6T-5

ベクトル空間モデルにおけるノルムによる単語重み決定

松浦 優彦 † 上原子 正利 † 小柳 滋 †
† 立命館大学情報理工学部

1 はじめに

情報検索に用いられる代表的な検索モデルの一つにベクトル空間モデルがある．これは文書集合や検索質問を多次元ベクトルによって表現し，ベクトル間の相関量を求めることによって文書間や検索質問との類似検索を行うものである．計算機上では文書集合を文書や単語を行や列，ある文書における各単語の出現頻度を要素とする行列として容易に表現できる．

類似度を求める際には TF・IDF という重み付けがよく用いられる [1]．しかし，IDF は単語の出現頻度の数値である TF にどの程度影響させれば適切かを一般的に決定することが難しい．また，事前に不要語のリストを作成し文書ベクトルの要素から不要語を除去する操作が必要となる．以上の難点を解消するために，本論文では単語も文書と同様にベクトルと見なし，単語ベクトルに文書ベクトルと同様の操作を適用する．

2 ベクトル空間モデルの概要

本章ではベクトル空間モデルの一般的な処理の流れを説明する．

2.1 文書行列の構成

ベクトル空間モデルでは検索対象となる文書集合を計算機上で扱い易くするために次のような行列として表現する．

$$\begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{pmatrix} tf(1,1) & tf(1,2) & tf(1,3) & tf(1,4) \\ tf(2,1) & tf(2,2) & tf(2,3) & tf(2,4) \\ tf(3,1) & tf(3,2) & tf(3,3) & tf(3,4) \end{pmatrix} \end{array} \tag{1}$$

これは各行が文書を，各列が単語を表し，単語の各文書における出現回数を要素として持つ行列である．このような単語の出現頻度を TF(Term Frequency) と呼び，TF の値によって構成された行列を以下では TF 行列と呼ぶ．

文書集合を行列により表現した場合，文書も単語もベクトルとして捉えることができる．以下では文書のベクトルを文書ベクトル, 単語のベクトルを単語ベクトルと呼ぶ．

Term weighting by Euclidean norm for vector-space model

† Masahiko MATSUURA(matsurua@cpsy.cs.ritsumei.ac.jp)
† Masatoshi KAMIHARAKO(m7i@mail.goo.ne.jp)
† Shigeru OYANAGI(oyanagi@cs.ritsumei.ac.jp)

2.2 単語の重み付け

単語の重みを決定するには文書集合全体から各単語の重要性を示す値を計算に加える必要がある．代表的な尺度に IDF がある． 以下では文書を特徴づける上で役に立つ単語を特徴語，不要語リストには含まれないが役に立たない単語を一般語と呼ぶ．

単語 t の IDF は次のように定義される [1].

$$idf(t) = \log \frac{N}{df(t)} + 1 \tag{2}$$

ただし $df(t)$ は文書集合全体で単語 t が出現する文書の総数，N は総文書数である．

2.3 文書の正規化

TF は文書長の違いの影響を受けやすい．これに対処するため，文書ベクトルのノルムを 1 に揃えるように正規化する操作が一般に行われる．ノルムは次のように定義される． ここで M は (1) の列の数を表す．

$$\|d_i\| = \sqrt{\sum_{j=1}^{M} tf^2(i,j)} \tag{3}$$

$\|d_i\|$ で d_i の各要素を割ることで各文書ベクトルのノルムを正規化することができる．

2.4 類似文書の検索

与えられた文書集合から類似文書を検索するには，以下の処理を行う．

1. 文書集合から TF 行列を作成する．
2. 手順 (1) の行列の単語を，IDF で重み付けする．
3. 手順 (2) で得られた行列を文書正規化する．
4. 指定された文書間の内積を求め，それらの類似度とする．

手順 (3) で正規化し手順 (4) で内積を求めることは，余弦の計算に相当する．

3 既存の重み付け手法の問題点とその解決

3.1 既存の重み付け手法の問題点

IDF による重み付けは単語の重要性の決定を適切に行えるとは限らない． この原因は，IDF を TF 行列にどの程度影響させれば適切かが文書集合によって異なり，一般的に決定できないことにある．影響力が小さすぎると特徴語の重みが一般語を上回れず，一般語の除去は不要語リストの性能に大きく依存してしまう．

逆に影響させすぎるとノイズのような単語の重みを大きくしてしまう.

