DEWS2007 E1-6

ベクトル空間モデルにおける単語重み決定の一般化

松浦　優彦† 　上原子正利† 　小柳 　滋†

† 立命館大学情報理工学部
〒 525-8577 滋賀県草津市野路東 1 丁目 1-1
E-mail: †matsuura@cpsy.cs.ritsumei.ac.jp, ††m7i@mail.goo.ne.jp, †††oyanagi@cs.ritsumei.ac.jp

あらまし　ベクトル空間モデルでのベクトルは一般的に文書を表し, 単語を明示的にベクトルとして捉えることは少ない. 本論文では, 明示的に単語をベクトルと捉え, 単語の重み付けを文書に対するものと同様の操作とみなす. これにより, 文書に対する類似度計算の方法をそのまま単語の類似度計算に用いることができる. また, この見方に立つと, 単語の重み付けに文書の正規化と同じ操作を用いることが自然になる. この単語正規化の考え方は既存の文献でも取り上げられているが, 現在はあまり議論されない. 本論文では, 複数の重み付け方法で類似文書検索を行い, その結果を比較した. 用いた方法は IDF, 信号／雑音比, 単語正規化である. その結果, 信号／雑音比は性能が悪く, IDF と単語正規化は性能の優劣が文書によって変わった. また, IDF と単語正規化は異なる観点で類似した文書を出力した. 本論文ではまた, 分野情報を架空の単語ベクトルとみなし, このベクトルとの類似度を重み付けに用いる類似文書検索を行なった. この重み付けは特定の分野にだけ頻出する単語を重要視する. これを IDF あるいは単語正規化と組み合わせると, それらを単独で用いるより性能が優れ, また, 異なる観点で類似した文書を検索することができた.
キーワード　ベクトル空間モデル, 重み付け, 分野情報, 単語ベクトル

A generalization of term-weighting method for vector-space model

Masahiko MATSUURA†, Masatoshi KAMIHARAKO†, and Shigeru OYANAGI†

† Ritsumeikan University
Department of Computer Science, College of Information Science and Engineering
Nojihigasi 1-1-1, Kusatsu, Shiga, 525-8577
E-mail: †matsuura@cpsy.cs.ritsumei.ac.jp, ††m7i@mail.goo.ne.jp, †††oyanagi@cs.ritsumei.ac.jp

Abstract　In vector-space model, a word “vector” is usually used for a document, but it is not common to explicitly regard a term as a vector. In this paper we explicitly regard a term as a vector and term-weighting in the same way as a operation for a document. From this viewpoint, we can use a method of document similarity calculation for words. And it becomes natural to use a document normalization method for term-weighting. This idea of term-normalization is described in a old paper, but is not discussed today. In this paper we employ three term-weighting methods, IDF, signal-noise ratio and term-normalization, for similar document search and compare the results. The comparison shows that signal-noise ratio underperforms other methods, and the winner between IDF and term-normalization changes as to a base document used for similarity calculation. And the resultant similar documents of IDF have different viewpoint from ones of term-normalization. We also show some results of similar document search made by another term-weighting method which employs category data as pseudo term vectors and uses similarities between those vectors and term vectors as term weights. This weighting method puts much importance on vectors of selectively appearing in certain categories. When we combine this method with IDF or term-normalization, we get better results than the results produced by other methods, and the resultant similar documents of this method have different viewpoint from ones of IDF or term-normalization.
Key words　Vector-space model, Term-weighting, Category data, Term vector

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, ベクトル間の類似性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．

文書ベクトル間の相関を求めるには一般に余弦が用いられる．しかし, 余弦の計算に単語の出現回数を用いると, どの文書にも出現回数が多い一般的な単語の影響が大きくなる．そのため文書ベクトルの要素には, 単語の出現回数に対して各単語の重要性を示す尺度で重み付けを行う必要がある．

文書集合全体での各単語の分布を考慮した重みとして, IDF(Inverse Document Frequency) [1] がよく用いられる．これは各単語が文書集合全体の中でどれくらいの文書に出現するかという割合を求め, その逆数の対数をとったものである．つまり, わずかな文書にしか出現しない単語の重みは大きくなり, 多くの文書に出現する単語の重みは小さくなる．一般にこの尺度を単語の出現回数 (Term Frequency) に乗じる TF・IDF という重み付けがよく用いられる．

IDF の他に, 文書集合全体における各単語の出現回数のエントロピーから単語の重要性を示す尺度として, 信号／雑音比 (Signal-Noise Ratio) がある．これは単語の出現回数の合計の対数を信号, エントロピーの大きさを雑音とし, その比を単語の重みとする尺度である [2, pp.28-30]．また, 文書集合を行列として表現すると単語もベクトルと捉えることができる．そこで, 文書正規化 [3, pp.38-39] のように単語のベクトルの大きさを正規化することによって重み付けを行う方法もある [4]．本論文では, この方法を単語正規化と呼ぶ．

