5W-5

単語の頻度情報の偏りを用いた文書の自動分類手法の評価

村上 誠　　石野 明　　竹田 正幸　　松尾 文碩

九州大学大学院 システム情報科学府 情報理学専攻

1 まえがき

本稿では，英文科学技術文を対象とした文書の自動分類手法の評価を行う．単語には特定のカテゴリに偏って頻出するものと，カテゴリに依らず頻出するものとがある．そこで前者のようなカテゴリを特徴づける単語には高いスコアを付ける方法を示し，そのスコアを基に文書を分類する．

本稿では英文科学技術二次文献INSPECデータの1990年から1999年までの10年分を用いて分類を行なった．すべての文書は複数の分野コードを添付することで分類されている．

まず，ストップワードを除いた総頻度上位10,000語の単語についてカテゴリごとのスコア付けを行なう．単語のスコアは，カテゴリごとの出現確率と全体での出現確率の比で求める．次にそのスコアを基に，文書に対してカテゴリごとのスコア付けを行なう．文書のスコアは，その文書に含まれる単語のスコアの和として求める．最後に文書のスコアを基に分類を行ない評価する．比較のためにSVMによる分類を行なった．

2 文書の分類

本章では文書の自動分類手法について述べる．まずカテゴリごとに異なる単語の頻度情報を用いて単語のスコア付けを行なう．次にその単語のスコアを用いて文書のスコア付けを行ない，それに基づいて分類を行なう．

An Evaluation for Automatic Text-Classification Based on Differences in Word Frequencies over the Categories
Makoto MURAKAMI, Akira ISHINO, Masayuki TAKEDA, Fumihiro MATSUO
Department of Informatics, Kyushu University, 1-10-6 Hakozaki, Higashi-ku, Fukuoka-city 812-8581, Japan

2.1 単語のスコア付け

カテゴリ間で異なる単語の頻度情報を基に，単語に各カテゴリごとのスコア付けをする．

単語 x が全文書中で出現する確率を $P(x)$，単語 x のカテゴリ Ci に出現する確率を $P(x|C_i)$ とするとき，**単語 x のカテゴリ C_i におけるスコア** $S(C_i,x)$ を以下のように定義する．

$$S(C_i,x)=\frac{P(x|C_i)}{P(x)} \tag{1}$$

ここで $f(x)$ を単語 x の出現回数，$f(x,C_i)$ を単語 x のカテゴリ C_i での出現回数，$n(C_i)$ をカテゴリ C_i の総単語数，N を全文書中の総単語数とするとき，$P(x|C_i)$，$P(x)$ をそれぞれ以下の式で求める．

$$P(x|C_i)=\frac{f(x,C_i)}{n(C_i)} \qquad P(x)=\frac{f(x)}{N} \tag{2}$$

このスコア $S(C_i,x)$ が高いということはカテゴリを C_i と特定することで単語 x が出現しやすくなるということであり，そのカテゴリ C_i を特徴づける単語の1つであるといえる．以下，カテゴリを特徴づける単語を**特徴語**と呼ぶ．このスコア付けをストップワードを除いた総頻度上位10,000語について行なった．また，カテゴリの大きさの差を考慮して，スコアの分布を平均0，標準偏差1の正規分布に変換した．

2.2 文書の分類法

前節で求めた単語のカテゴリごとのスコアを用いて，以下のように定義する文書のスコア付けを行なうことで分類する．ある文書 T の構成要素である単語 $x_k(k=1,2,\cdots,n)$ のカテゴリ C_i におけるスコアを $S(x_k,C_i)$ とするとき，**文書 T の C_i におけるスコア** $S(T,C_i)$ を以下の式で定める．

$$S(T,C_i)=\sum_{k=1}^{n}S(x_k,C_i) \tag{3}$$

このスコアを用いて文書の属すカテゴリを判定する．

3 実験と考察

本手法と Support Vector Machines(SVM) による分類を行ない、評価と考察を行なう.

