2D1-01

wwwを用いた概念ベースにない新概念およびその属性獲得手法

The Method of Acquisition of The New Concept and Its Attribute Using The World Wide Web

辻 泰希[*1] 渡部 広一[*1] 河岡 司[*1]
TSUZI Yasuki WATABE Hirokazu 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

This paper proposes the method of acquisition of the concept which is not in the concept base, and its attributes at high speed and with high precision using World Wide Web which the amount of information is increasing explosively. Highly precise attribute acquisition is gained by using IDF and statistical IDF in Web. Finally, this paper evaluates accuracy of attribute and shows the validity of the proposed method.

1. はじめに

近年，人間のような状況に応じた柔軟な対応や，常識的・知的な判断を行える情報処理システムの登場が望まれている．知的な判断をコンピュータで実現するための知的メカニズムの一つである常識判断メカニズムの中心をなす要素として，概念ベース[渡部 2001]と呼ばれる知識ベースがある．

概念ベースでは，ある単語の意味（概念）をその単語に関連の深い単語群（属性）で定義する．例えば，人間は，「木」から「葉」や「枝」などの単語を連想できる．この場合，概念を概念ベースでは，「木」の概念を{葉，枝，植物，水，…}のように概念－属性群として格納している．

しかし，知的なコンピュータを実現するための核となる概念ベースに登録されている語は限られている．本稿では，概念ベースに存在しない概念およびその属性を，情報量が爆発的に増加している Web を用いて，高速かつ高精度に獲得する手法について述べる．

2. 常識判断メカニズム

知的な判断をするシステムとして，各常識判断メカニズムが提案されている．常識判断メカニズムとは，会話メカニズムによって人間と会話し，各種判断メカニズムによって判断を行うシステムである（図1）．各判断メカニズムに対する入力が自分の知識ベースにないときは背後に存在する概念ベースを利用し，知っている語の中でその語と最も関連の強いものと置き換える処理を行う．

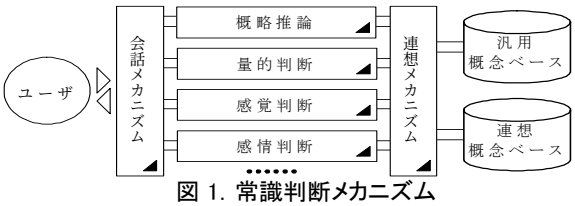

図 1. 常識判断メカニズム

また，概念ベースの利用法として，概念間の関連の深さを数値化した関連度計算が挙げられる．関連度計算によって導かれた関連度は，概念間の関連の強さを計るために使われ，各判断メカニズムで使用されている．

連絡先：辻泰希 同志社大学工学研究科 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

概念ベースに格納される概念の定義，および概念ベースの構造を定める．概念ベースの単純な機能として，一つの入力単語に対して複数の単語を返すという機能がある．そのために，概念を以下のように定義する．

［概念］：単語に対し定義された，その単語の意味特徴をあらわす単語（属性）の集合．
例） 木{葉，枝，植物，水，酸素，…}

概念ベースでは概念と属性の関連の深さを表す重みが付加されている．この属性と重みの対集合で一つの概念を構成している．概念ベース K におけるある概念 $Word_i$ を，意味特徴をあらわす属性 p_{ij}(j=1,2,…)と，概念 $Word_i$ と属性 p_{ij} の重み($q_{ij}\neq 0$)の対集合で以下のように定義する．

$$概念\ Word_i=\{(p_{i1},q_{i1}),\cdots,(p_{ij},q_{ij}),\cdots\}$$

現在，概念ベースは概念約 9 万個を保有しているが，問題の1つとして，固有名詞概念が少ないことがあげられている．この概念ベースに登録されていない語を未定義語と呼ぶ．

