テキストの絞り込み検索のための特徴抽出手法の検討

5Y－2

帆足 啓一郎　松本 一則　青木 圭子　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，さらに絞り込み検索を行うことが必要である．そこで本研究ではテキストの絞り込み検索に適した特徴量としてドキュメント間の類似性における単語寄与度を定義し，広く一般に使用されている特徴量である TF*IDF との比較実験を行い，単語寄与度の有効性を検証する．

2　従来の特徴量とその問題点

これまでドキュメントを表す特徴量としてはドキュメント中の単語(句)の頻度を要素とした特徴ベクトル(*term frequency*)，あるいはこの特徴ベクトルの各要素に重みを付加した TF*IDF (*Term Frequency * Inverse Document Frequency*) などといった特徴量が提案されており，一般に広く用いられている [1].

また，Iwayama らはこのような単語頻度の特徴量に基づき，SVMV (Single Random Variable with Multiple Values) という統計的テキスト分類手法を提案し，他のテキスト分類手法との比較実験の結果，その優位性を示した [2]. しかし，実際にこの手法を用いて検索を行ったところ，抽出されたドキュメント群中に不要ドキュメントが含まれる割合は依然として高いものであった．このことから，抽出ドキュメント群に対し，さらに絞り込み検索を行う必要があり，そのために有効な特徴量について検証する必要があると考える．

しかし，こうした特徴量はヒューリスティックなものであるため，その有効性を示すのは難しく，現にこれまでこのような特徴量の効果を検証した例は少ない．そこで本研究ではドキュメント間の類似度における単語の寄与度という概念を定義し，従来の特徴量との比較実験を通じてその有効性を検証する．

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

3　単語寄与度

ここでは 2 つのドキュメント間の類似度を計算する際，各ドキュメント中に出現した単語がその類似度に与える影響をその単語の寄与度とし，これを特徴量とする．以下，この寄与度の計算方法について解説する．

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，d_i, d_j 中の出現単語とその出現頻度を求める．ここでは単語として名詞のみを考える．これをもとに Iwayama らの手法で 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．本研究ではこの寄与度を d_i と d_j との比較を行う際の特徴量と捉える．

4　比較実験

単語の寄与度という特徴量の有効性を示すため，一般に用いられている特徴量である TF*IDF との実験を行った．

4.1　方法

ドキュメント d_{org} と 25,511 個のドキュメントからなる検索対象ドキュメント群 D との間の類似度を計算した．その結果，D 中のドキュメントのうち d_{org} との類似度が高いドキュメント 55 個 (以下，D_{top}) を主観評価により，d_{org} と類似性が高い $D_\circ$，類似性の無い $D_\times$，および $D_\circ$, $D_\times$ のいずれにも属さない $D_\triangle$ の 3 つのグ

表 1: 各グループの構成

Group	No. of docs	Ratio
$D_\circ$	8	14.5%
$D_\triangle$	16	29.1%
$D_\times$	31	56.4%

ループに分割した．各グループ中のドキュメント数を表 1に示す．

D_{top}の各ドキュメントと d_{org}間の類似度において寄与度の高い単語の上位 5 個を各ドキュメント毎に抽出し，これらをマージした単語リスト (48 単語) を生成した．各ドキュメントの特徴はこのリスト中の各単語の寄与度を要素としたベクトルで表す．

この特徴ベクトルの有効性を示すため，これを元に因子分析を行い，第 1 因子と第 2 因子を軸とした 2 次元平面上にプロットした．比較のため，D_{top}のドキュメントを TF*IDF によって特徴抽出し，同様に因子分析してプロットを行った．

4.2 結果

図 1，図 2にそれぞれ寄与度および TF*IDF による特徴抽出に基づいた因子分析の結果を示す．図の中の「○」「△」「×」はそれぞれ $D_\circ$，$D_\triangle$，$D_\times$のドキュメントを表している．

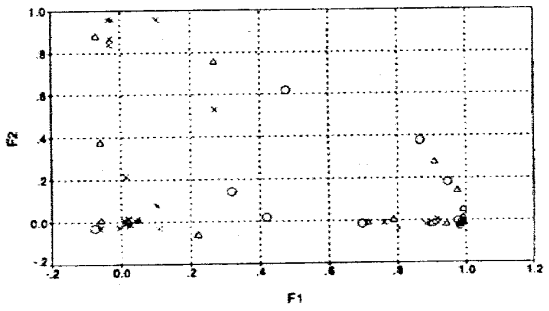

図 1: 寄与度による特徴の因子分析結果

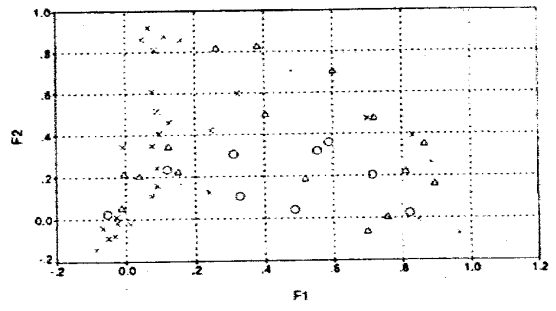

図 2: TF*IDF による特徴の因子分析結果

これらの結果を比較すると，図 1では同じグループのドキュメントが集中してプロットされているのに対し，図 2ではそれに比べて平面上にプロットが散らばっていることが明らかである．このことにより，寄与度の方が TF*IDF よりも主観評価によるドキュメント分類に適していると考えられる．

この結果を定量的に検証する．まず図 1および図 2の平面の両軸をそれぞれ N等分し，各平面を N^2個の矩形領域に分割した．そして，以下の式に基づき，寄与度および TF*IDF による分析結果の情報量を計算した．$p_{ij}(g)$ は領域 (i,j) 内に $g \in \mathcal{G} = \{\circ, \triangle, \times\}$ のプロットが現れる確率とする．

$$H(N) = \frac{1}{N^2} \sum_{i}^{N} \sum_{j}^{N} \left(-\sum_{g \in \mathcal{G}} p_{ij}(g) \log p_{ij}(g) \right)$$

$N = 25, 50, 100$ のときの情報量の計算結果を表 2に示す．

表 2: 寄与度，TF*IDF に基づく分析結果の情報量

N	$H(N)$(寄与度)	$H(N)$(TF*IDF)
25	0.753	0.755
50	0.585	0.660
100	0.337	0.387

この結果より，寄与度による分析結果の方が一様に情報量が少ないことが明らかであり，定量的にも本手法の有効性が示された．

5 まとめ

本研究ではテキストの絞り込み検索に適した特徴量として，ドキュメント間の類似度における単語の寄与度を定義した．この寄与度に基づく特徴量と TF*IDF の特徴量に基づく因子分析結果の情報量の比較により，寄与度の有効性を示すことができた．今後は本研究で提案した特徴量をもとに検索対象ドキュメントから類似文書を検出する手法などについて研究を進める予定である．

参考文献

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

[2] 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.