでは Rocchio の手法の条件を修正し, 提案手法との公正な比較実験を行った。以下, この実験結果について報告する。

5.3.1 非適合文書情報使用の有無について

提案手法では Rocchio の手法と異なり, 検索式拡張の際に非適合文書の情報を利用してない。そこで, Rocchio の手法との比較条件を統一するため, Rocchio の手法のパラメータを $\alpha=3, \beta=2, \gamma=0$ と設定し, Rocchio の手法でも適合文書のみを使用した実験

を行った. 表 9 に各評価実験の結果を示す.

表 9 適合文書のみを使用した Rocchio の手法の平均精度
Table 9 Average precision of Rocchio's method without nonrelevant information

Exp condition	Avg prec
手法 A ($Num=10$)	0.3441
手法 A ($Num=20$)	0.3789
手法 B ($Num=20$)	0.2310

表 9 の結果から, Rocchio の手法で非適合文書を使用しないことにより, 手法 A ($Num=10$) においては非適合文書を使用した場合の平均精度を上回る結果が得られた. しかし, いずれの場合においても提案手法に基づく検索式拡張の平均精度の方が高い. この結果は, 提案手法の有効性が非適合文書情報使用の有無によらないことを示している.

5.3.2 検索式拡張時に使用される単語数

4章の評価実験では, Rocchio の手法において加えられる単語数を 1 つの入力文につき 20 単語に固定したのに対し, 提案手法では適合文書集合中の各文書からそれぞれ $N=10$ 単語を抽出している. そのため, 提案手法の方が入力文に加えられた単語数が多くなり, その結果高い検索精度が得られている可能性がある. 表 10 に, 各評価実験において検索式拡張時に各入力文に加えられた平均単語数を示す.

表 10 単語寄与度に基づく検索式拡張で加えられた平均単語数
Table 10 Number of words used in word contribution based QE

QE method	Avg # of words
WC10 (手法 A)	34.2
WC20 (手法 A)	55.0
WC20 (手法 B)	20.9
Rocchio	20

表 10 から明らかなように, 適合文書集合に含まれる文書数が多い手法 A の実験では, 提案手法の方が Rocchio の手法と比べ検索式拡張時に使用された単語数が多い.

そこで提案手法との比較を公平にするため, Rocchio の手法において各入力文毎に使用される単語数を変更し, 手法 A の評価実験を行った. 表 10 の結果に基づき, $Num=10$ の場合は拡張単語数を 35, $Num=20$ の場合は拡張単語数を 55 にそれぞれ変更した. また, Rocchio の手法のパラメータは $\alpha=3$, $\beta=2$, $\gamma=2$ と設定した. この結果を表 11 に示す.

表 11 から明らかなように, $Num=10, 20$ のいずれ

表 11 拡張単語数を変化させた Rocchio の手法による検索の平均精度

Table 11 Average precision of Rocchio's method with increased number of expanded words

# of expanded words	Num	Avg prec
35	10	0.3664
55	20	0.4385

の条件下でも拡張単語数を増やすことにより検索精度が向上したものの, いずれも提案手法での検索精度を超えていないことが明らかになった. すなわち, 提案手法の優位性が拡張単語数の差によるものではないことが確認された.

6. 考 察

ここでは wgt の最適化について考察を行う.

5.2節にて, wgt を -50 から -1200 に変化させた結果, Recall=0.0 の時点での精度が上昇するものの, 平均精度が下がるという現象がみられた. 詳しい分析のため, 図 8 に, 手法 B の実験での wgt の変化に伴う平均精度 (Avg Precision) ならびに Recall=0.0 の時点での精度 (Precision@Recall=0.0) の変化を示す.

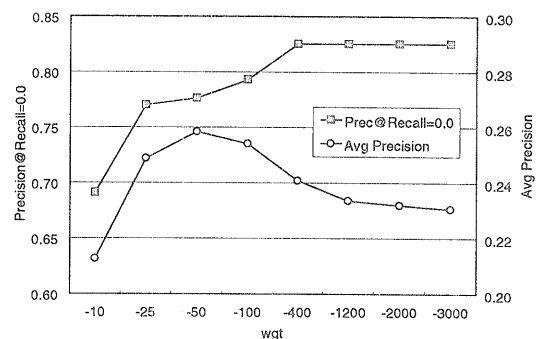


図 8 手法 B での wgt の変化と精度
Fig. 8 Precision and wgt in Exp B

図 8 より, wgt の絶対値を増大させるにつれ, Recall=0.0 での精度も増大していることが明らかである. しかし, 平均精度は $wgt=-50$ をピークに減少している. このことから, 検索式拡張で使用した単語の影響を増大させるにつれ, 類似度が上昇する適合文書と, ある wgt のピークを境に類似度が下降する適合文書が存在することがわかる.