4. WWW 検索エンジン

4.1 WWW 検索エンジンの機能と特徴

WWW検索エンジンは，ユーザが必要としている情報がインターネット上のどの Web ページで提供されているかを見つけ出すことを支援するシステムであり，大きく，“ディレクトリ型検索エンジン”と“ロボット型検索エンジン”に分けることが出来る．

ユーザは検索要求を複数のキーワードで表現し，検索エンジンに入力する．複数あるいは一つのキーワードで構成される入力を検索質問文という．実用化されている検索エンジンの多くは，検索質問文中のキーワードを含む Web ページを素早く見つけるために，あらかじめ世界中から Web ページを取得し，索引付け処理を行ったあとデータベースに登録している．

現在，Google[Google]はロボット型検索エンジンの代表的検索エンジンといえる．Google の大きな特徴として，1つめに保有ページ数(全言語において 2004 年 4 月現在 3,307,998,701 ページ)が非常に多いことあげられる．2つめに検索のつど一致するテキストの抜粋を，検索語句にハイライトをつけて結果リストに表

示することがあげられる. これらの特徴から本稿では検索エンジンとして Google をもちいた.

4.2 NOT 検索質問文

WWW には多種多様な Web ページがあり, その全てが有益な Web ページとは言えない. そこで WWW 検索エンジンでは, あるキーワードを含む Web ページを検索する通常の機能に加え, あるキーワードを含まない Web ページを検索するという NOT 検索という機能がある. NOT 検索を行うためには, 通常の検索質問文に加えて, 表示したくない Web ページを特異付けるキーワードを"-"を付けて付加する. NOT 検索質問文は式(2)のように定義する.

"Word －Word1 －Word2 … －Wordn" (2)

式(2)は「"Word"というキーワードを含み, かつ"Word1 ", "Word2", …, "Wordn"というキーワードを含まない Web ページ」を検索するための検索質問文である. この NOT 検索質問文を用いることにより, 不適切な Web ページを除去でき属性獲得精度が向上すると考える.

5. 属性獲得手法

WWW 情報空間には, 意味のある語に関する情報はほぼ存在する. つまり, ある語を検索すると, その語に関する情報が「数」「質」は保証しないにしろ受け取ることができる. 本稿では, WWW 情報空間に存在する多くの情報を利用し, 未定義語の属性を獲得する手法を提案する.

属性獲得手法は4.2で説明した NOT 検索質問文を用いて検索を行い, 検索結果の上位 L 件の Web ページを取得する. 取得した Web ページは HTML 形式であるため, そのままでは解析対象の文書としてはふさわしくない. そこでまず, スクリプト部分を取り除き, タグ除去を行うことにより文書を整形する. そして整形した文書を茶筌[松本 2002]で形態素解析を行う. しかし, 解析時に解析を失敗する場合がある. その場合, 失敗を補正するためのルールに従って適切な属性を取得し, 必要であれば再結合を行い属性候補とする. そして取得した属性候補にそれぞれ TF・IDF[徳永 1999]による重み付けを行い精錬することにより概念の属性を取得する(**図 2**).

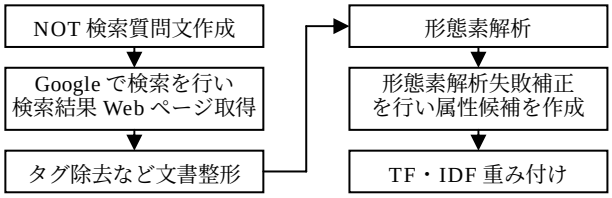

図 2. 属性候補獲得の流れ

6. 属性獲得精度向上手法

6.1 形態素解析補正

本研究ではタグを除去した文書に対し形態素解析を行い, 属性を取得した. しかし, 形態素解析には,「死亡者」→「死」／「亡者」や「ガイドライン」→「ガイド」／「ライン」のような単語の切れ目の誤りや「ストーカー」→「スる」／「トーカー」のように形態を間違うといった問題がある.

形態素解析の失敗は大きく次の3つのパターンに分けることが出来る.