以上の 3 つの方法のうち, 一般的には IDF が最もよく用いられている．我々は, その原因を調べるために実際のデータに対して 3 つの方法によって重み付けを行い類似文書を検索した．その結果, 基準文書によって各方法の適合率の優劣が変わった．適合した文書の内容を調べると, 重み付けの方法を変えることで類似性の観点が変わることがわかった．

文書間の類似性について考えると, 一般に類似性の観点には様々なものがあると思われる [5]．異なる重み付け方法によって異なる観点から類似した文書を検索できることから, 目的に応じた重み付け方法の使い分けが重要になると考えられる．

さらに, 類似性の判断に関連していれば異種データでも重み付けに使えると考えられる．最も利用しやすい異種データとしては分野情報がある．本論文では, 分野情報を単語の重み付けに利用する方法を示す．その方法の基本的な方針は, 特定の分野にだけ頻出する単語を重要視し, ベクトル空間モデルによってその重要度を求めるというものである．この重み付けを用いると, IDF などとは異なる観点の類似文書を高い適合率で検索することができた．

本論文の構成は次の通りである．2 章ではベクトル空間モデルの概要について述べる．ここでは, 文書集合の行列表現を定義し, 単語の重み付けである IDF, 信号／雑音比, 単語正規化と, 文書正規化を説明する．3 章では IDF, 信号／雑音比, 単語正規化の 3 つの方法を用いて重み付けを行い, 類似文書検索を行なった結果を確認する．4 章では単語の出現回数以外のメタデータから単語の重み付けを行う方法を定義する．5 章では既存の重み付けと新たに定義した重み付けを類似文書の検索に適用した結果を比較する．6 章では残された問題について述べる．

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による類似文書検索の一般的な処理の流れを説明する．この処理では、まず各文書を単語に分解して文書と単語の行列を作り、次に単語を重み付けと文書の正規化を行い, 最後に文書間の類似度を求める。

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．

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

これは各行が文書を, 各列が単語を表し, 各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．例えば文書 d_1 は単語 t_1 を $tf(1,1)$ 個, 単語 t_2 を $tf(1,2)$ 個, 単語 t_3 を $tf(1,3)$ 個含む．このような単語の出現回数は TF(Term Frequency) と呼ばれる, 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．

2.2 単語の重み付け

TF は局所的な文書の特徴を示すと言えるが, 文書集合全体から大局的に捉えれば a や the のような多くの文書に出現するような単語は文書を特徴付ける上では役に立たない．このような単語は不要語と呼ばれ、重み付けの前にはあらかじめ作成された不要語リストに従って不要語を除去する操作が行われる．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尺度を計算に加える必要がある．一般的にはこれらの尺度を単語ベクトルの各要素に乗じた値を各単語の重みとする．既存の尺度に IDF と信号／雑音比, 単語正規化がある．

IDF は単語の重要性を示すのに最もよく用いられる尺度である．単語 t の IDF は次のように定義される [2]．

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

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．この式の基本的な考え方は, 単語 t が全文書中の多くの文書に出現する場合は一般語とみなし値を小さくし, わずかな文書にしか出現しない場合は特徴語とみなし値を大きくするというものである．

信号／雑音比はある単語が各文書にどれくらいばらついて出

現するかの尺度, すなわちエントロピーを雑音と考え, 雑音が小さいほど重要性の高い単語と捉える尺度である [2]. 雑音は次のように定義される.

$$n_t = -\sum_{i=0}^{N} \frac{tf(i,t)}{\sum_{j=1}^{N} tf(j,t)} \log_2 \frac{tf(i,t)}{\sum_{j=1}^{N} tf(j,t)} \tag{3}$$

信号／雑音比の値は雑音の増減を逆転させたものであり, 次のようになる.

$$s_t = \log_2 \sum_{k=1}^{N} tf(k,t) - n_t \tag{4}$$

IDF や信号／雑音比は, 特徴語ベクトルのユークリッドノルム (以下ノルム) を一般語よりも大きくしようとするものである. しかし, 全ての単語ベクトルのノルムを正規化することによって, 文書ベクトル間の類似性計算に対する一般語ベクトルの影響を小さくする方法がある [4]. 正規化を行うとノルムの大きい一般語ベクトルの各要素の値は小さくなり, ノルムの小さい特徴語ベクトルの各要素の値はあまり変わらない. そのため, 文書ベクトル間の類似性計算において特徴語ベクトルが重要視されるようになる. 以下では単語ベクトルを正規化する操作を単語正規化と呼ぶ. 単語ベクトル t_j のノルムは次のように定義される. ここで N は総文書数を表す.

$$\|t_j\| := \sqrt{\sum_{i=1}^{N} tf^2(i,j)} \tag{5}$$

$\|t_j\|$ で t_j の各要素を割ったものが各単語の重みとなる.

