自然言語処理 129－6
（1999. 1. 21）

個人の選好に応じた単語の重要度の学習

持橋大地† 加来田裕和††

†奈良先端科学技術大学院大学 情報科学研究科
daiti-m@is.aist-nara.ac.jp

††ＮＴＴソフトウェア研究所
kakuda@slab.ntt.co.jp

概要

情報検索・メッセージ処理などにおいて，単語に重みづけを行うことは基本的で重要な課題である．従来このための手法として tf・idf が用いられてきたが，tf・idf は文脈を考慮していないため，重要な語を落としてしまう可能性がある．本研究では，単語の重要度の基準として周辺分布に着目し，頻度と組み合わせた形での指標を提案する．この手法はテキストが文書に分かれない環境でも重みづけが可能であり，学習データによって適応的な重みづけが得られる．また，電子メールの重要性判定に適用することで，内容による優先度判定やフィルタリングが行えることが示唆された．
[キーワード] 重みづけ，tf・idf，フィルタリング，情報検索，適応．

Adapted Automated Learning of the Weights of Words

Mochihashi Daichi
Graduate School of Information Science
NAIST

Kakuda Hirokazu
NTT Software Laboratories

Abstract

Contrary to recent developments on Information Retrieval techniques, word weighting mechanism is still based on simple tf.idf model solely based on the statistics of word occurrences. According to this property, system may overlook some important words and is not applicable to the environments in which texts are not segmented into files.

This paper proposes a new method of word weighting to solve these problems, based on word co-occurrence contexts and frequencies. Experimental application of this method to the precedence decision of electronic mail suggested its validity and adaptiveness.
[keywords] word weighting, tf・idf, filtering, Information Retrieval, adaptation.

1 はじめに

ネットワーク上のコミュニケーションの発達により，われわれは大量の文書を処理する必要に迫られている．一度に処理できる量は有限であるから，われわれは各々の基準によって心的にこれらに優先順位をつけることで，作業の効率化を図っていると考えることができる．

しかし，目で見て優先順位をつけられる範囲には限りがあり，また負荷も大きいことから，工学的に個人

の基準に応じて文書の重要度を判定することが求められる．このことは情報検索の分野では特に大きな問題となっており，適合性フィードバックを行うなど様々な手法が開発されているが [1]，近年の進歩にもかかわらず，それらは基本となる単語の重み付けについては未だ古典的な tf·idf モデルに依っており，個人の選好を考慮しない一般的な重み付けしか得られないことや，テキストの文書への分割に依存するなどの問題がある．

本論文では連続的なテキストに対する重み付け手法として単語の周辺分布に着目し，頻度と組み合わせた形での指標を提案する．この指標では，学習テキストの指向性に応じた重み付けが得られ，連続的なテキスト環境の下で適応的に学習を行うことができる．個人が選択的に読むテキストはその個人の指向性を反映していると考えられることから，本研究では学習テキストとして個人の受け取る電子メールを用いた．

本稿の構成について述べる．2 章で古典的な tf·idf モデルとその限界について述べ，3 章で結節度 (connectivity) に基づいた単語の重み付け手法を提案する．4 章で電子メールの学習結果から新たなメールの重要度を判定する実験の結果を示し，情報のフィルタリングに応用できることを明らかにする．最後に 5 章で，まとめとモデルの精密化について述べる．

2 tf・idf モデル

tf·idf モデル [2] は，Salton(1983) によって提案された単語の重要度を示すパラメータであり，次式で定義される．

$$\mathrm{tf}_{ij} = \frac{\text{単語}w_i\text{の出現回数}}{\text{文書}j\text{に含まれる単語数}} \tag{1}$$

$$\mathrm{idf}_{ij} = \log \frac{N}{\mathrm{df}_j} \tag{2}$$

(N : 文書総数, df_j : 単語w_jが現れた文書数)

