2D1-02

関連度計算を用いた短文と長文の意味的近さの定量化

—文献検索への適用—

Quantification of the Semantic Closeness of a Short Sentence and a Long Letter Using the Degree of Association -Application to Retrieving Documents-

倉田 篤史[*1] KURATA Atsushi　渡部 広一[*1] WATABE Hirokazu　河岡 司[*1] KAWAOKA Tsukasa

[*1] 同志社大学大学院 工学研究科 知識工学専攻
Department of Knowledge Engineering and Computer Sciences, Graduate School of Engineering, Doshisha University

Abstract: Recently, the amount of information which we humans can acquire is becoming immense by the development and the spread of computers and cellular phones. Information retrieval is indispensable to extract only required information efficiently out of immense information. This proposing method of retrieval is not by using the notation of input words but by understanding the meaning of input words. Concept-base and calculation of the degree of association enable to quantify the degree of the relation between concepts. By applying the calculation of the degree of association to a sentence and a letter, the degree of the relation between a short sentence, a need of a user, and a long letter, contents of a document, is quantified. In this paper, it is proposed the new method applicable to information retrieval.

1. はじめに

近年パソコンや携帯電話の発達により，ユーザが入手できる情報は，莫大なものとなってきている．多くの情報から必要なものだけを効率よく得るためには，情報検索が必要不可欠となる．

情報検索では，キーワードの表記情報を用いて，そのキーワードを含む情報を検出する．たとえ表記が一致しなくても，意味のよく似た語もある．単語には意味があり，決して表記のみでは，十分な情報とはいえない．そこで，コンピュータが語の意味を捉えることが望まれる．そのためには，コンピュータが言葉の意味を理解しなければならない．

既存の情報検索として，単語の表記情報のみを手がかりに検索を行うものがある．単語を記号として捉えるのではなく，意味を含めて概念として捉えることが必要となる．

本稿では，汎用知識ベースである概念ベース[廣瀬 2001]を用いて短文と長文の関連の強さを定量化する関連度計算[渡部 2001]手法を提案する．

2. 意味を捉えた文献検索

既存の文献検索システムは，ユーザがキーワードを入力し，そのキーワードが文献の題名などに含まれている文献を検出するものである．または，自然言語による文の形式で入力するものもある．しかし，これらは入力語の表記しか考慮していない．

入力語が表す意味を捉え，文献内容の意味と合うものを検出することが望ましい．表記は異なるが意味は近い語や，同じ表記でも全く異なる意味を示す多義語のような語が多く存在するからである．

そこで，概念ベースと関連度計算を用いることで，ユーザの入力する短文と，文献内容または概要である長文の関連の強さを定量化する．ユーザの要求と文献内容の関連の強いものを抽出すれば，ユーザの求める文献が検索できる．

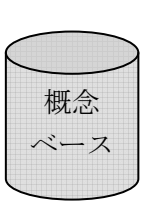
連絡先：同志社大学大学院工学研究科知識情報処理研究室，〒610-0394 京都府京田辺市多々羅都谷 1-3，Tel：0774-65-6944

3. 概念ベースと関連度計算

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（図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．1つの概念あたり，平均 30 個の属性が存在する．



概念:約９万語

概念	{属性, 重み}
医者	{(医師, 0.3), …, (治療, 0.1)}
太陽	{(日光, 0.2), …, (月, 0.05)}
美術	{(芸術, 0.4), …, (鑑賞, 0.07)}
:	:

図 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．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（表1）．

表 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	病院	‥	治す	一次属性
	医者	病人	医院	‥	治療	
	診察	包帯	患者	‥	医療	二次属性
	病院	看病	手術	‥	癒す	
	:	:	:	‥	:	
	保健	治療	施設	‥	病気	

3.2 重み付き関連度

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，重み付き連度計算を利用している．その計算方式のアルゴリズムを説明する．

まず，次の (1) 式のような概念 A と B を考える．

$$A=\{(a_1,wa_1),(a_2,wa_2),\cdots(a_M,wa_M)\}$$
$$B=\{(b_1,wb_1),(b_2,wb_2),\cdots(b_N,wb_N)\} \quad (1)$$