2.3 文書の正規化

文書集合に含まれる文書には長さに違いがある. そのため, TF の値は同じ 10 でも, その単語が 10000 語の文書で 10 回出現する場合と 1000 語の文書で 10 回出現する場合とでは, その重要性が異なる. このように, 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 は式 (1) の列の数を表す.

$$\|d_i\| = \sqrt{\sum_{j=1}^{M} tf^2(i,j)} \tag{6}$$

$\|d_i\|$ 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) の行列の単語を何らかの方法で重み付けする.

（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. 各重み付け方法による類似文書検索の実験と考察

本章では、実際のデータに対して IDF, 信号／雑音比, 単語正規化の 3 つの方法で重み付けを行い, 類似文書を検索した結果を比較する.

3.1 文書集合の構成

文書集合には UCI KDD Archive [6] で公開されている NSF Research Awards Abstracts(注1) の Part1.zip(注2)を用いた. この文書集合から TF 行列を以下の方法で作成した [7, p.46].

（1） 各文書からタイトル部と概要部の文字列を抜き出す. 概要部が空の文書は無視する.

（2） 全ての文書の文字列を空白で単語に分けて小文字に変換し，語幹を取り出し，全ての語幹の集合を作る.

（3） 全ての語幹をソートして，語幹に番号を付ける.

（4） 各文書を，出現する語幹の番号とその語幹の出現回数のペアを要素とする集合に変換する. 全ての文書にこれを行う.

以上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, 密度 0.0014 の疎行列であった [7, p.47]. 本実験では大きな疎行列に対して行と列の両方を操作するため、行列データ構造に AHDM を用いる [7, 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次の 5 文書を用いる. ただし, a9000006 などは文書の番号である.

（1） a9000006『CRB: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitochondrial DNA and Historical Demography』

（2） a9221396『U.S.-Kenya Workshop to Develop an African Drought Research Agenda; December 11-14, 1992; Nairobi, Kenya』

（3） a9012737『U.S.- Jordan Cooperative Research: Electrical Studies of Ferroelectric Thin Films』

（4） a9014726『Gas Phase Photolysis with Matrix Isolation: A Quest for Novel Reactive Intermediates』

（5） a9312851『Experimental and Fracture Mechanics Fatigue Studies of Double Angle Railway Bridge Connections』

検索結果の評価は, 各重み付け方法に対してプーリング [3, pp.23-24] と呼ばれる手法を用いて行う. この手法は, 検索結果の上位に順位付けされる文書の適合性だけを調べ, 適合情報を作成するものである. 本論文では検索結果の上位 10 件だけの適合性を調べる. また, 適合率を次式のように定義する.

$$\text{適合率} = \frac{\text{検索結果上位 10 件中の適合した文書数}}{10} \tag{7}$$

図 1 は 5 つの各基準文書を IDF, 信号雑音比, 単語正規化の 3 つの方法でそれぞれ重み付けを行い, 類似文書を検索した結果の適合率を示したものである. ただし IDF の対数の底には 10 を用いている. 各重み付け方法によって重視された特徴語上位

(注1)：http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html
(注2)：http://kdd.ics.uci.edu/databases/nsfabs/Part1.zip

表 1　a900006 の各重み付け方法による特徴語リスト

順位	IDF	信号／雑音比	単語正規化
1	whale	year	humpback
2	popul	world	bowhead
3	the	wide	drove
4	humpback	variat	magnific
5	of	univers	palumbi
6	genet	two	mysticet
7	exploit	subdivid	whale
8	mysticet	stephen	subdivid
9	and	southern	exploit
10	size	somewhat	crb

表 2　a900006 の IDF による類似文書の検索結果

順位	文書名	文書の title
基準	a9000006	CRB: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitochondrial DNA and Historical Demograph
1	a9024592	U.S.-Sweden Cooperative Research: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitol Historical Demographic Studies
2	a9208369	Public Attitudes to Whales and Whaling: An International Study
3	a9113342	Evolutionary Ecology of Structured Populations
4	a9000063	Population Biology of Tropical Rain Forest Trees
5	a9207278	Doctoral Dissertation Research in Geography and Regional Science
6	a9212583	Dissertation Research: Importance of Genetic Factors on Fecundity and Survival of Small Populations
7	a9207558	Likelihood Methods for Population Samples of Sequences
8	a9307694	Nonadditive Genetic Variance: The Genetical Consequences of Population Structure
9	a9211945	ABR: Developments in Matrix Population Analysis
10	a9100002	CRB: Population Size and Density Effects on Population Viability: A Case Study of Two Cirsium Species