文書中の頻度が高い語は基本的に重要であると考えられるが，そのような語については tf の値が高くなり，重要性を示すことができる．しかし一方，頻度の高い語には「の」，「する」などそれ自体では特に意味を持たない機能語も含まれるため，これらの重要度は下げる必要がある．

ここで idf を用いれば，機能語は多くの文書に共通して現れるため，idf の値は低くなり，tf と idf の積 tf·idf を指標とすることで高頻度語のうち機能語を除くことができる．

また逆に，tf の値が低い低頻度語であっても，文書中での出現が稀で idf の値が高ければ，tf·idf の値は高くなり，低頻度の内容語に比較的大きな重みを付けることができる．

このように tf·idf は単語の重要度を判定する手法として大きな妥当性を持っているが，次のような単語については問題を含んでいる．

- どのテキストにも平均して少数回現れる専門用語 (tf—小, idf—大) は低い重みしか与えられない．
- 専門用語であるが，分野の中ではありふれた単語 (tf—大, idf—大) に高い重みが割り当てられる．

この原因は，モデルが (テキスト内における) 単語の生起のみを見ており，言語的な文脈を考慮していないためだと考えることができる．

また，tf·idf モデルはテキストデータの文書への分割に依存しているため，テーマの分散した短い文書が多数ある場合には検出力を発揮するが，長い文書に対しては 1 つの文書が様々な単語を含む可能性があるため，検出力が弱まるという問題がある．

また，データ文書に分かれない通信などの環境で適応的に学習することも考えた場合，文書への分割に依存しない学習法が必要であるといえる．

3 提案する手法

それでは，どのように学習を行えば良いのだろうか．前節の議論から，単語の重要度の判定は

(i) 機能語と内容語の分離．
(ii) 頻度に比例した内容語に対する重みづけ．

の 2 段階に分けられることがわかる．

以下 3.1節で機能語と内容語の分離について結節度 (connectivity) という指標を定義し，3.2節で頻度との結合による重要度の指標について述べる．

3.1 機能語 - 内容語の分離

単語には「の」「が」などの機能語 (function word) と，意味を持つ内容語 (content word) がある [3]．情報検索・分類などによって単語の重要度を考慮する際，前者には低い重みを，後者には高い重みを割り当てる必要がある．

一般に機能語には助詞や助動詞，内容語としては名詞や動詞が多いため，品詞によって分類することがまず考えられるが，これはうまくいかない．名詞や動詞の中にも，「これ」「する」「来る」など機能語が含まれるし，助詞や接続詞の中にも「だけ」「しかしながら」など内容的要素の強い語もあるからである．

古くから提案されているように，stop list を用いて，あらかじめ機能語のリストを作成しておき，重み付けの際に除くという方法 [3] も考えられるが，この方法では

単語	結節度
の	0.3399
に	0.2140
を	0.1989
は	0.1834
が	0.1780
と	0.1474
する	0.1348
だ	0.1342
ます	0.0881
も	0.0730
で	0.0649
のだ	0.0564
ない	0.0538

:	:
源氏	0.0006
厳選	0.0006
原田	0.0006
熊谷	0.0006
和田	0.0003
役	0.0003
特典	0.0003
同時	0.0003
電	0.0003
ショップ	0.0003
サテライト	0.0003
てき	0.0003
それでは	0.0003

表 1: 結節度

- 判断の基準が作成者によって曖昧である.
- 人手ですべての単語についてリストを作成しなければならない.

といった問題点がある. 実際, 現在の主要な検索エンジンではこれらの理由に加え, 機能語を含んだ語句を検索する必要から stop list を作成していない [5].

また, 機能語 - 内容語の区別は利用者によって異なるのに加え, そもそも明確に分離されるものではなく, 連続的なものであると思われる.