M，N は，それぞれ概念 A，B の属性数である．また，$a_i,wa_i(1\le i\le M)$ は，概念 A の属性とその重みである．同様に概念 B も $b_j,wb_j(1\le j\le N)$ で表される．二次属性についても，次の(2) 式のように表現する．

$$a_i=\{(a_{i1},wa_{i1}),(a_{i2},wa_{i2}),\cdots(a_{im_i},wa_{im_i})\} \quad (2)$$
$$b_j=\{(b_{j1},wb_{j1}),(b_{j2},wb_{j2}),\cdots(b_{jn_j},wb_{jn_j})\}$$

このとき，一次属性 a_i と b_j の重み付き一致度 $MatchW(a_i,b_j)$ を次の(3)式のように定義する．

$$\mathrm{MatchW}(a_i,b_j)=(S_a/n_a+S_b/n_b)/2 \quad (3)$$
$$\left(S_a=\sum_{a_{i\alpha}=b_{i\beta}}wa_{i\alpha},n_a=\sum_{\alpha=1}^{n}wa_\alpha,S_b=\sum_{a_{j\beta}=b_{j\beta}}wb_{j\beta},n_b=\sum_{\beta=1}^{n}wb_{j\beta}\right)$$

これは，関連度計算で用いる属性数を n 個としている場合である．そして，概念 A の属性の並びを固定し，重み付き一致度が最大になるように，概念 B の属性を並べ替える．属性を n 個で打ち切っているわけだが，属性がどれほど多く存在しようとも，30 個で打ち切った場合に，関連度の精度と処理時間の観点から最も適切であることが実験的に分かっている．

重み付き一致度から概念 A と B の重み付き関連度を(4)式のように $ChainW(A,B)$ を求める．

$$\mathrm{ChainW}(A,B)=(S_A/n_A+S_B/n_B)/2 \quad (4)$$
$$\left(S_A=\sum_{i=1}^{M}wa_iMatchW(a_i,b_i),n_A=\sum_{i=1}^{M}wa_i\right)$$
$$\left(S_B=\sum_{i=1}^{M}wb_iMatchW(a_i,b_i),n_B=\sum_{i=1}^{N}wb_i\right)$$

4. 関連度計算の文への適用

4.1 関連度計算を文に適用する方法

関連度計算とは，もともと概念と概念，つまり単語と単語の関連性を定量化する目的で，考案されたものである．そこで関連度は，単語（概念）の意味特徴を表す属性を利用して，計算される．

この方式を文へ適用するには，まず一連の文章を1つの概念とする．そして，文を単語に分割し，その各単語を概念の属性とする．構文解析ソフトの茶筌[Chasen 奈良先端科学技術大学院大学 2003]を利用し，文の自立語を抽出し，それらを文の属性とする．また，各単語に対する重み付けの方法に関しては，6章で説明する．

4.2 関連度計算方式を文に適用する問題点

前章で述べた重み付き関連度など，従来の関連度計算方式では，概念と概念，つまり単語と単語の関連性を定量化するものであるため，両概念の属性数がほぼ同じだけ存在することが前提となっている．

しかし，本研究の目的である短文と長文の関連度を求めることを考えると，短文と長文から得られる属性数に偏りが生じてしまう．そこで，新たな関連度計算方式を考案する必要がある．

5. 属性平均関連度計算方式の提案と検証

5.1 属性平均関連度計算方式の提案

(1)式の概念 A，B が(2)式のような属性をもつとする．ただし，概念 A を短文から得られた概念とする．よって，両概念の属性数の関係は，$M<<N$ とする

(4) 式の重み付き関連度を利用して，属性平均関連度 $AAR(A,B)$ を(5)式のように定義する．

$$AAR(A,B)=\frac{\sum_{i=1}^{M}wa_i\cdot ChainW(a_i,B)}{M} \quad (5)$$

この方式では，属性数が少ない概念 A の属性 a_i を概念とみなすため，概念ベースに存在する概念 a_i は，属性を平均 30 個もつため，$ChainW(a_i,B)$ を求めるとき，属性数の偏りが生じることはない．また，文章に出現する概念ベースにない語は，関連度計算を行うときの属性として，採用しない．

5.2 属性平均関連度計算方式の検証

提案した関連度計算方式を検証するために文献データベース（以下，文献 DB）を作成した．この文献 DB には，情報処理学会論文[情報処理学会 2003]の題名と概要が 150 件格納されている．