3.2節で述べられた分析結果より, 単語寄与度が大きく負の値を持つ単語は, 各検索対象文書につき数単語程度しか存在していないことがわかっている. 一方, 提案手法による検索式拡張では各検索文書から $N(=10)$ 個の単語を抽出しているため, ここで抽出された N 個

の単語のうち、単語寄与度の絶対値が高い単語と低い単語の単語寄与度には大きな差がある。これは、表1に示された単語寄与度のデータ例からも明らかである。

wgt の絶対値が増加するにつれ、拡張された検索式で使用された単語のうち単語寄与度の絶対値が高い単語と低い単語の影響度、すなわち式(7)で計算されたスコアの差が拡大するのは明らかである。したがって、 wgt の絶対値が増大すると、スコアの低い単語の影響度が大きくなるため、スコアの低い単語を含む適合文書の類似度が増加する反面、このような単語を含まない適合文書の類似度が相対的に低くなる。このことから、 wgt の最適値を導出するためには、スコアの低い単語が存在する適合文書と存在しない適合文書とのバランスを考慮する必要がある。

7. 結 論

検索式拡張の代表的な手法である Rocchio の手法では、適合文書集合における平均的な重要度のみを基準として検索式拡張の際に使用される単語の抽出を行っているため、検索式拡張において適切な単語が抽出されていない可能性がある。これに対し、本論文では入力文と個々の適合文書との間の類似度に対する各単語の影響度を考慮して検索式拡張を行うことにより、一層の検索精度向上が期待できると考えた。

そこで本論文では、適合文書に出現する各単語の単語寄与度を算出することによって適合文書の特徴を顕著に表す単語が抽出できることを示し、これを応用した検索式拡張手法を提案した。さらに、提案手法の有効性を実証するため、Rocchio の手法との比較実験を行った。評価実験では、適合フィードバックが十分に与えられる状況と、より現実的な量のフィードバックが得られる状況をシミュレートし、それぞれの状況下で従来手法との比較実験を行った。その結果、いずれの状況下においても提案手法が有効であることを証明した。

謝辞 日頃御指導頂く KDD 研究所村谷所長、鈴木副所長、および知識情報処理グループの皆様へ深謝致します。また、本論文で述べた評価実験において多大な御協力を頂いた早稲田大学の大平茂輝氏ならびに HTW Dresden の Marko Herzog 氏に心から感謝申し上げます。

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 Con-

ference", 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, NIST SP 500-242, pp 239-251, 1999.
- 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.

(平成 11 年 6 月 20 日受付)

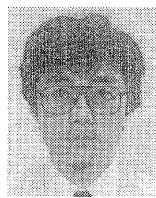
(平成 11 年 9 月 27 日採録)

(担当編集委員 仲尾 由雄)



帆足啓一郎（正会員）

平7 早大・理工・情報卒。平9 同大学院修士課程了。早稲田大学ヒューマノイドプロジェクトに携わり、音声対話、マンマシンインタフェースの研究に従事。同年国際電信電話(株)入社。現在、(株)KDD 研究所において情報検索、情報フィルタリング等の研究に従事。



井ノ上直己（正会員）

昭57 京大・工・電子卒。昭59 同大学院修士課程了。同年国際電信電話(株)入社。昭62～平3 ATR 自動翻訳電話研究所に出向。知識ベース、自然言語処理の研究に従事。平3 より、KDD 研究所において機械翻訳、音声認識、情報検索の研究に従事。工博。平3 年度学術奨励賞受賞。平7 年度日本音響学会技術開発賞受賞。電子情報通信学会、日本音響学会各会員。



松本 一則（正会員）

1984 京大・工・情報卒。1986 同大学院修士過程了。同年国際電信電話(株)入社、研究所所属。現在、同研究所知識情報処理グループにて、時系列データ処理、類似検索の研究開発に従事。特に実例からの知識獲得手法に興味を持つ。電子情報通信学会会員。



橋本 和夫（正会員）

1977 東北大・工・電子卒。1979 年同大学院修士課程了。同年国際電信電話(株)入社、研究所所属。現在、同研究所知識情報処理グループ(リーダ)。自然言語処理、知識表現、エキスパートシステム等の研究開発に従事。電子情報通信学会、人工知能学会各会員。