社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－NL－171 (4)
2006／1／12

概念ベースと関連度計算を用いた記事関連度計算方式

倉田 篤史† 渡部 広一† 河岡 司†

† 同志社大学大学院 工学研究科 知識工学専攻

〒610-0321 京都府京田辺市多々羅都谷 1-3
E-mail: {akurata,watabe,kawaoka}@indy.doshisha.ac.jp

本稿では，情報検索や文書分類に応用できる文書間の意味的近さの定量化手法を提案する．この提案手法は，単語の表記情報のみで定量化するのではなく，単語の意味を理解することによるものである．単語の意味属性を登録した概念ベースを用いることによって，文書間の関連の強さを定量化することができる．そのため，単語の表記的な揺らぎに影響されることがない手法である．この提案手法を NTCIR2 によって，ベクトル空間モデルと比較評価し，提案手法では数%の精度向上が得られることを示した．

キーワード：概念ベース，関連度計算，情報検索，文書分類

Calculation of the Semantic Closeness between Documents using the Concept-Base and the Degree of Association between Concepts

KURATA Atsushi† WATABE Hirokazu† KAWAOKA Tsukasa†

† Department of Knowledge Engineering and Computer Sciences, Doshisha University

1-3 Miyakodani, Tatara, Kyotanabe, Kyoto 610-0321, Japan
E-mail: {akurata,watabe,kawaoka}@indy.doshisha.ac.jp

This paper proposes the quantification technique of the semantic closeness between documents which can be applied to information retrieval and the documents classification. This proposed technique is not the one by using the notation of words but the one by understanding the meaning of words. The concept-base and calculation of the degree of association enable to quantify the degree of the relation between documents. Therefore this technique isn't influenced by the notation of words. This proposed technique is evaluated by NTCIR2 to compare vector space model and we showed that the proposed technique has several percents accuracy more than VSM.

Key words: concept-base, degree of association, information retrieval, documents classification

1. はじめに

近年コンピュータに関する技術の発達やネットワークの拡大により，ユーザが入手可能な情報は膨大なものとなってきている．それらの情報の中には，ユーザにとって必要となる適切な情報もあれば，必要のない不適切な情報もある．そのため，あらゆる情報の中から必要な情報だけを抽出しなければならない．そこで，ユーザが効率良く情報を入手するためには，情報検索技術や文書の要約技術，分類技術が必要となる．しかし，現在の情報や文書を整理する技術というものは，その中に出現する単語の表記情報のみを手がかりとしている．これでは，単語またはその集合である文書の表す意味を捉えることはできない．意味が同じ，または意味が似通っている単語が存在するにもかかわらず，表記が異なるために検索・分類などの結果に反映されないことになる．これでは，ユーザが求める情報を検索したり，意味的に近い文書をまとめて分類したりする場合に活かされない．

本研究では，単語（概念）の意味特徴を定義した概念ベース [1] を用いることによって，記事間の意味的近さを定量化する手法を提案し，その手法の有効性をNTCIR2によって検証する．

2. 関連技術

本章では，情報検索の分野で用いられる一般的な技術について述べる．

2.1 文章の表現

自然文の文章Sは，文章内の語k_iとその重みw_iの対の集合として，以下の式により定義される．

$$S=\{(k_1,w_1),(k_2,w_2),\cdots,(k_L,w_L)\} \tag{1}$$

Lは，文章S内の語数である．ただし，$\{k_i \mid 1\le i\le L\}$に重複はないものとし，特に関連度計算に用いる場合は，$\sum_{i=1}^{L} w_i=1$とする．