題名と概要からそれぞれの自立語とその重み（次章参照）を取得する．そして，題名（短文）と概要（長文）との関連度を計算し，実際の論文の題名と概要の組み合わせの関連度が最大となれば正解として，評価を行った．ただし，重みはすべての自立語（属性）に同じ値が割り振られている．結果は，従来の関連度計算方式よりも 4.7%の正解率の向上が見られた．（図2）

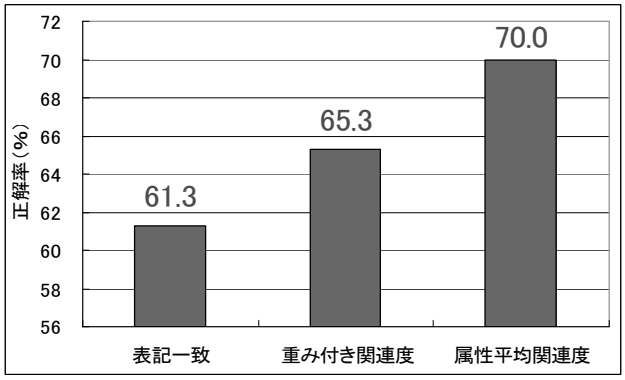


図 2 関連度計算方式の比較結果

6. 重み付け手法

6.1 長文に対する重み付け

長文に対する重み付けには，情報検索の分野で用いられている TF･IDF 重み付け[徳永 1999]を使用する．

(1) TF

TF とは，出現頻度を表す値であり，語の網羅性を示している．文書 d 中に出現する語 t の頻度を $tf(t,d)$ と表す．この値を採用するのは，同じ文書中に何度も繰り返し出現する語は，その文書で重要な語であるという考えからである．

文献 DB では，ある1つの論文の概要の中にその語が何回出現するかを TF の値とする．

(2) IDF

IDF とは，ある語がどの程度その文書に特徴的に現れるのかという特定性を表す尺度である．そのため，ある文書内だけではなく，文書群が存在する空間全体での語の分布を調べる必要がある．語 t の IDF の値 $idf(t)$ は，次の(6)式のように定める．

$$idf(t) = \log(N/df(t)) \quad (6)$$

ここで，N は文書群に存在する全文書数，$df(t)$ は語 t が出現する文書数である．

文献 DB では，$df(t)$ はある論文の概要に出現した語が他のいくつの論文に出現するかを示し，N は全論文数の 150 となる．

6.2 短文に対する重み付け

短文に対する重み付けとしては，長文と同様に TF･IDF を用いることができない．そこで，TF に替わる重みを提案する．IDF に関しては，長文と同じものを適用する．

(1) 意味的にかかる語の数

短文 s 中に出現する語 t に係る語数を $Modify(t,s)$ とすると，TF に相当する重み $Stf(t,s)$ は，次の(7)式のように定義する．

$$Stf(t,s) = Modify(t,s) + 1 \quad (7)$$

この $Stf(t,s)$ を具体的に“赤くて丸いボール”を使って説明する．ここで，どの語がどの語に意味的にかかるかを知るために構文解析ソフト CaboCha[CaboCha 奈良先端科学技術大学院大学 2003]を使用する．

“丸くて赤いボール”という表現を CaboCha で解析すると，“丸い”と“赤い”は，“ボール”にかかっていることが分かる(図3)．

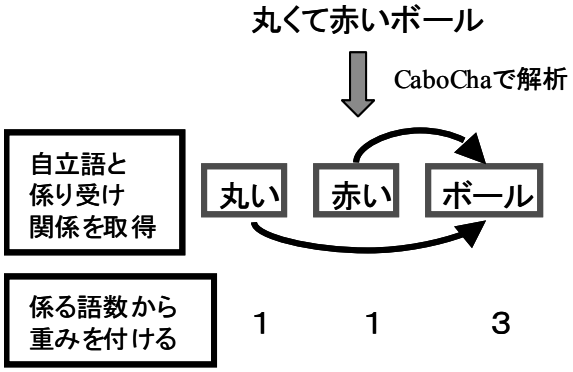


図 3 意味的にかかる語の数

語の意味的な係り受けを調べることにより，この表現は“丸いボール”であり，かつ“赤いボール”であるという意味である．よって，“ボール”という表現が重ねて出現するため，TF の考え方より“丸い”，“赤い”より“ボール”という語がこの表現の中で重要であり，大きな重みを付与する．かかる語の数に 1 を加えているのは，意味的にかかる語がない“丸い”や“赤い”の重みを 0 ではなく，1 とするためである．

