

概念ベースの情報検索への適用 —概念ベースを用いた検索の特性評価—

熊本 睦 島田 茂夫 加藤 恒昭

NTT コミュニケーション科学研究所

〒619-0237 京都府相楽郡精華町光台 2-4

E-mail: kumamoto, shimada, kato@cslab.kecl.ntt.co.jp

あらまし

文書の類似検索では、文書における単語の出現頻度など統計情報を利用し、検索要求と文書の類似性を判断するのが一般的である。本稿では辞書の語義文や文書群の単語の共起を元に自動作成した概念ベース(辞書ベース、コーパスベース)を利用した方法、および、従来手法(tf・idf方式)の類似判別能力を調べた。tf・idf方式は検索要求中の単語が文書に含まれる場合は有効であるが、そうでない場合は無力であった。一方、概念ベースを用いると、単語が文書に含まれない場合でも適切な類似判別能力を持つことが確認された。さらに、両者を組合せることで、より高い類似判別能力が得られることが分かった。

キーワード 情報検索、概念ベース、ベクトル空間モデル、類似検索

An Application of Concept Bases to Information Retrieval — An Evaluation on Characteristics of Information Retrieval using Concept Bases —

Mutsumi Kumamoto Shigeo Shimada Tsuneaki Kato

NTT Communication Science Laboratories

2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237

E-mail: kumamoto, shimada, kato@cslab.kecl.ntt.co.jp

Abstract

In this paper, we evaluate the abilities of similar information retrieval(IR) methods: (1)a conventional information retrieval method (the tf・idf method), and (2)an information retrieval method based on concept bases (knowledge of word meanings). The tf・idf method is very effective when the relevant documents include words in a query, but is ineffective when the relevant documents do not include words in a query. The IR method based on concept bases is effective when the relevant documents do not include words in a query. The combination of both methods is even more effective.

key words Information Retrieval, Concept Base, Vector Space Model, Similarity

1 はじめに

近年のインターネットの普及や電子化文書の増加に伴い、大量の文書から必要とする情報をすばやく取り出す必要性が高まっており、情報を検索要求と似ている順にランクづけする類似検索技術が重要になってきている。

類似検索技術としては、ベクトル空間モデル[1]が一般的に使われている。ベクトル空間モデルは、文書(あるいは検索要求)を n 次元のベクトル空間上の点にマッピングし、それらの間の距離の大小により文書同士(あるいは検索要求と文書)の類似性を計るものである。

古典的な手法として、ベクトル空間の軸とその軸の成分の値に、それぞれ、文書に出現する単語、その単語の出現頻度から得られる統計情報(tf・idf)を用いることが多い。しかし、このtf・idfに基づくベクトル空間モデルには、単語の意味が考慮されていないという問題がある。

我々は、想起型情報検索方式¹ [2]において、単語の意味知識である概念ベース(単語を他の単語との意味関係で表したベクトル)を用いて文書(あるいは検索要求)のベクトルを定義することにより、類似検索を行ない、この問題に対処している。具体的には、概念ベースは2種類用意している。ひとつは国語辞書の各見出しの語義文に含まれる単語頻度を元に作成した概念ベースであり[3]、もうひとつは、文書中の単語の共起頻度を元に作成した概念ベースである[4]。

本稿では、従来のtf・idfに基づく方式(以下、tf・idf方式)と上記2種類の概念ベースに基づく方式(以下、概念ベース方式)の類似判別能力を評価しその特性を考察する。tf・idf方式と概念ベース方式はそれぞれ異なる特性を持ち、これらを組合せることにより類似判別能力が向上することを述べる。また、評価方法として、見出しと本文が対になった文書群を用いて、見出しを検索要求とした場合に対応する本文が現れる順位を用いる方法を提案する。

¹ 想起型情報検索方式を適用した検索システムの研究開発はIPA 創造的ソフトウェア育成事業による。

これは、従来の適合率と再現率による検索結果の適合性評価とは異なる新しい手法であり、正解の作成のコストがかからないという特徴を持つ。以下、2節では、類似検索の各手法について、3節では、評価方法について、4節では、評価結果と考察について述べる。

