2005－NL－165 (10)
2005/1/12

概念ベースと関連度計算を用いた新聞記事の分類

若月 紀之　渡部 広一　河岡 司
同志社大学大学院工学研究科

電子化された情報の溢れる現代社会において，効率よく目的の文書を探し出すことは重要である．そこで本稿では記事間の関連の強さを定量化する手法を提案する．この手法を利用することで記事を自動的に分類することができ，我々の情報活用に役立つと考えられる．本稿では概念ベースを用いることにより，概念の意味属性を利用した関連度計算が可能になることを述べる．概念ベースを利用することで，語の表記の一致だけでは見出せない概念間の関連性に基づいた定量化が可能になる．

Classification of Newspaper Articles using The Concept-Base and The Degree of Association between Concepts

Noriyuki Wakatsuki　Hirokazu Watabe　Tsukasa Kawaoka
Graduate School of Engineering, Doshisha University

In the modern society with which the information processed electronically overflows, it is important to discover the target document efficiently. Then, in this paper, the technique of quantifying the strength of the relation between reports is proposed. A report can be automatically classified according to using this technique, and it is thought that it is useful to our information practical use. In this paper, what the degree of relation between reports understands using the semantic attribute of a concept is described. By using Concept-base, the quantification based on the relevance between the concepts which cannot be found out only by coincidence of the notation of a word is attained.

１．はじめに

近年のコンピュータに関する技術は著しい発展を遂げており, ネットワークも飛躍的な拡大を見せている．IT という言葉を目にしない日はなく, デジタル家電, ユビキタスといった用語がメディアを飛び交う. ビジネス現場から家庭内までデジタル機器が深く浸透しつつある現代は高度情報化社会と言われる. 世の中の情報の多くは電子化され, 以前には考えられなかったような大量の情報に容易にアクセスできるようになっている.

このように現代社会には電子化情報が溢れているが, すべての文書に目を通すことは不可能である. そのため, いかに効率よく目的の情報に辿り着くかが重要になる. そこで人間の情報収集のサポートを可能な, 文書の検索や分類の技術の果たす役割は大きい.

本稿では, 2 つの記事の間の意味的な関連の強さを数値化する手法を提案する. これによりコンピュータによる記事の検索や分類が可能になると考えられる. ここでは, 概念ベース[1]と概念間の関連性を判断する関連度計算[2]を

利用して記事間の意味的関連性の強さの数値化を行う．概念ベースは，概念とその意味特徴を表す属性の集合による知識ベースである．

２．概念ベース

概念ベースは，複数の辞書等から機械的に構築された大規模で汎用的なデータベースである．本稿では，不適切なデータを削除し，必要なデータを追加する精練処理[3]を行った概念ベースを利用して研究を行った．

概念 A は，概念の意味を表す属性 a_i と，属性の重要性をあらわす重み w_i の対で表される．概念 A の属性数を N 個とすると，概念 A は以下のように表すことができる．

$$A=\{(a_1,w_1),(a_2,\ w_2),\cdots,(a_N,w_N)\}$$

ここで，属性 a_i を概念 A の一次属性と呼ぶ．

概念 A の一次属性もまた，属性 a_i も概念ベースに登録されている 1 つの概念である．従って，a_i からも同様に属性を導くことができる．a_i の属性 a_{ij} を概念 A の二次属性と呼ぶ．概念 A を二次属性まで展開した様子を図 1 に示す．

電車	電車,0.36	汽車,0.21	線路,0.10	…
	電車,0.36	汽車 0.42	線路,0.38	…
	汽車,0.21	電車,0.23	鉄道,0.25	…
	線路,0.10	列車,0.11	電車,0.18	…
	⋮	⋮	⋮	

二次属性

図 1：概念「電車」（二次属性まで展開）

本稿で実験に用いた概念ベースには，87,242 語の概念があり，総属性数は 2,539,582 個である．属性数は概念ごとに異なるが，1 概念あたりの平均属性数は約 29.1 個である．また，重みは情報量や概念間規則を用いて 0 から 1 の実数値で与えられている．