また，本稿で述べる文章とは，文の集合であり，ある特定の情報を有するものとする．

2.2 TF・IDFによる重み付け

TF・IDFによる重み付け [2] とは，語の出現頻度と特定性を表す尺度に基づいた重み付け手法である．文書dにおける語tの重みは，以下の式で定義される．

$$w_t^d=tf_d(t)\cdot idf(t) \tag{2}$$

$tf_d(t)$とは，文書d内での語tの出現頻度$tf(t,d)$と文書d内の全ての語数から求められる相対頻度である．以下の式で定義される．

$$tf_d(t)=\frac{tf(t,d)}{\sum_{s\in d} tf(s,d)} \tag{3}$$

また，$idf(t)$は語tが出現する文書数によって決まり，以下の式で定義される．

$$idf(t)=\log\left(\frac{N}{df(t)}\right) \tag{4}$$

ここで，Nは検索対象群での全文書数であり，$df(t)$とは，検索対象群で語tが出現する文書数である．以下，全ての手法において，この重み付け手法を用いる．

2.3 表記一致方式

本節では，単純な単語の表記のみを活用した文章間の類似度計算方式について述べる．

この方式は，単純に単語の表記が一致した割合から類似度を求めるものであり，以下の式で定義する．

$$Score(X,Y)=\frac{m/x+m/y}{2} \tag{5}$$

ここで，xは文章Xの単語数，yは文章Yの単語数，mは文章XとYの両方に出現する単語数である．

また，TF・IDFによる重み付けを行い，その重みの割合から，類似度を求める方法を重み付き表記一致方式と呼ぶ．

2.4 ベクトル空間モデル

ベクトル空間モデル [3] は，ユーザから入力された検索質問文と文書をそれぞれベクトルで表現し，ベクトルの類似度により情報ランキングを行う検索モデルである．文書は，以下の式で定義される．

$$doc=\{(d_1,w_1),(d_2,w_2),\cdots,(d_a,w_a)\} \tag{6}$$

ある文書 doc は，語 d_i と重み w_i の集合であり，a 個の語からなる．検索対象群に含まれるすべての語の数を M とすると，検索対象群の文書は，すべて M 次元の重みベクトルとして，以下の式で定義される．

$$doc = \{w'_1, w'_2, \cdots, w'_M\} \tag{7}$$

同様に，ユーザから入力された検索質問文も以下のように定義される．

$$Q = \{(k_1, v_1), (k_2, v_2), \cdots (k_b, v_b)\} \tag{8}$$
$$= \{v'_1, v'_2, \cdots, v'_M\} \tag{9}$$

k_i は，検索質問文中のキーワードであり，文書 doc と同様に語と重みの集合で表される．しかし，検索質問文はユーザによる手入力のため，検索質問文中にあまり多くの語が出現することは考えにくい．このとき，検索質問文から得られるベクトル Q の要素はほとんど 0 である．

2 つのベクトルである，文書 doc とユーザからの要求 Q との類似度 $Vector(Q,doc)$ は，以下の式で定義される．

$$Vector(Q, doc) = \cos\theta \tag{10}$$
$$= \frac{Q \cdot doc}{|Q||doc|} \tag{11}$$

θ は 2 つのベクトル doc と Q のなす角度で，類似度はその 2 つのベクトルのなす角度の余弦値である．

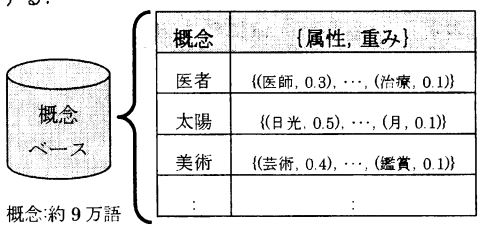
3. 関連度計算を用いた手法

本章では，概念ベースを利用して，関連度計算を行うことによる文章間の意味的近さの定量化手法について述べる．

3.1 概念ベース

概念ベースとは，ある概念の意味特徴を表す属性とその属性の概念における重要度を示す重みとの対の集合からなる知識ベースである（Fig. 1）．

概念ベースは，国語辞書などから自動構築され，現在約 9 万語の概念が収録されている．また，1 つの概念当たり，平均 30 個の属性が存在する．



概念	{属性, 重み}
医者	{(医師, 0.3), ···, (治療, 0.1)}
太陽	{(日光, 0.5), ···, (月, 0.1)}
美術	{(芸術, 0.4), ···, (鑑賞, 0.1)}
：	：

Fig. 1 概念ベース

ある概念の属性をまた概念として，その属性を展開することもできる．ある概念の属性のことを一次属性と呼び，一次属性の属性のことを二次属性と呼ぶ（Table 1）．つまり，ある概念はその属性を n 次元展開で表現することができる．ただし，以降に述べる関連度計算手法では，二次属性までを使用するアルゴリズムとなっている．