2 類似検索方式

2.1 tf・idfに基づく類似検索

Salton[1]は文書 D_i を次のベクトル $\mathbf{d}_i$ で表現した。

$$\mathbf{d}_i = (d_{i1}, d_{i2}, \dots, d_{in})$$

ここで、 d_{ij} は文書 D_i に対する単語 W_j の重みで、次式で計算される。

$$d_{ij} = tf_{ij} \log(N/df_j)$$

tf_{ij} は文書 D_i 中の単語 W_j の出現頻度、 N は全文書数、 df_j は単語 W_j の出現文書数である。検索要求のベクトルも文書ベクトルと同様に定義される。ベクトルの長さはすべて1に正規化され、類似度はベクトルの内積で計算される。この方法では、別の表記の単語は違うものとして扱われるため、単語の意味が考慮されず、同義語のような同じ意味の単語も別々の単語として判断される。

2.2 概念ベースに基づく類似検索

我々は、単語同士の類似性を文書の類似性に反映させるために、単語に関してもベクトルで表現している。具体的には、単語 W_i を次のベクトル $\mathbf{w}_i$ で表現する。

$$\mathbf{w}_i = (w_{i1}, w_{i2}, \dots, w_{in})$$

ここで、 w_{ij} は単語 W_j に対する重みであり、この表現は、単語を別の単語との意味関係で定義するものである。このようにして、作られた単語ベクトルの集合を概念ベースと呼ぶ。具体的には、次の2種類の概念ベースを用いている。

- 国語辞書に基づく概念ベース (辞書 CB)

国語辞書の見出し語を概念ベースの用語とする。見出し語の語義文に含まれる単語を属性とし、その頻度に基づき重みを決定する。重みは辞書の孫引き等を用いて類似検索の精度を向上させている。また、属性数は、日本語語彙大系 [5] の意味属性を用いて約 3000 まで圧縮している [3]。

- コーパスに基づく概念ベース (コーパス CB)

文書中に含まれる単語の内、記号、助詞などを除き概念ベースの用語とする。高頻度語を属性として、その語と共起する頻度に基づき重みを決定する [4]。

辞書 CB は単語の語義に基づいて、コーパス CB は単語の使われ方に基づいて単語間の意味の類似性を定義するものだといえる。

このようにして定義した単語ベクトルに基づき、文書については単語の羅列と捉え、文書 D_i のベクトル d_i を文書中の単語 W_j に対応するベクトルの総和により定義する。

$$d_i = \sum_{j=1}^m w_j$$

また、ベクトルの長さはすべて 1 に正規化し、文書や単語間の類似度はベクトルの内積で計算する。

3 類似判別能力の評価

3.1 評価方法

類似判別能力を評価する方法として、検索要求と正解の対を作成し、検索結果の適合率と再現率により評価する方法が考えられる。統計的に意味のある評価を行なうためには多くの検索要求と正解の対を作成する必要があるが、正解の作成に非常にコストがかかるため、多くの検索要求と正解の対を作成することは難しい。例えば、情報処理学会のテストコレクション BMIR-J2[7] でも 60 対と少数である。

そこで、次のような方法を提案する。すなわち、見出しと本文が対となった文書群を用意し、見出し

を検索要求とした時に、対応する本文が現れた順位により、各方式を評価する方法である。本方法は正解の作成のコストがかからず、また、正解が人によりばらつくという問題もないという特徴がある。見出しを本文に対する検索要求とする是非については、『見出しは本文全体の意味を表しているとは限らない』という指摘もある [6] が、例えば、我々が記事の内容を読むかどうかを判断する時に見出しを参考にしているという現実を考慮すると、見出しは本文を特徴的に示している場合が多いと考え、是とした。