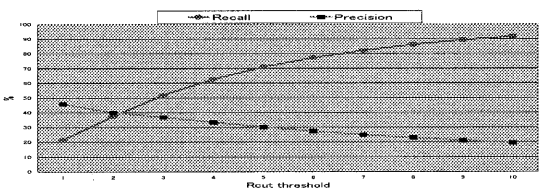
3.1 SVMについて

SVM による分類 [1] は, 訓練データの文書を多次元空間上のベクトルとして表現し, それらのベクトルを正例と負例に 2 分する超平面を求め, その超平面に基づきテストデータを分類する手法である. 文書に対応するベクトルは, 各次元に基底とする単語を, 各成分は TFIDF により重みづけされたスコアを割り当てることで生成する. 基底とする単語はストップワードを除いた総頻度上位 10,000 単語とした.

3.2 実験

実験に用いた訓練データ, テストデータはそれぞれ INSPEC データから無作為に選定した 10,000 文献とした. 本手法,SVM による分類を行い, 適合率 (Precision), 再現率 (Recall) を算出した. 本手法は順位付け手法 (ranking system) であり,SVM は二分類手法 (multiple binary classification task) であるため, そのままでは比較できない. そこで本手法を二分類手法として扱うために, 文書に対して付けられた各カテゴリのスコアを高い順に並べ, 任意の順位より高いカテゴリにその文書は属し, 低いカテゴリには属さないとする手法である Rcut[2] を用いた. 図 1 に結果を示す. 二分類手法においては, 各評価値がカテゴリごとに求まる. そこで microaveraging[2] を用いて手法としての 1 つの評価値を算出した. その結果, 適合率, 再現率はそれぞれ,50.8%,52.9%となった.

図 1: 本手法の各評価値



3.3 考察

文書は通常複数のカテゴリに属し, その数は文書ごとに異なる. そのため Rcut のように, 全ての文書を同じ順位を基に属す, 属さないを割り当てるのはふさわしくないと考えられる. そこで文書ごとに異なる割り当てを行なうため, 各文書のスコアの変化に着目する. ここではスコアの減少が最も大きい場所で決定する. 以下の例においては, スコアの減少が最も大きい 2 位と 3 位の間で分ける. すなわち 2 位以上の C31 と C32 に属し, 3 位以下のすべてのカテゴリには属さないとなる

例 1：各カテゴリにおける入力文書のスコア

We employ the HTS magnet interaction in the mechanical design of a vibration isolator. One common element of space structures is the coupling between multiple substructures or mechanical parts often ..... The concept can also be used as an isolation platform and can combine with the active vibration isolation technology so as to attenuate the vibration of all frequencies.

分野コード : C31 C33

分野コード	スコア	分野コード	スコア
C31	159.36	C42	-11.76
C32	144.84	C01	-17.09
C13	86.95	C51	-21.89
C33	60.17	C55	-24.24
C53	45.23	C56	-25.23
C74	21.04	C71	-41.34
C54	14.96	C73	-43.43
C41	10.37	C78	-47.78
C12	-0.90	C03	-54.62
C11	-9.00	C72	-66.86
C52	-10.38	C02	-87.15
C61	-11.57		

また, 文書が特徴語をどの程度含んでいるかということにも着目する. 文書が特徴語を多く含んでいる文書はカテゴリ間でのスコアの差が大きいため分類しやすいと考えられる. そこでテストデータを文書の 1 単語当たりのスコアが高い文書と低い文書に分け, それぞれ評価値を算出した. 表 1 に結果を示す.

表 1: 各手法での評価値

	本手法		SVM	
	Precision	Recall	Precision	Recall
全体	19.2%	51.3%	50.8%	52.9%
特徴語多	49.5%	37.7%	50.9%	58.9%
特徴語少	15.7%	59.3%	49.9%	50.4%

表 1 から, 本手法は特徴語を多く含む文書には高い精度の分類が行なえるのに対し, SVM ではそのような差は見られない. これは単語のスコア付けの段階で,SVM では 1 つの単語に 1 つのスコアを付けるのに対し, 本手法ではカテゴリごとに異なるスコア付けをしているためであろう.

4 まとめ

本稿では, 単語の頻度情報のカテゴリ間での偏りを用いた文書の自動分類手法を提案し,SVM による分類との比較を行なうことで評価した.

参考文献

[1] Thorsten Joachims:"Text Categorization with Support Vector Machines:Learning with Many Relevant Features", Proceedings of the European Conference on Machine Learning, Springer,1998

[2] Yiming Yang,Xin Liu:"An Evaluation of Statistical Approaches to Text Categorization",Journal of Information Retrieval,1999, Vol,No.1/2,pp67-88