表 3　a900006 の信号／雑音比による類似文書の検索結果

順位	文書名	文書の title
基準	a9000006	CRB: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitochondrial DNA and Historical Demograph
1	a9000063	Population Biology of Tropical Rain Forest Trees
2	a9113342	Evolutionary Ecology of Structured Populations
3	a9221175	Ecological Determinants of Genetic Structure in a Plant Meta-population
4	a9006285	Mathematical Models of Geographic and Phenotypic Variation
5	a9024592	U.S.-Sweden Cooperative Research: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitol Historical Demographic Studies
6	a9415669	Genetic Structure of Sea Urchin Populations Across a Biogeographic Boundary
7	a9212583	Dissertation Research: Importance of Genetic Factors on Fecundity and Survival of Small Populations
8	a9100860	Dissertation Research: Dispersal in Metapopulations of Butterflies: Implications for the Dynamics and Genetic Structure of Local Populations
9	a9214040	Diversification of Populations and Species of Bacillus
10	a9213184	Dissertation Research: An Analysis of the Speciation Process in Cave Spiders of the Genus Nesticus

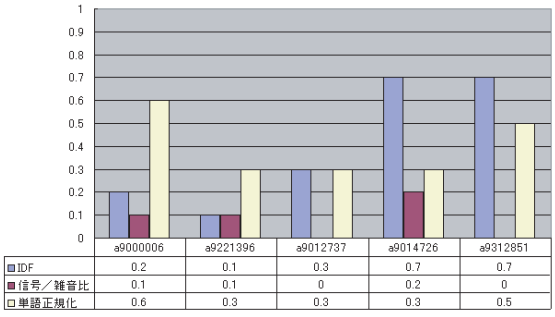

図 1　各方法による検索結果の適合率

10 語と検索結果上位 10 件は a9000006 についてのみ表 1, 表 2, 表 3, 表 4 に示す．信号／雑音比はどの基準文書でも最も適合率が低い．IDF と単語正規化を比較すると a900006 と a9221396 では単語正規化の方が適合率が高く, a9014726 と a3312851 を比較すると IDF の方が高い．

a9000006 は絶滅危惧種であるヒゲクジラの保護と管理のために, 個体数と他種のクジラとのミトコンドリア DNA の関係を研究する文書である．タイトルに含まれる CRB は Conservation and Restoration Biology の略であり, 生態系の保護や回復に関する文書であることを意味する．この文書の場合, IDF では語幹 popul, 単語正規化では語幹 whale や crb で文書間が関連付いていた．語幹 whale や crb で関連付いていることは明らかに妥当である．語幹 popul は判断が難しい．なぜなら, これは多くの文書に出現し得る一般語であると考えられるが, 一方で個体数の調査という観点から考えれば妥当でないとは言い難いためである．

a9012737 はアメリカとヨルダン (Jordan) が共同で行った強誘電性薄膜研究の文書である．この基準文書の場合, IDF でも単語正規化でも適合率は等しいが, IDF では語幹 Jordan が地名として重視されており, 単語正規化では語幹 Jordan が人名として重視されていた．

表 4　a900006 の単語正規化による類似文書の検索結果

順位	文書名	文書の title
基準	a9000006	CRB: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitochondrial DNA and Historical Demograph
1	a9354913	Mechatronics in Machine Tool Research. (Focus Area-Machine Tool Research)
2	a9024592	U.S.-Sweden Cooperative Research: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitol Historical Demographic Studies
3	a9208369	Public Attitudes to Whales and Whaling: An International Study
4	a9313627	Dissertation Research: Differential Resource Access in a Thule Eskimo Whaling Community
5	a9419898	RUI - Phylogenetic Analysis of the Cetacean Basal Cranium and Inner Ear and Study of the Cetacean Relationships
6	a9000162	Community Structure and Dynamics in a Newly Discovered Deep Sea Reducing Habitat: Lipid-Rich Whale Bones
7	a9107144	Dissertation Research: Technological Variability in the Portuguese Magdalenian
8	a9096227	Behavior of Hatchling Sea Turtles During Their Off-Shore Migration
9	a9214940	Dissertation Research: Late Magdalenian Technology
10	a9253377	Whales, Sharks & Things In The Dark

このように, 重み付け方法によって文書ごとの適合率の優劣が変わり, また得られる文書の類似性の観点も変わる. そのため, 各重み付け方法は単に適合率の優劣によって性能の優劣を決定できるものでなく, 検索する側の観点によって重み付けの方法を使い分けることが重要であると考えられる.

