1U－8

単語寄与度に基づく検索式拡張手法の検討

帆足 啓一郎　松本 一則　井ノ上 直己　橋本 和夫

KDD 研究所

1　はじめに

有効な検索結果を得るためには効果的な検索式(query)の作成が重要であることは明らかである．近年，この検索式(query)を自動的に拡大する「検索式拡張(query expansion)」の技術を採用したシステムがTREC等の会議で多く発表されている[1]．本研究では筆者らが定義した単語寄与度[2]という概念に基づいた新たな検索式拡張手法を提案し，従来手法との比較実験により提案手法の有効性を検証する．

2　従来手法

現在，最も有効な検索式拡張手法の一つとして，Rocchioの語の重み付け手法に基づくものがあげられる[3]．Rocchioの手法はベクトル空間モデルを前提とした語の重み付け手法であり，検索式のベクトルを類似文書のベクトルに近づけつつ非類似文書から遠ざけることを目的とする．この手法を表す数式は以下の通り：

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

ただし，R，Nはそれぞれ類似文書ならびに非類似文書の数を表し，α，β，γは任意の係数である．本手法を用いた検索式拡張は，上記ベクトル変換の結果，元の検索式に含まれない語のうち重みの値が高い語を抽出し，その重みとともに元の検索式のベクトルに加えるという手法で実現される．TREC-7では$\alpha = 3, \beta = 2, \gamma = 2$の係数値で上記検索式拡張手法を採用した検索システム"SMART"が発表され，高い検索精度を得ていた[4]．

Rocchioの手法の効果は多くの論文などで確認されているが，この手法では類似文書に出現する個々の単語の持つ影響について考慮していない．このため，検索式拡張の際に必ずしも有効な単語が抽出されていない可能性がある．

Word Contribution Based Query Expansion.
Keiichiro HOASHI (hoashi@kddlabs.co.jp), Kazunori Matsumoto, Naomi Inoue and Kazuo Hashimoto.
KDD R&D Laboratories, 2-1-15 Ohara, Kamifukuoka-shi, Saitama 356-8502 JAPAN.

3　単語寄与度による検索式拡張手法

3.1　類似文書との類似度における単語寄与度

図1はある検索式とそれに類似している文書との類似度における全出現単語の寄与度を大きい順に左から並べたものである．

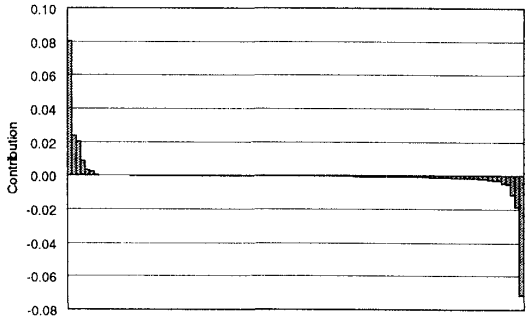

図1: 入力文と類似文書の類似度における単語寄与度

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と考えられる．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と思われる．そこでこのような単語を入力文に加えることにより，従来以上に有効な単語と重みの抽出が可能であると考える．

3.2　提案手法

ここで，単語寄与度に基づいた検索式拡張手法を以下のように提案する．

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．

4 評価実験

手法 提案手法の有効性を示すため，Rocchio の手法との比較実験を行った．本実験では検索用データとして TREC-6 の入力文 50 件と検索対象文書 528,150 件を使用し，類似度は TF*IDF ベクトル間の cos 値によって算出する．

まず，各入力文毎に初期検索を行い，類似度の上位 1000 件の文書を抽出する．各入力文 qについて抽出された文書のうち，実際に類似している文書の上位 Num 個を入力文に対する類似文書集合 $D_{rel}(q)$ とし，これらの文書から前節で述べた手法により検索式拡張を行う．今回の実験では各文書から抽出される単語数 Nを 10 に固定した．最終的な検索結果はここで拡張された検索式に基づいて導出される．

Rocchio の手法においては，類似文書集合として上記手法と同じ $D_{rel}(q)$ を，非類似文書集合として初期検索で類似度順位が 500 位以下の文書を利用する．ただし，この非類似文書集合に実際は類似している文書が含まれる場合，その文書は非類似集合から除外される．また，TREC-7 の SMART と同様に各係数の値は$\alpha = 3, \beta = 2, \gamma = 2$ とし，重みの上位 20 単語を元の検索式に加えることとする．

4.1 結果

表 1に Num=10,20 および wgt = −100, −400, −800, −1200 の提案手法による検索式拡張，および Num=10,20 で Rocchio の手法による検索式拡張，ならびに初期検索 (Baseline) の平均 Precision の値を示す．

この結果より，提案手法は初期検索に比べて 157%～197%以上の検索精度向上が得られたことがわかった．また，wgt の絶対値の増大，すなわち新しく拡張された単語の影響力の増大にともない精度が向上しており，さらに 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重み付けが有効であることが示された．図 2に Num=20 の場合の各検索手法の Precision-Recall 曲線を示す．この図より，提案手法はどの Recall においても Rocchio の手法の Precision 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証明された．

表 1: 各検索式拡張手法の平均 Precision

wgt	-100	-400	-800	-1200
Num=10	0.3696	0.3837	0.3837	0.3844
Num=20	0.4265	0.4475	0.4504	0.4528
Rocchio10	0.3140			
Rocchio20	0.3758			
Baseline	0.1433			

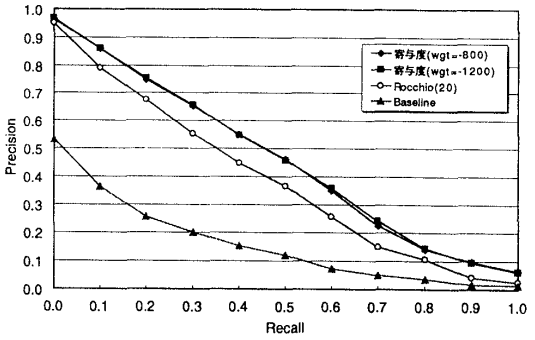

図 2: 各手法の Precision と Recall($Num = 20$)

5 結論

本研究では，単語寄与度に基づいて類似文書から単語を抽出して検索式に加える検索式拡張手法を提案した．評価実験で抽出された単語の影響力の増加にともない検索精度が向上したことから，提案手法による検索式拡張のための単語およびその重みの抽出が有効であることが証明された．また，従来手法との比較の結果からも，提案手法の有効性が示された．

参考文献

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

[2] 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.

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

[4] 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)