ここで原点に戻って考えてみると, われわれがある語を機能語 (的要素が強い) と考えることができるのは, その語が

- 様々な文脈で現れる

からであると考えることができる. 例えば, 「が」は前に様々な語をとって現われるが, 「情報」は前後に現れる語が限られている. そこで, 単語 w_i について, 次式 (3) により結節度 (connectivity) を定義する.

$$\text{結節度} c_i = \frac{\text{単語} w_i \text{と共起した単語数}}{\text{全単語数}} \qquad (3)$$

$(0 < c_i < 1$, 単語数は異なり数$)$

実際のテキスト (電子メール 300 通) について学習を行い, c_i を計算した結果が表 1 である. 本研究では以下, 共起として前後 2 単語幅の窓を用いた. 「は」「に」などの助詞に加え, 「する」「ます」なども結節度が高く, 機能語とみなされていることがわかる.

c_i は機能語 - 内容語の区別に連続的な指標を与えることができるが, 結節度にはさらに, 内容語について学習データに応じた適応的な重要度を与えられるという利点がある. 例えば, 「情報」という言葉の重要度は一般には低く, 情報科学者にとっては高いと考えられるが, 結節度を用いれば" 情報" は一般には様々な語と共起するが, 情報科学者の場合には特定の単語としか共起しないため[1], 前者に関しては低く, 後者に関しては高い重みを与えることができる.

実際には, c_i の値は 0 近傍に片寄り, そのままでは分布が扱いにくい. そのため, 次式 (4) により指数をとったもの C_i を扱い, これを以下結節度コストと呼ぶ.

$$C_i = \alpha e^{-\beta c_i} \qquad (4)$$

$(\alpha, \beta$ は定数$)$

3.2 頻度ファクタ

3.1 で単語のカテゴリに関する連続量を定義したが, 実際はこのような基本コストを持った語の頻度が高いほど学習テキストにおける重要度が高いと考えることができる.

しかし, 重要度は単純に頻度に比例するのではなく, むしろ機能語については何回聞いても重要度は上がらないと考えた方が自然である. そこで, 繰り返し聞くことによる重みの上がり幅は結節度の逆数に比例すると考え, 次のように頻度ファクタ f を定義する.

$$f_i = k \cdot \frac{1}{c_i} \cdot \log n_i \qquad (5)$$

$(n_i : w_i$ の生起回数, $k :$ 定数$)$

これを用いて, 単語の重み W_i は次のように定義される:

$$W_i = f_i \cdot C_i \qquad (6)$$

同じ学習データに対して, 提案する手法および tf·idf による重み付けの上位 20 語と下位 20 語の例を表 2 に示した.

4 実験

われわれは文書を一見してそれが自分によって重要であるかそうでないかを判断することができるが, これは含まれている単語についての判断がその大きな割合を占めていると考えることができる.

そこで, 文書の重要度をそこに現れている単語の重要度の平均 (相加平均) だと見なして, 電子メールの重要度分類実験を行った. 加来田 [4] で用いられたメールについて, 同一組織内での無作為に抽出した電子メール 300 通を学習データとし, 算出した順位と人による順位を比較すると図 1 のようになった. 横軸が学習による順位, 縦軸が人による順位である. スピアマンの順位相関係数 r_s を計算すると, $r_s = 0.3127$ と比較的よい相関を示した.

この中には通常の私的メールだけでなく, SPAM[2]的な多数への情報メールも含まれてい

[1] 充分な量のテキストを与えた場合, これは正しくない. この議論については, 5章を参照されたい.

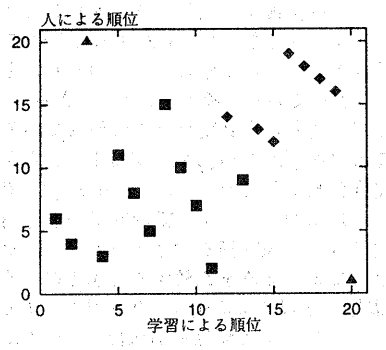
[2] 不特定多数への広告を主目的としたメールの総称.