4. 分野情報とベクトル空間モデルを用いた重み決定

前章では, 単語の重み付け方法によって異なる観点から関連付いた文書を検索できることを示した. また, 単に適合率の高い重み付け方法を使うのではなく, 目的に応じて重み付け方法を使い分けることが重要であることも述べた. 本章ではさらに進み, IDF や単語正規化とは異なった観点から重み付けする方法を示す.

4.1 分野情報からの観点

ここまでに述べた重み付けは TF 行列内の値の分布だけに基づいたものである. したがって, 文書間の類似性は文書内の単語の分布から判断される. しかし, 予め分類され分野情報の付随した文書であれば, 分野の観点から重み付けを行うことで文書間の類似性を判断することもできる [5].

分野の観点を反映した重み付けを行う場合, 特定の分野の文書集合でしか頻出しないような単語は文書集合に関わらず重要であると考えられる. 以下ではこのような単語を分野特徴語と呼ぶ. また, 分野特徴語の重要性を示す値を分野特徴値と呼ぶ.

分野特徴語の抽出は文書の自動分類に用いられることがある [8]. それに対して本論文では分野特徴値を, IDF などと同様の単語に対する重みの決定に用いる. この方法で決定した値を重み付けに用いると, IDF などとは異なった観点から単語の重要性を決定できる. この方法は文書に対して行う類似性の計算を単語に対して行うものであり, 新しい概念を必要としない単純なものである.

4.2 単語の分野特徴値の決定

本論文では, 分野特徴値の決定にベクトル空間モデルをそのまま用いる. 通常のベクトル空間モデルは文書検索に用いられ, 検索質問ベクトルとして架空の文書ベクトルが作成される. しかし, 文書集合を行列として捉えれば行と列の操作を入れ換えることで単語検索に用いることができるため, 本論文では検索質問ベクトルとして架空の単語ベクトルを作成する. このベクトルと各単語ベクトルの類似度を求めることで, 各単語のその分野に対する分野特徴値を決定することができる. 例として次のような TF 行列を考える.

$$
\begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \\ d_8 \\ d_9 \end{array}
\begin{array}{c}
\begin{array}{ccccc} t_1 & t_2 & t_3 & t_4 & query \end{array} \\
\left(\begin{array}{ccccc}
0 & 3 & 1 & 2 & 1 \\
0 & 2 & 0 & 3 & 1 \\
2 & 3 & 0 & 2 & 1 \\
0 & 0 & 2 & 2 & 0 \\
0 & 2 & 1 & 1 & 0 \\
1 & 0 & 3 & 0 & 0 \\
0 & 0 & 2 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0
\end{array} \right)
\end{array}
\tag{8}
$$

ここで $query$ は分野 c の検索質問ベクトルである. d_1 から d_3 までの文書は分野 c に属し, d_4 から d_9 までの文書は c 以外の分野に属する. $query$ では分野 c に属する文書の行が 1 に, それ以外の行は 0 になる. 分野 c のみに目立って出現する単語は t_2 であり, $t3$ は分野 c 以外で、t_1 と t_4 は分野 c の中にも外にも出現する単語である.

単語ベクトルと検索質問ベクトルとの類似度を求めるには, 文書検索と同じく, 文書正規化と単語正規化を行う必要がある. 正規化を行う順序は文書検索の逆となり, 文書正規化の次に単語正規化を行う. 検索質問に対しては単語正規化のみを行う. その後に各単語ベクトルと検索質問ベクトルとの内積を求める. これをその分野に対する各単語の分野特徴値とする. これを全ての分野に対して行う.

表 5　分野 c の分野特徴値リスト

単語番号	特徴値
t_1	0.49
t_2	0.79
t_3	0.10
t_4	0.52

表 5 は分野 c における各単語の特徴値である．分野 c 内でのみ頻出する t_2 の特徴値は最も高く, 分野 c 以外でのみ頻出する t_3 の特徴値は最も低い．これらの値は分野特徴値の性質をよく表しており, 妥当な結果である．

特徴値は単語ベクトル間の余弦に相当するため, 0 から 1 までの範囲に正規化されたものである．しかし分野ごとの特徴値には, 各分野に属する文書数などによって大きな差が出る場合がある．これに対処するためにその分野の最大特徴値で各特徴値を割ることで正規化する．表 5 の値を正規化したものが表 6 である．