3.2 単語の正規化による解決

2.3 節で文書長の違いが TF に与える影響を統一するための文書正規化を説明した．長い文書の文書ベクトルは要素に非ゼロ要素が多く値も高い．それに対して短い文書の文書ベクトルは非ゼロ要素が少なく値も小さい．この性質は一般語と特徴語の単語ベクトルの違いに等しい．この点に注目した重み付けとして，単語ベクトルの正規化がある [3]．以下ではこの操作を単語正規化と呼ぶ.

単語正規化は文書正規化と同じ操作を単語ベクトルに対して行うものである．単語 t のノルムは次のように定義される． ここで N は (1) の総文書数を表す.

$$\|t_j\| := \sqrt{\sum_{i=1}^{N} tf^2(i,j)} \tag{4}$$

$\|t_j\|$ で t_j の各要素を割ったものが各単語の重みとなる．単語正規化は不要語リストを使用せずに重要な単語に大きな重みを与えることができる．このことは 5 章で示される.

4 実験

本章では，実際のデータに対して IDF，単語正規化の 2 つの方法で重み付けを行った結果を比較する.

4.1 文書集合の構成

文書集合には UCI KDD Archive[†]で公開されている NSF Research Awards Abstracts[††] の Part1.zip[‡]を用いた．この文書集合から TF 行列を [2, p.46] の方法で作成した．この操作によって作成された TF 行列は文書数 49078，単語数 71969，非ゼロ要素数 4876169 の疎行列であった [2, p.47]．行列データ構造に AHDM を用いる [2，pp.14-15].

また以下の実験には，類似検索の基準となる文書に a900006『CRB: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitochondrial DNA and Historical Demography』を用いる.

4.2 IDF と単語正規化による重み付け

表 1 は，IDF と単語正規化を用いて重み付けされた基準文書の特徴語の上位 10 位である．ただし IDF の対数の底には 10 を用い, 非ゼロ要素数が総文書数の 1 割を超える単語ベクトルは TF 行列から除去した.

TF・IDF では一般語である popul に高い値が付いている．また，the，of，and を含んでいることから，この重み付けは不要語リストを必要とすることがわかる．それに対して，単語正規化による重み付けでは一般語と不要語に低い値を付けている.

表 1: IDF と単語正規化による単語の重み

順位	TFIDF	単語正規化
1	whale	whale
2	popul	subdivid
3	the	exploit
4	humpback	crb
5	of	officeseek
6	genet	somewhat
7	exploit	gray
8	mysticet	genealog
9	and	magent
10	size	migratori

これらの方法で重み付けされた TF 行列から基準文書と類似した文書を検索した結果を比較した．TF・IDF は語幹 popul が支配的であったのに対し，単語正規化は語幹 whale が支配的であった．語幹 whale は基準文書の特徴をよく表しており，この結果は妥当である．以上のことから，単語正規化は IDF の難点が解消していると言える.

単語正規化は文献 [3] の LSI を用いた実験において，IDF とエントロピーよりも性能が悪いとされている．この結果の違いは，LSI とベクトル空間モデルの違いに起因すると考えられる．LSI は予め文書が分類されていることが前提であり，実験の結果は分類の性能に依存する．本研究の実験では，基準文書が属する分野は Life Sience Biology である．しかし，文書の特徴を表す語としては語幹 whale が妥当であり，本実験では語幹 whale によって結びついた文書を関連付けている.

5 おわりに

本論文は，ベクトル空間法における単語を文書と同様のベクトルと見なし，単語ベクトルのノルムによる重み付けを行った． この方法を用いることで，IDF による単語重み付けの難点を解消した.

参考文献

[1] 徳永 “ 情報検索と言語処理 ”, 東京大学出版会 (1999).

[2] 上原子 “ 関連要素決定問題の行列表現とその解法 ”, 博士論文, 立命館大学 (2006).

[3] Dumais, Improving the retrieval of information from external sources, Behavior Research Methods, Instruments, & Computers, 23 (2), 229-236(1991).

† http://kdd.ics.uci.edu/

†† http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html

‡ http://kdd.ics.uci.edu/databases/nsfabs/Part1.zip