3V－3

類似検索における単語寄与度に基づく重要語選択手法の検討

帆足 啓一郎　青木 圭子　松本 一則　橋本 和夫

KDD 研究所

1　はじめに

現在までに，大量のドキュメント中から類似する情報を検索するための様々な類似性の尺度が考案され，これらに基づく検索手法が提案されている．しかし，どの手法を用いても不要なドキュメントが検索される割合は依然として高く，検索精度向上のための工夫が必要である．このような工夫の一つとして，類似検索の際に重要だと思われる語を選択して検索を行う手法がある．ここではこの手法を「重要語選択」と呼ぶ．

ドキュメント同士の類似度を測定する場合，一般的には各ドキュメントに出現する単語の頻度などを抽出し，これを特徴とする．しかし，ここで出現する全ての単語がドキュメント間の類似度に影響を与えるとは考えにくく，逆に類似度に無関係な単語までもドキュメントの特徴として考慮することにより，検索精度が低下してしまうおそれがある．そこで，検索入力ならびに検索対象のドキュメント集合に出現する単語の中から類似検索に必要な単語を抽出する手法が重要語選択である．

本稿では筆者らが提案したドキュメント間の類似度における単語の寄与度を用いる重要語選択手法を提案し，χ二乗による手法との比較実験を通してその有効性を検証する．

2　従来の重要語選択手法

重要語選択の手法として，各単語のχ二乗値に基づいて重要語を選択する手法[1][2]が提案されている．検索対象ドキュメント集合D_tを，入力ドキュメントiに対する正解ドキュメント集合CとCに属しない集合$\overline{C}$の2つの集合に分け，各単語wのχ二乗値を以下の数式により算出する．

$$\frac{(N_{(C+)}+N_{(C-)}+N_{(\overline{C}+)}+N_{(\overline{C}-)})(N_{(C+)}N_{(\overline{C}-)}+N_{(C-)}N_{(\overline{C}+)})}{(N_{(C+)}+N_{(C-)})(N_{(\overline{C}+)}+N_{(\overline{C}-)})(N_{(C+)}+N_{(\overline{C}+)})(N_{(C-)}+N_{(\overline{C}-)})}$$

ただし，$N_{(C+)}$はC中のドキュメントで単語wを含むドキュメントの数，$N_{(C-)}$はC中のドキュメントで単語wを含まないドキュメントの数とする．このように計算されたχ二乗値の上位M個の単語を重要語として選択する．

しかし，ここで求められるχ二乗値はCと$\overline{C}$に含まれるドキュメント数の比率に極端な差がある場合，有効ではないと考えられている．そのため，類似検索のように大量の検索対象データからわずかな類似データを検索するタスクにおいては不向きであると思われる．

2.1　単語寄与度に基づく手法

前述の問題に対応するため，筆者らが提案したドキュメント間の類似度における単語の寄与度[3]を利用した重要語選択手法を提案する．

2つのドキュメントd_i，d_j間の類似度における単語wの寄与度を求めるものとする．まず，Iwayamaらの手法[4]により，d_i，d_j間の類似度$Sim(d_i,d_j)$を計算する．

ここで$d_i'(w)$をd_iからwを除いたものとし，d_i，d_j間の類似度におけるwの寄与度$Cont(d_i,d_j,w)$を以下のように定義する．

$$Cont(d_i,d_j,w)=Sim(d_i,d_j)-Sim(d_i'(w),d_j'(w))$$

このようにしてd_i，d_j中の全ての出現単語についてその寄与度を求めることができる．

重要語選択のためには，入力文書iに対して類似しているドキュメントの集合(以降，正解ドキュメント集合)$C=\{c_1,\cdots,c_N\}$中の各ドキュメントとiとの類似度における全ての単語wの寄与度$Cont(i,c_k,w)$を算出する．その後，全ての単語wに対し，寄与度の総和：

$$SumCont(i,w)=\sum_{k=1}^{N}Cont(i,c_k,w)$$

A Research on a Feature Extraction Method Based on Word Contribution.
Keiichiro Hoashi(hoashi@lab.kdd.co.jp), Keiko Aoki, Kazunori Matsumoto and Kazuo Hashimoto.
KDD R&D Laboratories, 2-1-15 Ohara, Kamifukuoka-shi, Saitama 356-8502 JAPAN.