- 複合語の過多分割(例:「基本情報処理技術者試験」「死亡者」).
- 新しい英単語への未対応(例:「BSE」「KHV」).
- 新しいカタカナ語への未対応(例:「ストーカー」「ピッキング」).

この3つの問題を解決するために, "必要でない形態素, あるいは区切り文字が出現するまで結合を続ける"ことを基本とした補正ルールを作成した.

本稿では出来る限り, 複合語や固有名詞などの語をひとつの形態素として取得することにより, 属性獲得精度の向上を目指した. 但し例外として語尾が「～する」で終わるサ変動詞については形を変え語幹のみとした. 例えば「走行」と「走行する」のように, 品詞は異なるが概念としては同一と考えて差し支えないと考えた. このようにサ変名詞＋「する」については,「～する」の部分を除いた名詞と同一の単語とみなして処理した.

6.2 TF･IDF を用いた重み付け

索引語(文書の内容を表現する語) t の文書 d における重み $w(t,d)$として, 頻度 $tf(t,d)$と索引語の網羅性 $idf(t)$によって計算される TF･IDF[Steve 1998]は次の式(1)の通りである.

$$w(t,d) = tf(t,d) \cdot idf(t) \quad (1)$$

本稿では後述の2種類の IDF(SWeb-idf と Web-idf)を使用した.

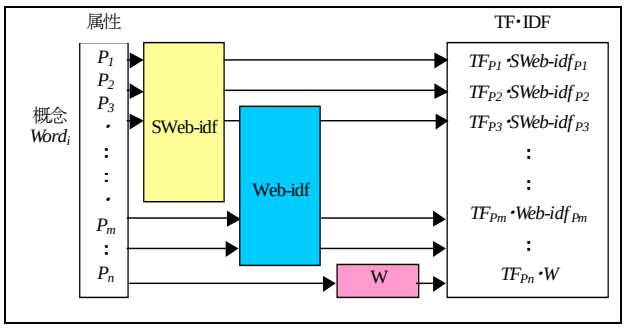

図 3. 重み付けルール

具体的手順は, DB を参照し属性候補の語が SWeb-idf に登録されていれば, IDF 値を SWeb-idf とする. 登録されていなければ Web-idf を参照し, 登録されていれば IDF 値を Web-idf とする. SWeb-idf と Web-idf ともに登録されていなかった属性に対しては定数重み W を IDF の代わりとする(**図 3**).

(1) Web-idf

Web-idf は WWW 上での語の大局的な重みである. 検索対象文書数 N を Google が保有している日本語のページ数とする.

ただし, Google が全言語において保有しているページ数は公開されているが, 日本語のページとして保有している数は公開されていない. そこで日本語の文書として最も使われている「は」で検索を行ったヒット件数[大森 2000](416,000,000)を Google が保有している日本語の全ページ数とし, Web-idf の検索対象文書数 とする. 索引 t が出現する文書の数 $df(t)$は, 索引語 t を Google で検索を行った時のヒット件数とする.

これを概念ベース内の全概念と, シソーラス[NTT コミュニケーション科学研究所 1997]のリーフとして登録されている語全て(合計約 14 万語)について行い, $(t, Web\text{-}idf_t)$のセットの形で DB に登録をした. $Web\text{-}idf_t$とは索引語 t の Web-idf 値である.

(2) SWeb-idf

SWeb-idf とは Web 上の語の IDF を統計的に調べた IDF 値である. まず, 無差別に選んだ固有名詞(語数 1132)を作成する. この作成した語, 全てについて個々に Google で検索を行い, それぞれの検索結果のページを取得する. ここから得られたページ数は 1132 ページとなる. この 1132 ページを Web の全文書空間と見なし, ページ数を検索文書の数Nとする. その後, 通常の IDF と同様に索引語 t が出現する文書の数 とする. これらにより得られた語 t を, (t, $SWeb\text{-}idf_t$)の形で DB に登録をした. $SWeb\text{-}idf_t$とは索引語 t の SWeb-idf 値である.