3.2 評価結果と考察

3.2.1 各方式単独の場合

文書集合として、情報処理学会のテストコレクション BMIR-J2²の記事集合を用いた。本コレクションは、毎日新聞の記事からなり、総数は 5,080 である。文書ベクトルの作成には、名詞や動詞等の主語や述語になりうる単語を用いている。辞書 CB 方式、コーパス CB 方式、tf・idf 方式の各方式の特性を見い出すために、2 者の順位比較を行なった結果を表 1 に示す。概念ベース方式と tf・idf 方式を比較すると、tf・idf 方式の順位が辞書 CB 方式の順位より高くなる場合が 43.2%、その逆が 9.5% ある。また、tf・idf 方式の順位がコーパス CB 方式の順位より高くなる場合が 31.1%、その逆が 14.1% ある。

次に、各方式がどのような場合に優れているかという特性を見るために、個々の事例を調査した。

例 1 辞書 CB 方式の順位が tf・idf 方式の順位より高い場合

見出し『トラック火事で住宅類焼 男性が巻き添え死—東大阪市 【大阪】 (記事 ID=00085170)』に対応する本文は、辞書 CB 方式で 1 位、tf・idf 方式で 17 位で検索された。tf・idf

²(社) 情報処理学会・データベースシステム研究会が、新情報処理開発機構との共同作業により、毎日新聞 CD-ROM '94 データ版を基に構築した情報検索システム評価用テストコレクション BMIR-J2 を利用。

表 1: 各方式の順位の比較結果

数値は、X の順位が Y の順位より高いものの件数。

X \ Y	辞書 CB	コーパス CB	tf・idf
辞書 CB	-	947(18.6%)	484(9.5%)
コーパス CB	1988(39.1%)	-	718(14.1%)
tf・idf	2195(43.2%)	1581(31.1%)	-

方式での順位が下がったのは、見出しにある「火事」「類焼」という単語が対応する本文中になかったためと考えられる。一方、辞書 CB 方式では、本文中にある「全焼」「出火」等が見出しにある「火事」「類焼」と類似度が高いと判断できたので、上位となったと考えられる。ところで、この見出しに対して tf・idf 方式で第 1 位として検索されたのは『住宅建設費を 3 割削減 「ハウス・ジャパン」構想推進——通産省、2000 年めどに (記事 ID=00751760)』の本文であった。本文中の「住宅」の tf・idf 値が高く、そのために主題が異なっているにもかかわらず順位が高くなってしまったと考えられる。

例 2 コーパス CB 方式の順位が tf・idf 方式の順位より高い場合

見出し『[みんなの広場] 原発推進、自信を持って=会社員・田中××47 (記事 ID=00831250)』に対応する本文は、コーパス CB 方式で 1 位、tf・idf 方式で 63 位で検索された。tf・idf 方式での順位が下がったのは、対応する本文中に「原発」という単語がなかったためと考えられる。一方、コーパス CB 方式では、本文中にある「原子力発電所」「電力」「発電所」等が見出しにある「原発」と類似度が高いと判断できたので、上位となったと考えられる。ところで、この見出しに対して tf・idf 方式の第 1 位として検索されたのは『××販売会社社長の田中××さんが自宅前で射殺される——京都 【大阪】 (記事 ID=00267680)』の本文であった。本文中の「田中」、「会社」の tf・idf 値が高く、そのために主題が大きく異なっているにもかかわらず順位が高くなってしまったと考えられる。

でも順位が高くなってしまったと考えられる。

例 3 tf・idf 方式の順位が辞書 CB 方式の順位より高い場合

見出し『「ディスカバリー」が「ミール」と交信 ロシア人飛行士搭乗し 【大阪】 (記事 ID=00096280)』に対応する本文は、tf・idf 方式で 1 位、辞書 CB 方式で 108 位で検索された。今回使用した辞書 CB は国語辞書から作成したために、「ディスカバリー」「ミール」「ロシア」等の固有名詞が含まれておらず、辞書 CB 方式で順位が下がったと考えられる。