３．概念間の関連度計算

概念間の関連性の強さを表す数値を関連度という．関連度は，0 から 1 の間の実数値で表現される．関連度が 1 となるのは，二つの概念が同一の場合のみである．関連度の計算方法としては，概念の意味属性の一致度と重みを利用する意味関連度計算方式[2]や，表意文字である漢字の表記特徴を利用して関連性を評価する漢字関連度計算方式[4]などがある．本稿では，前者の意味関連度の計算方法について述べる．

3.1　一致度の計算方法

概念 A，B を，その一次属性を a_i，b_j，重みを u_i，v_j とし，属性がそれぞれ L 個，M 個($L \leqq M$)とすると

$$A=((a_1,u_1),(a_2,u_2),\cdots,(a_L,u_L))$$

$$B=((b_1,v_1),(b_2,v_2),\cdots,(b_M,v_M))$$

と表現する．概念 A, B の一致度 $MatchW(A,B)$ は以下のようになる．

$$MatchW(A,B)=\left(s_A/n_A+s_B/n_B\right)/2$$

$$s_A=\sum_{a_i=b_j}u_i \qquad s_B=\sum_{a_i=b_j}v_j$$

$$n_A=\sum_{i=1}^{L}u_i \qquad n_B=\sum_{j=1}^{M}v_j$$

重み付き一致度 $MatchW$ は概念 A から見たとき，概念 B の属性と一致した属性の重みの割合と，概念 B から見たときの概念 A の属性と一致した属性の重みの割合の平均を表している．

3.2　関連度の計算方法

意味関連度は，対象となる一次属性の組み合わせについて一致度を計算し，一次属性どうしの対応を決定することにより計算する．

概念 A，B のうち属性数の少ない概念を A($L \leqq M$)とし，概念 A の一次属性の並びを固定する．

$$A=((a_1,u_1),\ (a_2,u_2),\ \cdots,\ (a_L,u_L))$$

概念 B の各一次属性を対応する概念 A の各一次属性との一致度($MatchW$)の合計が最大になるように並べ替える．ただし，対応にあふれた概念 B の一次属性(b_{xj}，$j=L+1$，・・・，M)は無視する．

$B_x=((b_{x1},v_{x1}),\ (b_{x2},v_{x2})\ ,\cdots,\ (b_{xL},v_{xL}))$

概念 A と概念 B との関連度 $ChainW(A,\ B)$は，

$$ChainW(A,B)=\left(s_A/n_A+s_B/n_B\right)/2$$

$$s_A=\sum_{i=1}^{L}u_i MatchW(a_i,b_{xi})$$

$$n_A=\sum_{i=1}^{L}u_i$$

$$s_B=\sum_{i=1}^{L}v_i MatchW(a_i,b_{xi})$$

$$n_B=\sum_{j=1}^{M}v_j$$

である．すなわち，重み付き関連度は，概念 A から見たときの一致している属性の重みの割合 s_a/n_a と概念 B から見たときの一致している属性重みの割合 s_a/n_a の平均になる．

４．記事関連度計算

今回は上に述べた「概念ベース」と「概念間の関連度計算」を用いた記事関連度計算の手法を提案する．本稿では記事間の関連性の強さを表す数値を記事関連度といい，0 から 1 までの実数値で表す．

ここでは，「概念間の関連度計算」を記事間の関連度計算に適用する手法について述べる．

概念ベースは，

(a)概念

(b)概念の意味を表す属性

(c)属性の重要度を表す重み

の 3 つで構成されており，この 3 つを用いて概念間の関連度計算を行う．さて，ここで各記事から

(a)記事

(b)記事の意味を表す属性

(c)属性の重要度を表す重み

を得ることができれば，それぞれを概念の場合に対応させることが可能である．そして，記事を概念に対応させたものに対して「概念間の関連度計算」手法を適用すれば 2 者の関連度を求められるが，これは記事間の関連度と考えられる．