(3) Now-Web-idf

Web-idf はあらかじめ値を求めておき, DB に登録をした IDF 値である. しかし, 登録されている語は有限であるため, Web-idf のみで全ての語の重み付けを行うこと不可能である.

Now-Web-idf は属性候補獲得時に全属性候補に対し, Web-idf と同様の手法で IDF 値を取得する. この逐次 IDF を取得する方法を Now-Web-idf と定義する. この IDF の特徴は, 全ての属性候補に対して IDF による重み付けを行うことが可能である. 但し IDF 値を得るためには, 属性候補の数だけヒット件数を得るための通信が必要となる. 属性候補が 100 個存在すれば 100 回通信が必要となり, 膨大な時間が必要となる.

6.3 Not 検索質問文

7. 実験と評価, 考察

未定義語で構成される概念数 120 語を評価セットとする(**表 1**). 今回の実験では属性として重みの大きい順に属性を 50 個取得した.

表 1. 評価セット一部

BSE	サイクロン掃除機	自爆テロ
SARS	チェルノブイリ原発事故	小泉純一郎
カブトミジンコ	ロード・オブ・ザ・リング	おれおれ詐欺

この評価セットを用いて属性候補獲得を行い, 属性がその概念と関連が深いかを基準に人手で評価した. 評価としては, 重みの大きい順に属性を 50 個取得し, 概念と関係が深いと判断した属性を適切と判断した. 評価基準として, 固有名詞概念の中に, 適切な固有名詞属性が含まれる場合は適切な属性と見なした. 評価の一例を**表 2**に示す.

全属性候補の内, 適切な属性が占めている割合を属性獲得精度と呼ぶ. **表 2**の場合, 属性獲得精度は 75.00%(6/8)となる.

表 2. 評価例

概念 BSE			
属性	評価	属性	評価
BSE	○	牛海綿状脳症	○
狂牛病	○	検査	○
リーフレット	×	米BSE	○
関係	×	牛肉	○

7.1 処理速度評価実験

提案手法と他の手法の速度評価実験を行った(**図 4**). 処理速度の評価方法として, 無作為に選んだ固有名詞 30 語に対して属性獲得を5回行い, 平均の処理時間を測定する. 測定方法として, プロファイルソフトの DProf [5] を使用した. 処理速度の比較として重み付けに Now-Web-idf を用いた場合との比較を行った.

実験結果より, 全属性候補に対し通信を行い Google のヒット件数を調べることにより Web-idf 重み付けを行う Now-Web-idf と比較した場合, 処理時間は約 1/28 の 2.49 秒となり, 提案手法が速度面において非常に有効な重み付け方法であることがわかる.

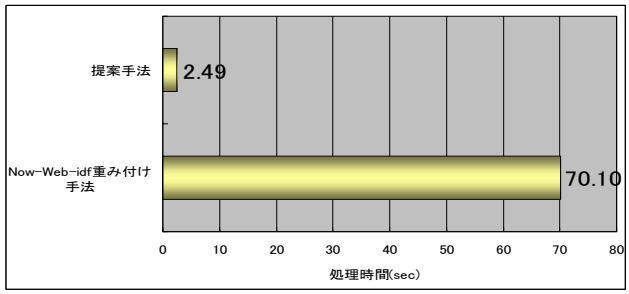

図 4. 処理速度実験結果

7.2 形態素解析補正の属性獲得精度への影響と評価

形態素解析失敗補正の属性獲得精度への影響を調べる実験を行った. 重み付けとしては TF 重みのみを使用した.