例 4 tf・idf 方式の順位がコーパス CB 方式の順位より高い場合

見出し『[みんなの広場] 医師射殺事件の底にあるもの=無職・××74 (記事 ID=00883500)』に対応する本文は、tf・idf 方式で 1 位、コーパス CB 方式で 108 位で検索された。ところで、この見出しに対して、コーパス CB 方式で第 1 位として検索されたのは『[特集] 1994 年・今年の重大ニュース 広がる「銃汚染」、企業テロも相次ぐ (記事 ID=01008300)』の本文であった。本文中の一部には、同事件の事柄も含まれており、そのために類似度が高くなってしまったと考えられる。

以上の結果をまとめると表 2 となる。なお、辞書 CB 方式とコーパス CB 方式との比較に関しては、辞書 CB の語彙数の問題もあり、他の特性も見い出せなかったため省略した。

表 2: 各方式の特性

方式	特性
辞書 CB	本文に見出し中の単語が含まれていなくても、同義語や類義語(「全焼」に対する「類焼」等)が含まれていれば効果がある。今回、辞書 CB は国語辞書から作成したために固有名詞等が含まれていなかったこともあり、他の方式より結果が悪くなった。これについては、様々な辞書や事典を用いて概念ベースを構築することにより改善可能と考えられる。
コーパス CB	本文に見出し中の単語が含まれていなくても、同義語や連想語(「原発」に対する「原子力」「発電所」「電力」等)が含まれていれば効果がある。様々な単語が連想されるので、tf・idf 方式のようなシャープさには欠けるが、大きく内容がはずれるものを検索することは少ない。
tf・idf	本文に見出し中の単語が含まれる場合には非常に効果的である。しかし、同じ単語が含まれていない場合は無力となり、その結果、主題が大きく異なるものが検索される場合がある。

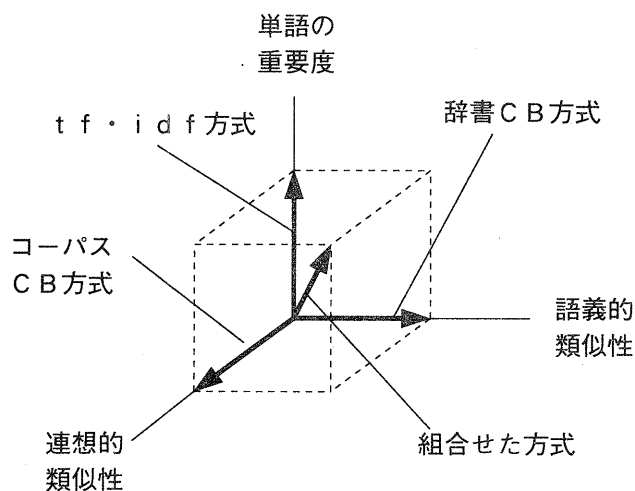


図 1: 類似判別能力の組合せのイメージ

3.2.2 概念ベース方式と tf・idf 方式を組合せた場合

上で述べたように、各方式は異なる特性を持つので、方式を組合せることで、互いの特性を活かせる可能性が考えられる。各方式の特性とその組合せのイメージを図 1 に示す。tf・idf 方式は単語の重要度という観点で、辞書 CB 方式は語義的類似性とい

う観点で、コーパス CB 方式は連想的類似性という観点で類似を判断していると考えられ、各方式をどのような割合で用いるかにより、組合せた方式の能力や特性が決まると考える。

今回は、辞書 CB 方式と tf・idf 方式、コーパス CB 方式と tf・idf 方式の組合せを試した。具体的には、両者を組合せた次式の類似度 S を用いる。

表 3: 概念ベース方式と tf・idf 方式を組合せた場合の順位の比較結果

数値は、X の順位が Y の順位より高いものの件数。

X \ Y	辞書 CB+tf・idf	コーパス CB+tf・idf	tf・idf
辞書 CB+tf・idf	-	688(13.5%)	678(13.3%)
コーパス CB+tf・idf	827(16.3%)	-	789(15.5%)
tf・idf	689(13.5%)	668(13.1%)	-