Table 1 概念「医者」の一次属性と二次属性（重みは省略）

概念	属性				
医者	医師	患者	··	治す	一次属性
	医者	病人	··	治療	
	診察	包帯	··	医療	二次属性
	病院	看病	··	癒す	
	：	：	··	：	
	保健	治療	··	病気	

3.2 重み比率付き関連度計算アルゴリズム

関連度計算とは，概念ベースを利用して，概念と概念の関連の強さを定量化する手法である．本研究では，重み比率付き関連度計算 [4] を利用している．本稿では，その計算方式のアルゴリズムを簡単に説明する．

3.2.1 重み比率付き一致度

関連度は，一次属性の一致度から計算される．よって，関連度計算を説明する前に一致度計算について説明する．

以下のような概念 A，B があるとする．

$$A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_M, w_M)\} \tag{12}$$
$$B = \{(b_1, v_1), (b_2, v_2), \cdots, (b_N, v_N)\} \tag{13}$$

M，Nは，それぞれ概念A，Bの属性数である．また，$a_i, w_i (1 \le i \le M)$は，概念Aの属性とその重みである．同様に概念Bも$b_j, v_j (1 \le j \le N)$で表される．二次属性についても，以下のように定義される．

$$a_i = \{(a_{i1}, w_{i1}), (a_{i2}, w_{i2}), \cdots, (a_{im_i}, w_{im_i})\} \quad (14)$$

$$b_j = \{(b_{j1}, v_{j1}), (b_{j2}, v_{j2}), \cdots, (b_{jn_j}, v_{jn_j})\} \quad (15)$$

このとき，一次属性a_iとb_jの重み比率付き一致度MatchW(a_i, b_j)は以下のように定義される．

$$\mathrm{MatchW}(a_i, b_j) = \sum_{a_{is} = b_{jt}} \min(w_{is}, v_{jt}) \quad (16)$$

3.2.2 重み比率付き関連度

前節で求めた一致度を一次属性全ての組合せに対して行い，一致度が大きいものから順に対応を決めていく．式（11）の概念Aに対して，一致度が最大となる組合せになるように，概念Bの属性を並べ替えたものを以下に示す．

$$B = \{(b_{x1}, v_{x1}), (b_{x2}, v_{x2}), \cdots (b_{xN}, v_{xN})\} \quad (17)$$

よって，これらの概念A，Bの重み比率付き関連度は次のようになる．

$$\mathrm{ChainW}(A, B) = \sum_i \mathrm{MatchW}(a_i, b_{xi}) \times (w_i + v_{xi})/2 \times (\min(w_i, v_{xi})/\max(w_i, v_{xi})) \quad (18)$$

3.3 記事の概念化

前節で説明した関連度計算方式を記事に対して行うことを考える．記事を概念化する流れは，以下のようになる．この方式を記事関連度と呼ぶ．

1. 記事から自立語を取得する
2. 自立語に重みを付与する
3. 自立語と重みの対の集合を概念とする

記事から自立語を取得するために，構文解析ソフト茶筌[5]を利用した．本研究では自立語として，「名詞」，「動詞」，「形容詞」，そして「未知語」を採用している．

概念ベース内での概念対概念の関連度計算と記事間の関連度計算を比較すると，概念と記事，一次属性と自立語，二次属性と自立語の一次属性がそれぞれ対応することになる（Table 2）．また，今回行った評価実験では，概念ベースに存在しない語も含まれる．そういった概念は，その概念自身のみを属性とする概念として扱う．

Table 2 記事と概念の対応

概念	記事 （自立語と重みの対の集合）
一次属性	自立語
二次属性	自立語の一次属性

4. 評価

本章では，関連度計算手法の有効性を検証するための評価方法について述べる．情報検索システムテストコレクション NTCIR2 [6] を用いて，提案手法の評価実験を行った．

4.1 評価方法

NTCIR2 は，学会発表論文の抄録など 736,166 件からなる文書データベースである．それに対する検索課題が 49 件用意されている．

今回の評価では，検索課題 49 件と 5,000 件の文書を使用し，評価実験を行った．検索課題は，検索要求説明を使用した．また，正解文書リストが存在し，各検索課題に対して，各文書が S（高適合），A（適合），B（部分的適合），C（不適合）の 4 段階の適合度が設定されている．S，A のみを正解文書とする評価方法をレベル 1 とし，S，A，B までを正解文書とする評価方法をレベル 2 とする．