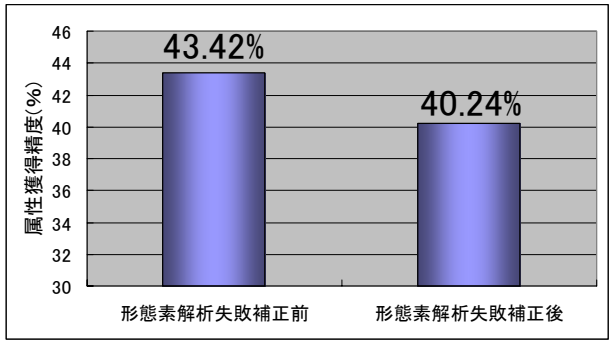

図 5. 形態素解析補正結果

実験の結果, 属性候補獲得精度としては約3%下がってしまった. しかし, 実際に属性候補を参照すると(表 3)雑音も含まれているが適切な属性が取得されていることがわかる.

表 3. 形態素解析失敗補正の属性候補例

概念 BSE	
補正せず	補正あり
牛	BSE
海綿	牛海綿状脳症
脳症	牛
対策	平成
平成	ハイライト表示しない

適切な属性が取得することが出来ているのにもかかわらず, 精度が下がった原因として次の2つが考えられる. 1つ目として, それぞれ適切な属性から構成される複合語が存在することがあげられる. これは表 3 のように,「牛海綿状脳症」は補正後の場合1語となるが, 補正前の場合は「牛」「海綿」「脳症」の3語となり, 上位 50 位の属性に対し, 占める割合の影響が大きくなることが考えられる. 2つ目として, 形態素解析時に茶筌が「unknown」と解析するものを許可したため, これまで取得していなかった不適切な語が含まれるようになったことが考えられる. しかし, これは適切な重み付けを行うことによって解決すると考えられる.

7.3 Web-idf, SWeb-idf の評価

6.2で説明を行った2種類の IDF の属性獲得精度についての影響について評価実験を行った.

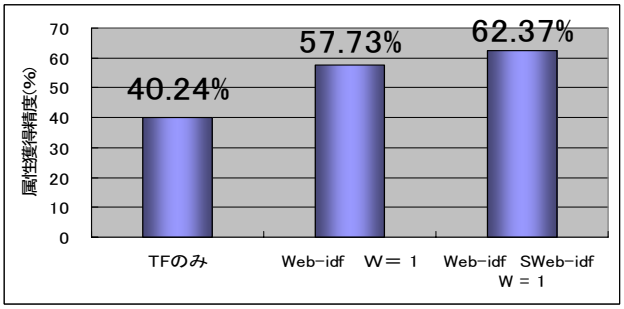

図 6. 2つの IDF による属性獲得精度

実験の結果(**図 6**), Web-idf 単独重み付け Web-idf と SWeb-idf を組み合わせた重み付けともに TF 重み付けよりも属性獲得精度が向上した. Web-idf と SWeb-idf に登録されていなかったときに使う定数重み W は. この実験では W=1 とした.

Web-idf 単独で重み付けを行うよりも, Web-idf と SWeb-idf を組み合わせた重み付けを行う方が 5%程属性獲得精度が良かったため, 本研究では Web-idf と SWeb-idf を組み合わせた重み付けを採用した.

また, Web-idf と SWeb-idf を組み合わせ, W の値を変化させて属性獲得精度を調査した結果, 定数重み W=10 の時がもっとも精度がよい結果となった.

7.4 提案手法の属性獲得精度の実験と評価

Web-idf と SWeb-idf を組み合わせ, W=10 とする重み付けを提案する. この提案手法の重み付け手法の有効性を調べるため, TF 重み付けと, 6.2 で説明した Now-Web-idf による重み付けとの比較実験を行った.

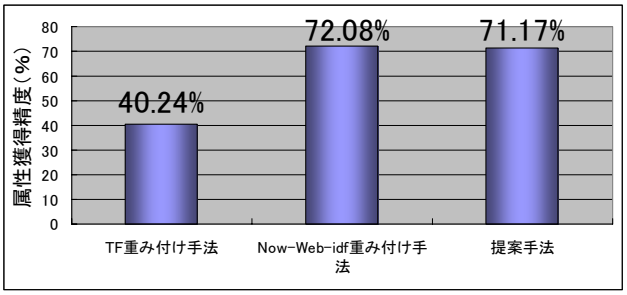

図 7. 提案手法の属性獲得精度

提案手法と TF 重み付けを比較した場合, 提案手法は TF 重み付けより約 25%向上し 71.17%となった.