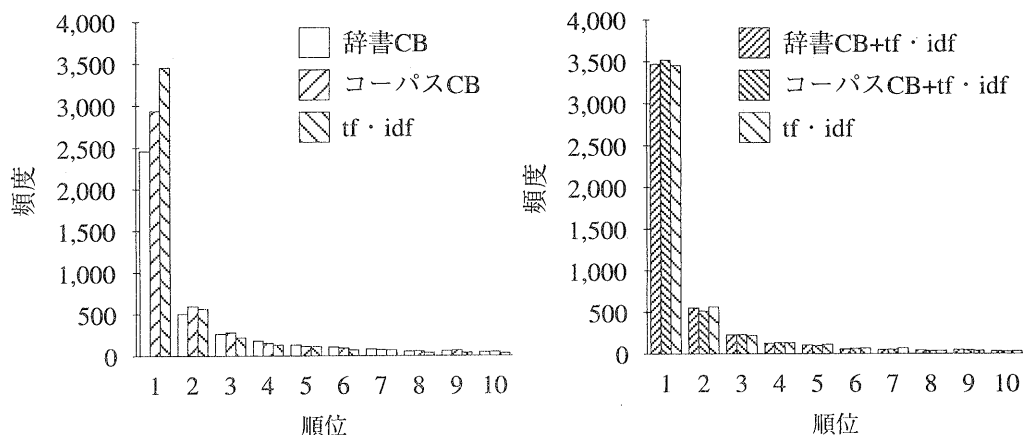


図 2: 方式の組合せによる改善の効果

$$S = \alpha S_c + \beta S_t$$

ここで、 S_c は概念ベース方式 (辞書 CB 方式あるいはコーパス CB 方式) による類似度、 S_t は tf・idf 方式による類似度である。例として、 $\alpha = 1, \beta = 1$ とした場合の結果を表 3 に示す。辞書 CB 方式に tf・idf 方式を組合せると tf・idf 方式単独と同じくらい、また、コーパス CB 方式に tf・idf 方式を組合せると tf・idf 方式単独よりよくなることが分かった。どのような組合せをすればよりよい結果が得られるかについては、今後の課題である。

次に、記事集合全体で見た時の方式の組合せの効果を見るために、横軸に対応する記事本文が検索された順位をとり、縦軸にその順位になったものの頻度をとったものを図 2 に示す。順位が上位 10 位までのものを描いている。この図からも複数の方式を組合せることで、単独の方式よりものより組合せ方式

の方がよくなっていることが分かる。

さらに、組合せた方式を使うことで、組合せに使った各方式の特性が活かされているかどうかを前述した各例を対象にして調べた。

例 1 単独方式では、記事 ID=00085170 の本文は、辞書 CB 方式で 1 位、tf・idf 方式では 17 位で検索されたが、これらを組合せた辞書 CB+tf・idf 方式では 1 位で検索された。辞書 CB 方式の特性が活かされ、tf・idf 方式で生じていた主題が異なるものが上位に出てくるとい問題は解消されていることが分かった。

例 2 単独方式では、記事 ID=00831250 の本文は、コーパス CB 方式で 1 位、tf・idf 方式では 63 位で検索されたが、これらを組合せたコーパス CB+tf・idf 方式では 8 位で検索された。この場合のコーパス CB+tf・idf 方式の第 1 位で検

表 4: 見出しや本文中の単語の違いによる各方式の良否

		見出し中の単語が本文中にある場合 に対応する本文が上位になる	見出し中の単語が本文になく、 関連語がある場合に対応する本文 が上位になる
単独	辞書 CB	△	○(同義語がある場合)
	コーパス CB	△	○(連想語がある場合)
	tf・idf	○	×
組合せ	辞書 CB+tf・idf	○	○(同義語がある場合)
	コーパス CB+tf・idf	○	○(連想語がある場合)

索されたのは『[みんなの広場]「原発」もっと学習してほしい=通信高校生・×× 17 (記事ID=00851660)』の本文であった。コーパス CB 方式から見れば順位は若干下がったが、コーパス CB 方式の特性が活かされ、tf・idf 方式で生じていた主題が大きく異なるものが上位に出てくるとい問題は解消されていることが分かった。