記事関連度の計算方法は，各検索課題に対して，5,000 件の文書全てとの関連度を計算し，その値を各文書のスコアとした．このスコアをもとに評価プログラム（trec_eval[7]）を使用する．

情報検索の分野でよく用いられる評価尺度，再現率と精度[8]を用いる．

4.2 表記一致方式と記事関連度の評価

まず，表記一致方式と記事関連度方式の比較を示す（Fig. 2，Fig. 3）．この評価は，既に報告されているものを引用する [9]．ただし，評価方法は検索課題 10 件に対して，それらの課題のレベル 2 での正解文書 567 件での評価結果である．また，採用している自立語は，概念ベースに存在する語のみである．記事関連度方式の重みは，TF・IDF を使用している．

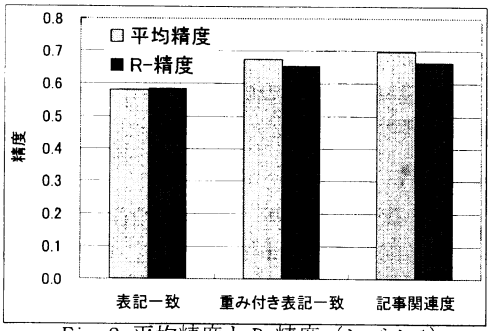


Fig. 2 平均精度と R-精度（レベル 1）

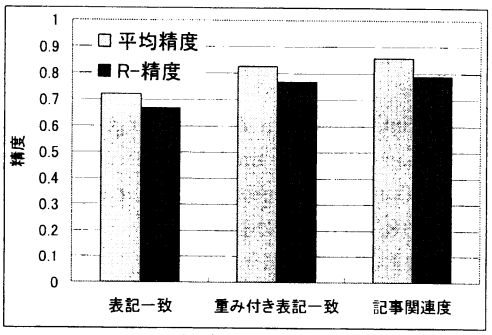


Fig. 3 平均精度と R-精度（レベル 2）

表記一致方式では，重み付きのほうが良い結果となっている．その重み付き表記一致方式と比べて，記事関連度はレベル 1 では平均精度が約 2%，R-精度が 1.3%，レベル 2 では平均精度が 3.5%，R-精度が約 2%の差が見られた．

平均精度とは，再現率が 0.0〜1.0 での間で 0.1 刻みでの 11 点における精度の平均値のことである．R-精度とは，正解文書 R 件あるとき，ランキングの上位 R 件での精度のことをいう．

4.3 ベクトル空間モデルと記事関連度の評価

記事関連度の有効性を示すために，同様のテストセットで重み付き表記一致とベクトル空間モデルの評価実験も行った．採用する自立語，重み付け（TF・IDF）は，全て同じである（概念ベースに存在しない語も使用）．ベクトル空間モデルと記事関連度の評価結果（11 点再現率・精度のグラフ）を次に示す（Fig. 4，Fig. 5）．