全属性候補に対し Now-Web-idf 重み付けを行った手法と比較した場合, 属性獲得精度がほとんど変わらなかった. しかし, 提案手法は7.1で述べた処理時間の実験では処理時間は非常に速く約 1/28 だった. このことにより本稿の提案手法は高速・高精度で属性候補を獲得できたと考える.

また, 全く別の評価セット(概念数 100 語)を用いて属性候補獲得精度を調べたところ 76.11%となり, 評価セットに依存していないことを確認した(図 8).

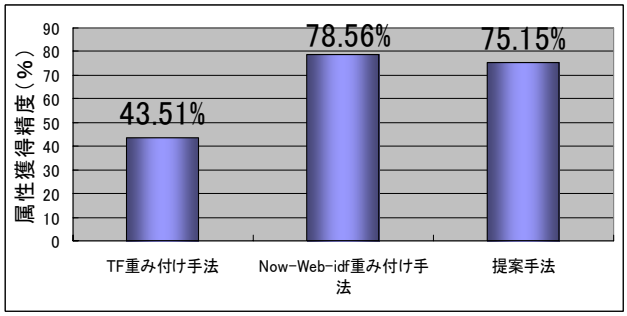

図 8. 別評価セットでの属性獲得精度の評価

8. おわりに

本稿では, WWW を用いて概念ベースに登録されていない概念の属性候補を, 高速・高精度に取得する手法を提案した. 提案手法を用いることにより, 概念ベースは登録されていない最新の時事語に対しても, 対応することが可能となり, コンピュータでより知的な判断を実現することが出来ると考える.

しかし, 実験を進めるに当たり2つの問題が発生した. 1つ目の問題としては, 獲得した属性には多数の仲間語が含まれている場合がある. 仲間語とは, 鳥{鳩, 雀, 鷹, 鷲, 鶴, …}のように, シソーラスの同一ノードに登録されている固有名詞などを指す. 多数の仲間語が属性に含まれると, 概念の骨格をなす属性の重みが相対的に薄れる問題があげられている. その為, 概念ベースに登録する場合には, このような仲間語の重みを相対的に下げるなどの内部的に属性の精錬をすることが必要となる.

2つ目の問題としては, 多義語の問題がある. 例として「スカイライン」という概念の属性を取得することを挙げる. 検索エンジンで「スカイライン」を検索した場合, 「自動車のスカイライン」に関するページと「道路のスカイライン」に関するページが両方提示されてしまい, 自動車と道に関する属性が入り乱れてしまうこととなる.

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

[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.

[Google] Google:http://www.google.co.jp

[松本 2002] 松本裕治,北内啓,山下達雄,平野善隆,松田寛,高岡一馬,浅原 正幸:日本語形態素解析システム『茶筌』使用説明書.奈良先端科学技術大学院大学 松本研究室,2002.

[徳永 1999] 徳永健伸:言語と計算 5 情報検索と言語処理,東京大学出版会, 1999

[Steve 1998] Steve Lawrence and C. Lee Giles, Searching the World Wide Web, SCIENCE, Vol.280, pp.98-100, 3 April 1998

[DProf] DProf:perl スクリプトの実行速度を分析するプロファイラソフト http://search.cpan.org/search?module=Devel::DProf

[大森 2000] 大森貴博:日本語 Web ページの統計的推定, 経営情報科学 Vol12 89, 2000/6

[NTT コミュニケーション科学研究所 1997] NTT コミュニケーション科学研究所:日本語語彙体系, 1997

[北 2002]北研二 津田和彦 獅々堀正幹;“情報検索アルゴリム”, 共立出版, 2002