上記(a)から(c)について具体的に対応をとる．1 つの記事を 1 つの概念に対応させる場合，記事中に出現する単語を「記事の意味を表す属性」とみなして「概念の意味を表す属性」に対応させる．なぜならば，記事はそれ自身を構成する単語の連なりによって意味を持つからである．「概念の意味を表す属性」を概念の一次属性と呼ぶのに対し，この「記事の意味を表す属性」を記事の一次属性と呼ぶことにする．またこれらを索引語と呼ぶ．

「概念間の関連度計算」では概念間で対応する一次属性の一致度を用いるが，この際，概念の二次属性を利用して一致度を求める．したがって「概念間の関連度計算」を記事に適用する場合，記事の二次属性を定義しておく必要がある．そこで記事の二次属性を「記事の一次属性が持つ属性」とする．したがって，概念ベースを利用して索引語（＝記事の一次属性）の一次属性を導けば，それを記事の二次属性とみなすことができる．また，二次属性の重みには概念ベースに格納されている重みをそのまま用いることができる．

ここで記事の一次属性の重みを得られれば関連度計算が可能となる．今回提案する手法では記事の一次属性の重みとして tf・idf を用いる．tf・idf についてはこの後で述べるが，これは情報検索の分野でよく用いられる方法の１つである．

ところで, 記事の一次属性はその記事に含まれる語, と述べたが, 本稿では関連度を計算する場合にさらに条件を加えている. それは「自立語」かつ「概念ベースに登録されている概念」であることである. 前者は, 自立語以外の語は大多数の記事に登場するため記事の特徴づけには適さず, 記事の意味内容に与える影響が比較的小さいことによる. 後者は, 概念ベースで定義されていない語に関してはその属性を取得できないため, 関連度計算ができないことによる.

しかし現在, 概念ベースに登録されていない語を含んだ場合にも関連度計算を適用する手法を準備中である. そのため今後は後者の条件を考慮する必要はなくなる予定である.

5. tf・idf

ここで, 本研究で使用した, 重み付け手法の１つである tf・idf 重み付けについて説明する.

5.1 tf

ある記事 d 中に出現する索引語 t の頻度を索引語頻度（term frequency）と呼び, $tf(t,d)$ で表す. 索引語頻度に基づく重み付けの背景には,「何度も繰り返し言及される概念は重要な概念である」という仮定がある. ただし, 記事が長くなると平均的に語の出現頻度も多くなる傾向にある. したがって, 同じ索引語でも長い記事に現れる索引語の方が重みが大きくなる. 本稿ではこのような絶対頻度を用いるのではなく, (1)式に示すように, 索引語の出現頻度を記事中の全索引語の出現数で割った相対頻度を重みとして利用することにする.

$$w_t^d = \frac{tf(t,d)}{\sum_{s \in d} tf(s,d)} \qquad (1)$$

5.2 idf

(1)式は各記事内の頻度は考慮していても, 記事集合内の他の記事の索引語の分布については考慮していない. ある索引語が, どの程度その記事に特徴的に現れるのかという特定性を考慮するためには, 他の記事中の索引語の分布も考慮する必要がある. 特定性を表すための尺度として idf（inverse document frequency）が知られている. idf はある索引語が全記事中のどれくらいの記事に出現するかを表す尺度で(2)式で定義される.

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

ここで, N は対象となる記事集合中の全記事数, $df(t)$ は索引語 t が出現する記事数である. (2)式からわかるとおり, idf はある索引語か少数の記事にしか出現しない場合に大きくなり, どの記事にも出現すると最小の値となる. すなわち, IDF を索引語の重みとして用いると, 特定の少数の記事に出現する索引語に大きい重みを与えることができる.