表 6　正規化された分野 c の分野特徴値リスト

単語番号	特徴値
t_1	0.62
t_2	1
t_3	0.13
t_4	0.66

4.3　分野特徴値による重み付け

以上の方法で決定された分野特徴値で重み付けを行うには, それらの値に 1 を加え各単語ベクトルに乗じていく．1 を加える理由は, 図 2 の (a) に分野特徴値による重みを加えることで (b) のようにするためである．分野特徴値の最大値が 1 であるため, 最終的な単語の重みは (a) から最大 2 倍までの範囲で増加する．ただし, 分野特徴値による重み付けは特徴語の重みが

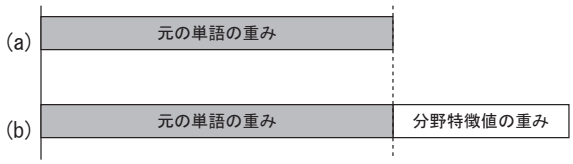

図 2　分野特徴値による重みの増加

一般語を上回ることを保証できないため, 一般語に対する処理として各単語に何らかの重み付けを先に行っておく必要があると考えられる．

分野 c についての重み付けは表 6 の t_n の特徴値を式 (8) の t_n ベクトルの各要素に乗じることで行われる．この操作を全ての分野で繰り返す．

4.4　分野特徴値による重み付けを用いた類似文書検索

以上の操作を用いて類似文書を検索する処理は, 次のようにまとめられる．

（1） 文書集合から TF 行列を作成する．

（2） TF 行列を文書正規化し, 次に単語正規化する．

（3） 分野の検索質問ベクトル c_i について以下の操作を繰り返す．

（a） c_i を単語正規化する．

（b） 手順 (3a) のベクトルと, 手順 (2) の行列の各単語ベクトルとの内積を計算する．

（c） 各単語の内積を, c_i の分野特徴値としてリスト l_i に保存する．

（d） l_i の最大特徴値で l_i の各特徴値を割る．

（4） 手順 (1) の行列を何らかの方法で重み付けする．

（5） 各 l_i について以下を繰り返す．

（a） l_i に含まれる各単語について, その分野特徴値に 1 を足したものを手順 (4) の行列の単語ベクトルにかける．

（6） 手順 (5) で得られた行列を文書正規化する．

（7） 指定された文書間の内積を求め, それらの類似度とする．

5.　分野特徴値を用いた類似文書検索の実験と考察

本章では, 実際のデータに対して分野特徴値のみ, IDF と分野特徴値の組み合わせ, 単語正規化と分野特徴値の組み合わせの 3 つの方法で重み付けを行い, 類似文書を検索した結果を比較する．

5.1　文書集合の構成

文書集合には 3 章のものを同様の手順で TF 行列に加工したものを用いる．基準文書にも 3 章と同様の 5 文書を用いる．

4 章で, 分野特徴値による重み付けは何らかの重み付け方法と組み合わせる必要があると述べたが, 本章では, IDF と単語正規化を用いる．そして, 3 章と同様に類似文書検索を行い, 検索結果を比較する．適合率の低かった信号／雑音比は今回用いない．評価方法は, 3 章と同様のものを用いる．

5.2　分野特徴語の抽出

前節で作成した TF 行列から分野 Physics について作成した分野特徴値リストの上位 20 位を表 7 に示す．ただし, 非ゼロ要素数が総文書数の 1/25 を超える単語ベクトルと, 非ゼロ要素数が 5 以下の単語ベクトルは TF 行列から除去した．また, 正規化をする前の特徴値が 0.1 を下回る単語についてはリストから除去した．

表 7 の単語は全て分野 Physics に関連している語幹であると考えられ, この結果は妥当であるといえる．その他の分野についても妥当な結果が得られた．

5.3　分野特徴値とそれによる類似文書検索

a9000006 を含めた 5 つの基準文書について, 分野特徴値と単語正規化の組み合わせ, 分野特徴値と IDF の組み合わせ, 分野特徴値の 3 つの方法で類似文書検索を行った結果の上位 10 件の適合率を図 3 に示す．