(2) IDF

長文から得られる自立語に対する重み付けの場合と同じく，短文の場合にも IDF を用いる．短文に出現する語が文書群の中で，いかに特定の文書を特徴付ける語であるかを知るためである．

文献 DB では，$df(t)$ は，題名に出現する語が全論文の概要の中で，いくつの論文に出現するかを示す．ただし，題名に出現した語が必ず概要に出現するとは限らないため，$df(t)$ が 0 となる場合は，0 に近い小さな定数とする．

6.3 重み付け手法の検証

どの重みを適用した場合が良い結果となるかを検証する．その方法は，5.2 節で述べた文献 DB を利用した方法である．

(1) 長文の自立語に対する重み付け手法の検証

長文から得られる自立語に対して，重みなし，TF，IDF，TF･IDF を重みとする 4 通りの方法で比較した．“重みなし”とは，全ての自立語に対して，一定の値が割り振られているということである．結果は，TF･IDF を重みとした場合が最も良い結果となった(図4)．このとき，短文から得られる自立語に対しては，一定の値が割り振られている．

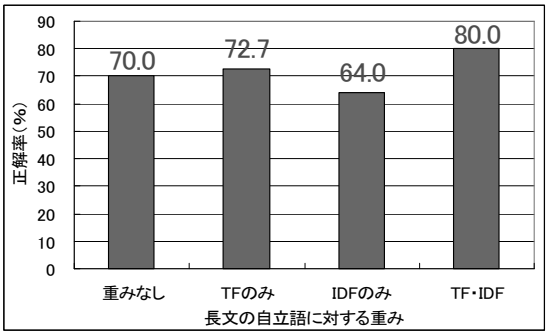


図 4 長文の自立語に対する重み付け結果

(2) 短文の自立語に対する重み付け手法の検証

短文から得られる自立語の重みに関しては，重みなし，意味的にかかる語の数に 1 を加えた STF，長文の場合と同じ IDF，STF･IDF，√(STF)･IDF の 5 通りの方法で比較した．結果は，STF･IDF を重みとした場合が最も良い正解率が得られた(図5)．ただし，長文から得られる自立語に対しては，一定の値が割り振られている．

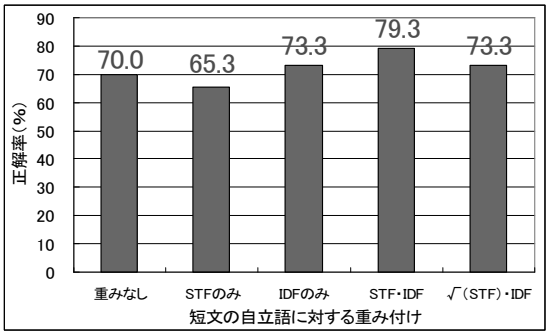


図 5 短文の自立語に対する重み付け結果

(3) すべての短文と長文に対する重み付け手法の検証

短文にも長文にも重み付けを行った結果，短文に対しては√(STF)･IDF，長文に対しては TF･IDF を重みとした場合が最もよく，正解率 84.7%であった(表2)．

STF と IDF の積を重みとする場合，STF と IDF のバランスを考慮して，STF の影響を軽減するために STF の平方根と IDF の積を重みとした場合が最も良い結果となった．

短文と長文に対する重みの組み合わせによって，STF･IDF が最も良い結果となる場合や√(STF)･IDF が最も良い結果と

なる場合がある. しかし, 短文に対しては IDF のみよりも STF を乗じた場合の方が良い結果となるといえる.

表 2 短文と長文の重みの組み合わせによる正解率

正解率(%)		題名 短文			
		STF	IDF	STF・IDF	√(STF)・IDF
概要 長文	TF	62.0	70.0	66.7	70.0
	IDF	58.7	72.0	71.3	72.7
	TF・IDF	67.3	84.0	80.7	**84.7**

7. 概念ベースにない語の活用

7.1 概念ベースにない語の活用方法の提案

概念ベースにない語(以下, 未定義語)は, 関連度計算を行うことができない. 概念ベースは, 汎用的知識であるため, 各分野の専門用語, 特にカタカナ語が格納されていない. しかし, 論文などでは専門用語がその論文を特徴付けていることはいうまでもない. そこで, 未定義語を関連度計算以外の方法で活用することを考える.