5.3 tf・idf

以上に述べた２つの尺度を組み合わせて索引語の重みを計算することが考えられる. 具体的には索引語 t の重み w_t^d として, tf と idf の積を用いる（(3)式）.

$$w_t^d = tf(t,d) \cdot idf(t) \qquad (3)$$

6. 記事関連度計算手法の処理手順

(ア)入力記事の形態素解析を行い, 名詞・動詞・形容詞等の自立語を索引語として抽出する.
(イ)(ア)で抽出した索引語のうち, 概念ベースで定義されていない語を除く.
(ウ)索引語の重複を除く.
(エ)(ア)から(ウ)により得られた索引語を記事の一次属性として取得し, それらに tf・idf 重み付けによる重みを付与する.
(オ)各索引語の一次属性, つまり記事の二次属

性とその重みを取得する.
(カ)関連度を求めようとする 2 つの記事それぞれについて(ア)～(カ)の処理を実行し終えたら, 3 節で述べた概念間の関連度計算手法を利用して記事関連度を計算する.

7．記事関連度計算の評価

情報検索評価コレクション NTCIR2 を用いて提案手法の評価を行った.
NTCIR2 は，(1)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，(2)49 の検索課題，(3)検索課題に対する正解文書リスト，から構成されている.

本稿では(1)のうち，(2)の課題 1～10 の正解文書にあたる 567 件を対象として評価実験を行った．なお，(3)では各文書についてＳ(高適合)，Ａ(適合)，Ｂ(部分的適合)，Ｃ(不適合)の４段階があるが，実験ではＳ，Ａ，Ｂに該当する文書を使用した.

評価実験では，検索課題 1～10 についてそれぞれ 567 件の文書と関連度計算を行い，算出された関連度を各文書のスコアとした. 評価プログラム(trec_eval)はこのスコアに基づいて文書のランキングを行う. したがって関連度計算の精度が高ければ, 正解文書ほど課題との関連度が高くなり, 上位にランキングされると考えられる. そのため評価プログラムによる提案手法の有効性の検討が可能である.

8．表記一致法との比較

今回提案した記事関連度の計算手法が有効であるかを調べるため, 単純表記一致法との比較を行う．これは，概念間の関連度計算を用いずに単純に単語表記が一致する割合によって記事間の関連の程度を求める方法である. 本稿では単純表記一致による記事の類似度として, 以下の式を用いた．記事 X 中の単語数を x, 記事 Y 中の単語数を y, 記事 X と Y の両方に出現する単語数を m とするとき,

$$(m/a + m/b) / 2$$

である．ただし，概念ベースで定義されていない固有名詞などの一致があると直接比較できないため，本稿では x, y, m ともに概念ベースで定義されている概念のみの数とした. この方法を表記一致法と呼ぶことにする.

さらに，表記一致法において，各索引語に tf・idf による重み付けを行い，全索引語に対する重みの割合として記事の類似度を求める方法を重み付き表記一致法と呼ぶことにする.

9．実験結果

以上に述べた表記一致法と提案手法について実験を行った．なお正解判定には，Ｓ，Ａのみを適合とするレベル 1 と，Ｓ，Ａ，Ｂを適合とするレベル 2 がある.

レベル 1 における 11 点再現率・精度のグラフを図２に示す.

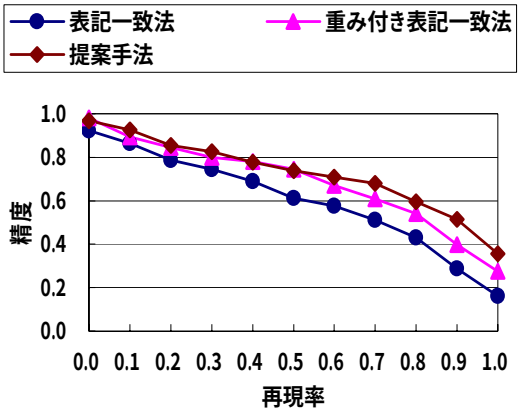

図２：11 点再現率・精度(レベル 1)

次にレベル 2 における 11 点再現率・精度のグラフを図３に示す.