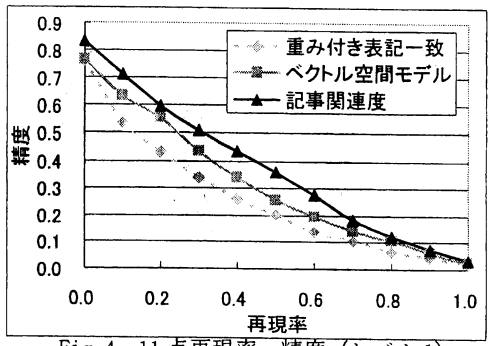


Fig. 4 11 点再現率・精度（レベル 1）

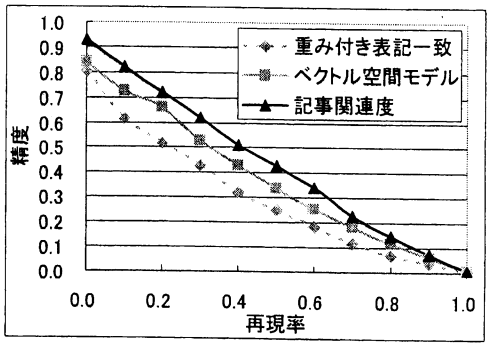


Fig. 5 11 点再現率・精度（レベル 2）

これらの結果より，記事関連度方式，ベクトル空間モデル，重み付き表記一致方式の順で精度が良いことが分かる．さらに，どの程度精度に差があるのかを示すために，平均精度と R-精度のグラフを載せる（Fig. 6，Fig. 7）．ベクトル空間モデルと記事関連度方式では，レベル 1 で，平均精度，R-精度ともに約 5.5%の差があり，レベル 2 で，平均精度 6.7%，R-精度 6.2%の差がある．

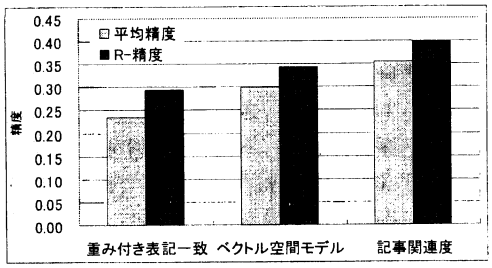


Fig. 6　平均精度とR-精度（レベル1）

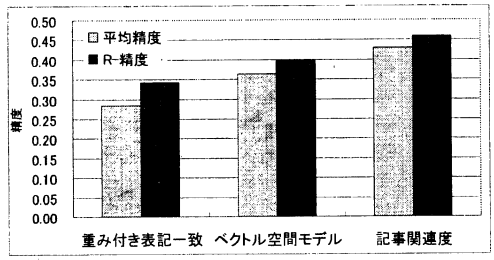


Fig. 7　平均精度とR-精度（レベル2）

5. 考察

NTCIR2を基に以前の報告[9]よりも多くの検索課題と文書を用いて評価を行い，表記一致方式とベクトル空間モデルとの比較により，記事関連度方式の有効性を示した．また，今回の実験では，概念ベースに存在しない語を使用した場合においても表記一致方式やベクトル空間モデルより良い結果となった．つまり，概念ベースを用いて，単語を表記のみで捉えるのではなく，その単語が表す意味を捉えた上での記事関連度方式が有効であると言える．

また，今回評価に用いたNTCIR2での文書は学会発表論文であったため，概念ベースに存在しない専門用語などが存在した．概念ベースの拡張や精錬によって，記事関連度方式の有効性はさらに大きくなるだろう．もしくは，記事関連度方式は概念ベースに存在する一般的な語で表現された文書空間においても，さらに有効な手段と言える．

6. おわりに

本研究では，単語の意味特徴を考慮した手法を提案した．今回扱ったテストセットは，学会論文の文書データベースであったため，専門用語などの概念ベースにない語が多く含まれていた．今後は，概念ベースの拡張・精錬によって，より精度が高い方法となることを期待する．

また，さらなる高速化を目指す．この精度を保ちつつ高速化できるアルゴリズムを考案する必要がある．

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクト「知能情報科学とその応用」における研究の一環として行った．

参考文献

1) 渡部広一，河岡司：「常識的判断のための概念間の関連度評価モデル」，自然言語処理，Vol.8, No.2, pp39-54, 2001.
2) Salton, G. and Buckley, C.: "Term-weighting approaches in automatic text retrieval", *Information Processing & Management*, Vol.41, No.4, pp.513-523, 1988.
3) Salton, G., Wong, A. and Yang, C. S.: "A vector space model for automatic indexing", *Communications of the ACM*, Vol.18, No.3, pp.613-620, 1975.
4) 井筒大志，渡部広一，河岡司：「概念ベースを用いた連想機能実現のための関連度計算方式」，情報科学技術フォーラムFIT2002, E-39, pp.159-160, 2002.
5) http://chasen.aist-nara.ac.jp/hiki/ChaSen/
6) http://research.nii.ac.jp/ntcweb/index-ja.html
7) ftp://ftp.cs.cornell.edu/pub/smart/
8) 徳永健伸：「情報検索と言語処理」，東京大学出版会，1999.
9) 若月紀之，渡部広一，河岡司：「概念ベースと関連度計算を用いた新聞記事の分類」，情報処理学会研究報告，2005-NL-165，pp.67-72, 2005.