例 3 単独方式では、記事 ID=00096280 の本文は、辞書 CB 方式で 108 位、tf・idf 方式で 1 位で検索されたが、これらを組合せた辞書 CB+tf・idf 方式では 1 位で検索された。tf・idf 方式の特性が活かされ、辞書 CB 方式で生じていた固有名詞の問題は解消されていることが分かった。

例 4 単独方式では、記事 ID=00883500 の本文は、記事 CB 方式で 108 位、tf・idf 方式で 1 位で検索されたが、これらを組合せたコーパス CB+tf・idf 方式では 9 位で検索された。この場合のコーパス CB+tf・idf 方式の第 1 位で検索されたのは、コーパス CB 方式の第 1 位と同じ『[特集] 1994 年・今年の重大ニュース 広がる「銃汚染」、企業テロも相次ぐ(記事ID=01008300)』の本文であった。コーパス CB 方式と比較すると順位は上がっており、コーパス CB 方式の特性をそねることなく、tf・idf 方式の特性が活かされていることが分かった。

以上のように、方式を組合せることで各方式の特性が活かされていることが分かった。見出しや本文中の単語の違いによる各方式の良否をまとめると表 4 のようになる。

4 おわりに

意味知識である概念ベースに基づく類似検索方式と tf・idf 方式の類似判別能力について評価し、特性を考察した。tf・idf 方式は、文書中に単語が存在する場合に非常に効果的であり、概念ベース方式は、文書中に単語が存在しなくても同義語や連想語がある場合に効果的であった。各方式の類似度を合成する方法により方式を組合せることで、各方式の特性を活かしてさらに効果を上げることができた。

また、評価方法として、見出しを検索要求とし対応する本文が検索される順位による評価方法を提案した。ただし、今回の評価で用いた文書集合は新聞記事であったので、記事中の単語が見出しに使われることが多く、tf・idf 方式に有利なものであったと思われる。この点に関しては、他の種類の文書集合で評価することも検討する必要がある。

また、今回は評価の観点に入れなかったが、文書集合の規模や更新頻度等の問題も考慮する必要がある。例えば、インターネット上の文書のような大規模かつ変更の多い領域を考えると、tf・idf 方式やコーパス CB 方式は、文書集合が変わると統計情報をとり直す必要があり頻繁に更新する必要がある。辞書 CB 方式に関しては、そのような問題がない

め有望である。

我々は、概念ベースの技術を核とし、検索者の検索意図が明確でない場合でも、検索者の想起、発想を支援しながら情報検索を進めることができることを目指した想起型情報検索方式の研究を行なっている。その中で、新たな連想や想起を促すような結果を得るために単語の意味を考慮した類似検索を用いている。各類似検索方式には、それぞれ特徴があるので、検索や発想支援といった目的に応じて、それらをどのように使い分けたり、組合せたりするかが今後の課題である。

参考文献

- [1] Salton, G. and Allen, J.: Text Retrieval Using the Vector Processing Model, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.
- [2] 飯田敏幸, 松澤和光, 池上徹彦, 石野福弥, 今井賢一: 想起型情報検索システムについて, 情処研究報告, 98-OS-77/98-DPS-87, pp.19-24, 1998.
- [3] 笠原 要, 松澤和光, 石川 勉: 国語辞書を利用した日常語の類似性判別, 情処論文誌, Vol. 38, No. 7, pp.1272-1284, 1997.
- [4] Schütze, H. and Pedersen, J. O.: Information Retrieval Based on Word Senses, 4th Annual Symposium on Document Analysis and Information Retrieval, pp.161-176, 1995.
- [5] 池原 悟 他: 日本語語彙大系 1 意味体系, 岩波書店, 845p., 1998.
- [6] 新世代通信網実験協議会, スタンフォード日本センター: 次世代情報網の利用研究と日本型連想検索システムの構築, メメックス研究会, 1998.
- [7] 木谷他: 日本語情報検索システム評価用テストコレクション BMIR-J2, 情報処理学会研究報告, DBS-114-3, pp.15-22, 1998.