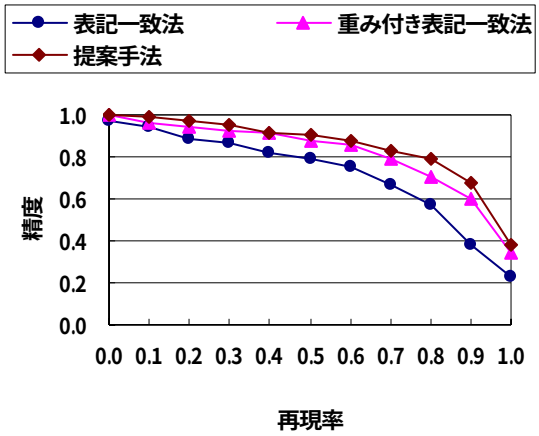

図３：11 点再現率・精度(レベル 2)

また, 平均精度を図４, R精度を図５に示す. R精度は, 正解文書数と同数出力した時点での精度である.

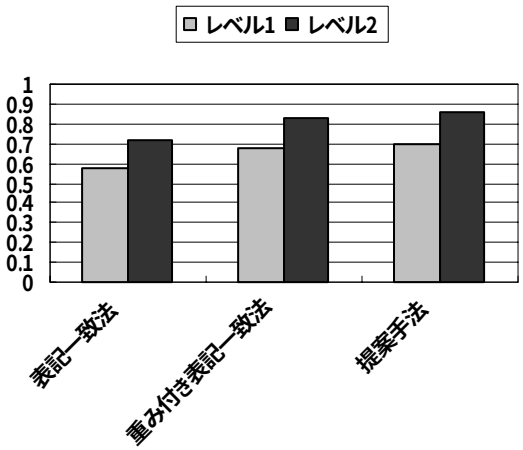

図４：平均精度

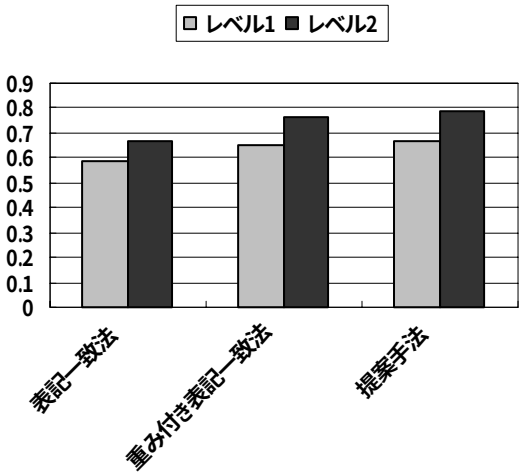

図５：R精度

いずれの結果を見ても, 表記一致の方法に比べて提案手法が有効であることがわかる. ただし今後さらに文書数, 課題数を増やして実験を行うことが必要である.

１０．おわりに

本研究では, 概念ベースを用いた記事の分類を実現するため, 記事と記事の関連性を数値化する手法について実験を行った. その結果, 概念ベース登録語を対象とした実験においては, 提案した記事関連度計算手法は有効であった. 今後, 概念ベースに登録されていない語も含めた手法の有効性を調べる予定である. 精度の高い記事関連度計算を利用することにより, 効率のよい文書の検索・分類が実現できると考えられる.

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]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 ”, 情報処理学会論文誌, Vol.38, No.7, pp.1272-1283 (1997)

[2]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, pp.159-160 (2002)

[3]広瀬幹規, 渡部広一, 河岡司, “ 概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精錬手法 ”, 信学技報, TL2001-49, pp.109-116 (2002)

[4]青田正宏, 東村貴裕, 渡部広一, 河岡司, “ 概念の漢字表記特徴を用いた関連度計算方式 ”, 第 16 回人工知能学会全国大会論文集, 1B3-03 (2002)