まず, 文から自立語を取得するときに, 概念ベースにあるかどうかで分ける. ある語に関しては, 関連度計算を行う際の属性に採用する. ない語に関しては, その語の表記を情報として, 表記一致する語の数をカウントする. (図6)

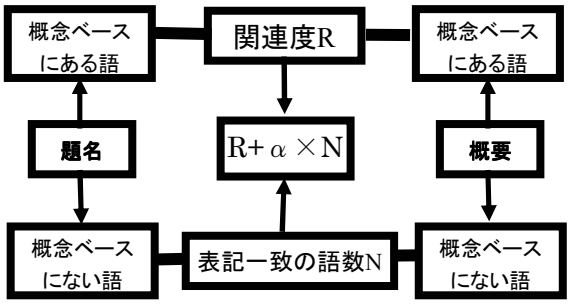


図 6 未定義語の活用方法

概念ベースにある語同士の関連度 R と未定義語同士での表記一致の数 N を R+α×N の値で短文と長文の関連性を評価する.

7.2 概念ベースにない語の活用方法の検証

α の値を 0 から 0.1 まで, 0.01 刻みで正解率の推移を文献 DB で調べた. α=0.01 のとき, 正解率が 85.3%となり, 最大となった(図7).

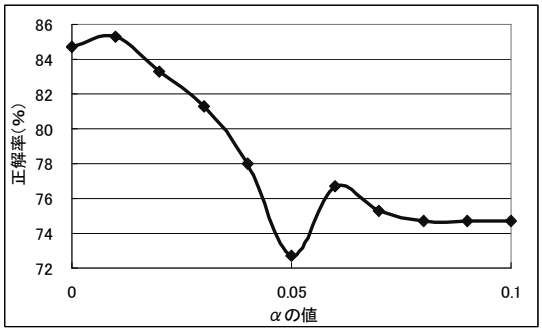


図 7 未定義語を活用した結果

8. 考察

表記のみを情報として検索するよりも, 概念ベースを利用し, 言葉の意味を理解した上での検索が, より適切な結果を導けることが分かった. しかし, この方法は検索時間がかかってしまう. 検索のシステムとするには, この方法はまだ改良の余地がある.

重み付けの方法に関しては, 短文という情報量が限られた中で, いかに出現頻度(TF)に替わる重み付けを行うかが難しい問題であった. 今回の実験では, IDF のみを短文の重みとするよりも, 考案した STF を用いることでより良い結果が得られた. ただ, 平方根をとるかどうかによって, IDF とのバランスを考える必要がまだ残っている.

未定義語の活用において, α の値を 0.1 以上にしたところ, 収束することが分かった. また, この α の値がこの文献 DB の固有のものでないか, 未定義語がどの程度の割合で出現しているかによっても変化する可能性がある. しかし, 未定義語を区別する方法は, 今後, 概念ベースの自動学習が実現することで, 未定義語の数が減少し, 解決されるだろう.

今後の課題としては, さらに大規模なデータベースに対する検索に対しても, 検索時間を考慮した方法を考案しなければならない. 重み付けに関しても短文から自立語に対する重みが正しく取れているかを異なるデータをもとに調査すること必要である. また, 論文の題名という限られた形式のデータではなく, よりユーザの入力文に近い, 自然言語を評価データとして採用しなければならない.

9. おわりに

本稿では, コンピュータに言葉の意味を理解させ, 文献検索などに応用できる検索手法を提案した. 莫大な情報があふれる情報化社会において, 情報検索は必要不可欠である.

今後, 長文から本当に必要な語だけを取り出し, 短文だけでは足りない情報を拡張することによって, より精度が高く, 検索時間を考慮した方法を考案していく予定である.

謝辞

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

[廣瀬 2001] 廣瀬幹規: 概念ベースの自動精錬と評価, 同志社大学工学部知識工学科卒業論文 ,2001.

[渡部 2001] 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.

[Chasen 奈良先端科学技術大学院大学 2003] 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座 松本研究室, http://chasen.aist-nara.ac.jp/hiki/ChaSen/, 2003.

[情報処理学会 2003] 社会法人情報処理学会 Information Processing Society of Japan, http://www.ipsj.or.jp/, 2003.

[徳永 1999] 徳永健伸, 情報検索と言語処理, 東京大学出版, 1999.

[CaboCha 奈良先端科学技術大学院大学 2003] 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座, http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/, 2003.