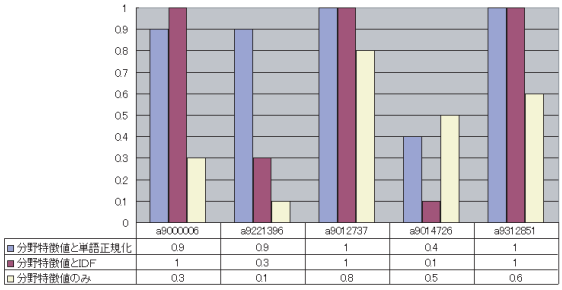

図 3　各方法による検索結果の適合率

基準文書 a9000006 の場合は, 分野特徴値と単語正規化の組み合わせでは CRB や遺伝に関する文書を高い適合率で検索し

表 7　分野 Physics の分野特徴値リスト

順位	語幹	分野特徴値	語幹の意味
1	nuclear	1	原子核の
2	atom	0.87	原子
3	quark	0.82	クォーク
4	quantum	0.73	量子
5	nucleon	0.72	核子
6	detector	0.69	探知器、検出器
7	elementari	0.68	元素
8	collis	0.60	衝突
9	fermilab	0.59	フェルミ国立加速器研究所
10	meson	0.57	中間子
11	nuclei	0.56	原子核
12	matter	0.56	物質、物体
13	relativist	0.55	相対性理論
14	decay	0.54	崩壊する
15	gravit	0.54	重力
16	physicist	0.54	物理学者
17	hadron	0.53	ハドロン
18	acceler	0.53	加速
19	scatter	0.52	散乱、散布
20	collid	0.52	衝突

表 8　a900006 の各重み付け方法による特徴語リスト

順位	分野特徴値と単語正規化	分野特徴値と IDF	分野特徴値のみ
1	humpback	polici	the
2	whale	isol	of
3	drove	pacif	whale
4	bowhead	whale	popul
5	magnific	dna	and
6	crb	crb	isol
7	palumbi	the	polici
8	mysticet	endang	pacif
9	endang	of	dna
10	subdivid	decis	will

ている．一方, 分野特徴値と IDF との組み合わせでは CRB や遺伝の他に, クジラに関する文書を多く検索している．

基準文書 a9221396 は, アフリカの干ばつ問題について開かれたナイロビの学会に関する文書である．分野特徴値と単語正規化の組み合わせではアメリカとケニアの共同研究や, アフリカの気候, ケニアに関する文書を高い適合率で検索している．一方, 分野特徴値と IDF との組み合わせではケニアや気候に関する文書を検索しているが適合率が低い．

基準文書 a9012737 は, 強誘電性薄膜に関するアメリカとヨルダンとの共同研究の文書である．分野特徴値と単語正規化の組み合わせでも分野特徴値と IDF の組み合わせでも全て薄膜に関する文書を検索している．これは基準文書の特徴がはっきりしており, 単語の重み付けの影響に関わらず関連文書を検索しやすいと考えられる．

基準文書 a9014726 は, 反応中間体 (Reactive Intermediates) というものに関する文書である．この基準文書ではどの重み付け方法でも適合率が低い．これは文書の特徴がはっきりしていないためと考えられる．

表 9　a900006 の分野特徴値と単語正規化による類似文書の検索結果

順位	文書名	文書の title
基準	a9000006	CRB: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitochondrial DNA and Historical Demograph
1	a9208369	Public Attitudes to Whales and Whaling: An International Study
2	a9354913	Mechatronics in Machine Tool Research. (Focus Area-Machine Tool Research)
3	a9024592	U.S.-Sweden Cooperative Research: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitol Historical Demographic Studies
4	a9322672	CRB: Conservation Genetics of Giant Galapagos Tortoises
5	a9300135	CRB: Conservation and Genetics of Pacific Salmonids
6	a9225127	CRB: Demographic and Genetic Factors in Extinction
7	a9424595	CRB: Should Molecular Genetic Diversity be Used as a Predictor of Evolutionary Potential?
8	a9122235	CRB: Diversity Theory and its Application
9	a9000162	Community Structure and Dynamics in a Newly Discovered Deep Sea Reducing Habitat: Lipid-Rich Whale Bones
10	a9296256	CRB: Conservation Genetics and Inbreeding Depression

基準文書 a9312851 は, 鉄道橋に関する文書である．分野特徴値と単語正規化の組み合わせでも分野特徴値と IDF の組み合わせでも橋に関する文書を高い適合率で検索している．これは a9012737 と同様に, 基準文書の特徴がはっきりしているからだと考えられる．

分野特徴値のみを用いた重み付けでは, 単語正規化などと組み合わせた場合と比べて適合率が低い．これは, 分野特徴値を TF 行列にどの程度影響させれば適切かを決定するのが困難であるからだと考えられる．したがって, 分野特徴値を用いて重み付けをするには事前に何らかの重み付けを行う必要があるといえる．

以上のことから, 平均的に適合率の最も高い分野特徴値と単語正規化を組み合わせた重み付けを行うことが望ましく, 別の観点から関連した文書を得たいときには IDF などの他の重み付け方法と組み合わせた重み付け方法を行うべきであると考えられる．

6. おわりに

本論文は, ベクトル空間モデルによる文書検索について, IDF, 信号／雑音比, 単語正規化の 3 つの重み付け方法を比較し, それぞれの重み付け方法ごとに異なる観点で関連した文書を検索