を求め，この値の上位 M 個の単語を入力 i に対する重要語とする．

3 評価実験

単語寄与度に基づく重要語選択手法の有効性を示すため，χ 二乗による手法との比較実験を行った．

3.1 実験データ

本実験で使用したデータは特許庁から発行されている特許データのうち，1993 年から 1998 年までに発行された公開公報データである．

入力ドキュメントとして同期間に KDD から出願された特許 20 件 $D_{in} = \{i_1, i_2, \cdots, i_{20}\}$ を設定し，特許検索業者にこれらの特許に類似している公開公報を検索させた．1 件の入力ドキュメントに対する正解ドキュメント集合に含まれるドキュメント数，すなわち 1 つの入力特許に対し，類似していると判断された公開公報の数の平均は 10.75 個である．以降，入力ドキュメント i_n に対する正解ドキュメント集合を C_n とする．

上記の公開公報データのうちの 10000 件を検索対象ドキュメント集合 (以降，D_t) とした．D_t には C に含まれる全ての公開公報データが含まれている．

3.2 実験内容

本実験では，χ 二乗と単語寄与度に基づいて選択された M 個の単語をもとに，Iwayama らの手法により D_t からの類似検索を行った．また，比較として全ての単語に基づく類似検索も行った．

3.3 結果

χ 二乗ならびに寄与度について，重要語の個数 $M = 100, 200, 300, 400, 500$ としたときの検索，および全ての単語に基づく検索 (All) の平均 Precision を表 1に示す．

表 1: 重要語の個数毎の平均 Precision

M	χ^2	$Cont$
100	0.1642	0.4400
200	0.1144	0.4983
300	0.1383	0.5277
400	0.1619	0.5394
500	0.1757	0.4856
All	0.3668	

また，単語寄与度 ($M = 400$)，χ 二乗 ($M = 500$) で選択された重要語による検索と全単語検索の Recall/Precision を測定した．その結果を図 1 に示す．

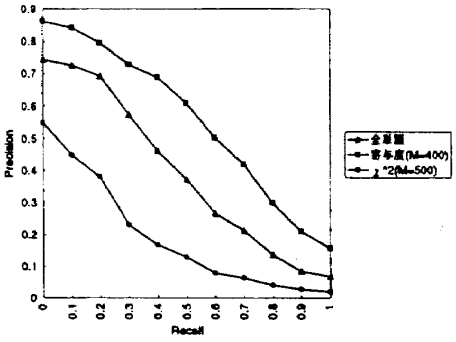

図 1: 各検索手法の Recall/Precision

以上の結果より，χ 二乗の重要語選択手法では検索精度が下がったのに対し，単語寄与度による重要語選択では精度向上が見られ，提案手法の有効性が示された．

4 まとめ

本稿ではドキュメント間の類似度における単語の寄与度に基づいた重要語選択手法を提案し，χ 二乗に基づく選択手法との比較実験を通してその有効性を確認した．

しかし，寄与度に基づく重要度選択による検索で高い検索精度を得ることができたものの，閉じられたデータに対する実験ゆえの結果という可能性もある．さらに厳密な評価のためには，本手法によって選択された重要語を用い，他の検索対象に対して検索するなどの実験を行わなければならない．

また，本研究では入力ドキュメントに対する正解ドキュメントの情報を用いて重要語選択を行ったが，正解が不明な新規入力ドキュメントに対する検索時の重要語選択手法についても検討する必要がある．

参考文献

[1] Schutze, Hull, Pedersen: "A Comparison of Classifiers and Document Representations for the Routing Problem", Proc of ACM SIGIR'95, 1995.

[2] Ng, Goh, Low: "Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization", Proc of ACM SIGIR'97, 1997.

[3] 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.

[4] Iwayama, Tokunaga: "A Probablistic Model for Text Categorization: Based on a Single Random Variable with Multiple Values", Proc of 4th Conference on Applied Natural Language Processing, pp 162-167, 1994.