	上位語			下位語	
順位	tf·idf	W_i	順位	tf·idf	W_i
1	教官	副	1	学位	書く
2	指導	主	2	確定	など
3	の	圭	3	各社	る
4	副	税	4	垣内	また
5	ご	ひひ	5	概略	ま
6	担当	生駒	6	概念	関
7	する	奈良県	7	外部	へ
8	法	ライティング	8	外周	行く
9	宏	利用者	9	外車	その
10	を	竹村	10	外国人	もの
11	だ	曼陀羅	11	界	見る
12	と	越す	12	海	より
13	が	知識工学	13	怪しい	出る
14	に	氏名	14	回線	下さる
15	受講	rapid	15	回数	中
16	ます	遺伝	16	回す	なら
17	は	削除	17	解決	後
18	英語	evolution	18	会社名	や
19	学生	田所	19	会議室	でも
20	ない	亮	20	雅人	ね

表 2: 単語重要度の学習例

る．いくつかの典型的なSPAMメールについて，単語の重み成長を負にするようにして学習を行った結果，図2のようになり，$r_s = 0.4722$ と精度が向上した．パーザが非常に短いメールに対して精度が悪く，図中▲のような外れ値が出ているため，これを除いてみると相関係数はそれぞれ $r_s = 0.7544$, $r_s = 0.7699$ とかなり高い相関[3] が得られることがわかった．

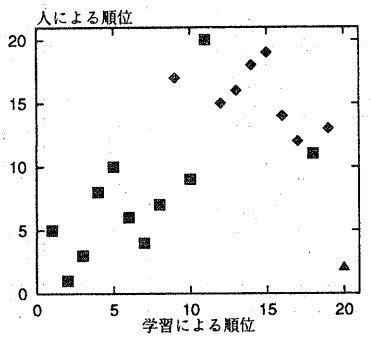
図 1: 順位付け結果



特に，SPAMメール (◆) をほぼ確実に下位順位に分類することができ，重みの絶対値を見ると負の値となって，明らかな分離を示した．なお，SPAMの中でも，個人の興味に一致する語句が含まれていた場合には値はそれほど低くならない．これは指向性を反映しているといえる．

[3]加来田 [4] での個人の判断による順位付けは絶対的な値とはいえず，揺れがあると考えられるため，相関は充分に高いといえる．

図 2: 順位付け結果 (負例を加えた場合)



これにより，これまで行われていなかった，内容によるメールのフィルタリングが行えることが示唆された．

5 まとめと精密化

本研究では単語の言語的な共起文脈に着目し，頻度と組み合わせた重みづけを行う指標を提案した．

このモデルは tf·idf と異なり，テキストが分割されていない場合でも学習が可能であり，また共起文脈を考慮することで個人毎に異なる単語の重み付けを行うことができる．個人の読むテキストはほぼその興味に沿っていると考えられるため，個人が読むメールなどのテキストに対して自動学習するようにすることで，指向性に応じた重み付けのデータベースが得られると思われる．

しかしながらモデルはまだ粗いものであり，例えば結節度の定義式 (3) では充分なテキストを学習させた場合，共起する単語数は特定の文脈で使う人かそうでないかに関わらず接近していくと考えることができる．共起した単語の異なり数だけでなく，その回数を記録し頻度を確率分布とみなせば，そのエントロピー＝(負の) 情報量を結節度の指標とすることができる．今後このような精密化を行うとともに，実際の応用としていかに組み込むかが課題になると思われる．

参考文献

[1] P. イングヴェルセン．情報検索 – 認知的アプローチ．トッパン，1995.

[2] Gerald Salton, *Automatic Text Processing*, Addison-Wesley, 1989.
[3] 伊藤哲郎. 情報検索. 昭晃堂, 1986.
[4] 加来田裕和, 角隆一. 情報取得のための優先順位判定要素と順位判定モデル. 情報処理学会研究会報告　知能と複雑系, ICS113-14, July, 1998.
[5] 原田昌紀. サーチエンジン徹底活用術. オーム社, 1997.