表 10　a900006 の分野特徴値と IDF による類似文書の検索結果

順位	文書名	文書の title
基準	a9000006	CRB: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitochondrial DNA and Historical Demograph
1	a9208369	Public Attitudes to Whales and Whaling: An International Study
2	a9024592	U.S.-Sweden Cooperative Research: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitol Historical Demographic Studies
3	a9225081	CRB: Effects of Dispersal on Demography and Genetic Varia- bility in Small Isolated Populations of the Northern Idaho Ground Squirrel: A Model System for
4	a9105791	Neural-Immune Interactions in the Beluga Whale
5	a9100002	CRB: Population Size and Density Effects on Population Viability: A Case Study of Two Cirsium Species
6	a9100057	CRB: Population Bottlenecks, Inbreeding and Estimation of Molecular Genetic Variation
7	a9000091	CRB: Genetic Variation and Estimates of Population Viability for a Rare Perennial Plant
8	a9317267	Neural-Immune Interactions in the Beluga, Delphinapterus leucas
9	a9419898	RUI - Phylogenetic Analysis of the Cetacean Basal Cranium and Inner Ear and Study of the Cetacean Relationships
10	a9300135	CRB: Conservation and Genetics of Pacific Salmonids

できることを示した．これにより, 各重み付け方法は単に適合率の優劣によって性能の優劣を決定されるものでなく, 検索する側の観点によって使い分けられるべきだと考えられる．

また, 本論文では分野情報を利用して新たな観点から重み付けを行う方法を示した．この方法と既存の重み付け方法を組み合わせた場合の類似文書検索の結果は, 既存の重み付け方法のみを用いた場合よりも適合率が高かった．さらに, 重み付け方法の組み合わせ方によって異なる観点で関連した文書を検索できることを示した．

残された問題には以下のものがある．本論文では 1 つのデータセットに対する結果しか示されていないため, 他のデータセットでも妥当性を検証する必要がある．また, 分野特徴値は TF 行列にどの程度影響させるかを一般的に決定できないという問題がある．そのため, 検索対象となる文書集合から TF 行列への影響のさせ方を一意に決定できる方法を考える必要がある．さらに, 分野情報以外のメタデータの利用も考えられる．

文　　献

[1] K. S. Jones: “A statistical interpretation of term specificity and its application in retrieval”, Journal of Documentation, **60**, 5, pp. 493–502 (2004).

表 11　a900006 の分野特徴値による類似文書の検索結果

順位	文書名	文書の title
基準	a9000006	CRB: Genetic Diversity of Endangered Populations of Mysticete Whales: Mitochondrial DNA and Historical Demograph
1	a9317695	Systematics and Biogeography of the Pempheridídae (Pisces: Perciformes), and Relationships of the Ocean Basins
2	a9006285	Mathematical Models of Geographic and Phenotypic Variation
3	a9112821	Mathematical Sciences: Analysis and Parameter Estimation of Models with Internal Structure
4	a9022821	Economic Opportunity in Urban America, 1850-1870
5	a9100002	CRB: Population Size and Density Effects on Population Viability: A Case Study of Two Cirsium Species
6	a9020126	Phylogeny, Speciation, and Systematics of Mielichhoferia (Musci)
7	a9000091	CRB: Genetic Variation and Estimates of Population Viability for a Rare Perennial Plant
8	a9105791	Neural-Immune Interactions in the Beluga Whale
9	a9016931	Synthesis and Study of Some Theoretically Interesting Molecules
10	a9424246	Astronomical Society of the Pacific (ASP) Symposium: Clusters, Lensing, and the Future of the Universe, College Park, Maryland, June 26-28, 1995

[2] 徳永：“情報検索と言語処理”, 東京大学出版会 (1999).
[3] 北, 津田, 獅々堀：“情報検索アルゴリズム”, 共立出版 (2002).
[4] S. T. Dumais: “Improving the retrieval of information from external sources”, Behavior Research Methods, Instruments, & Computers, **23**, 2, pp. 229–236 (1991).
[5] 大島, 小山, 田中：“文書群をクエリとした似て非なる文書の検索”, 電子情報通信学会第 17 回データ工学ワークショップ (2006).
[6] S. Hettich and S. D. Bay: “The UCI KDD Archive”, Irvine, CA: University of California, Department of Information and Computer Science (1999). `http://kdd.ics.uci.edu`.
[7] 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).
[8] 呉, 山田, 岸本：“文書自動分類のための分野関連語辞書の構成”, 情報処理学会論文誌, **2000**, 29, pp. 33–39 (2000).