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

2004－NL－159　(10)
2004／1／13

情報検索における未知語理解支援方式
～未知語のシソーラスノードへの分類～

伊藤 俊介†　渡部 広一†　河岡 司†

情報があふれる社会の中で, 必要な情報を効率よく的確に得る情報検索が必要となっている. しかし現状の検索エンジンによって必要な情報を得ることは，候補として示された多くのドキュメントを読み，理解しなければならないために時間および努力を要求する．そのためユーザにとって現状の情報検索システムを使うことはかなりの負担となっている．キーワードによる検索要求に対して，その語が属するシソーラスのノードを提示し，ユーザの理解を支援することは，検索にかかるコストを大幅に削減すると考えられる．本論文では，既存のシソーラスと概念ベースと Web を用いて, 未知語をシソーラスの最適なノードにマッピングする手法について述べる．

Unknown Word Understanding Support System for Intelligent Information Retrieval.
～Method of Classifying an Unknown Word to a Thesaurus Node～

Shunsuke ITOH†　Hirokazu WATABE†　Tsukasa KAWAOKA†

In this information-oriented society that we live in, a method of retrieving necessary information is needed. Acquiring required information by information retrieval system requires the cost, in order to have to read and understand many documents shown as a candidate, therefore the present information retrieval system requires much time and effort. It is expectable that offering a node of a thesaurus that an unknown word belongs to cuts down the cost concerning a user's information retrieval sharply. This paper describes a method of mapping the unknown word to the optimal node of a thesaurus by using the concept base and web.

1. はじめに

今日，テレビや雑誌等で Web(World Wide Web)やインターネット関連のニュースが頻繁に取り上げられ, 情報収集，発信の有用な手段として注目されている．

コンピュータやネットワークの専門家でない科学者向けに，インターネット上に在る様々な情報を手軽に収集可能にすることを目的として考案された Web は，繁雑なコマンドを覚えることなく Hypertext による手軽な情報へのアクセスを実現している．Web の GUI 環境はインターネット上にある情報の利用者を急速に増やしつつある. 急速な Web の流行の最大要因は，情報検索の容易さではなく，情報発信の容易さにあるともいえる．

利用人口の増大, 利用者層の広がりは, 情報量の増加，多様化を引き起こした．様々な情報が世界中に散在しており，情報量は日々増え続けている．

情報には自分にとって良い情報と悪い情報, あるいは，必要な情報と不要な情報が存在しており，自分にとって必要な情報を短時間に的確に得ることが必要となってくる．このような有用な情報を取得するために「情報検索」という研究が盛んに行われ実用化されている．

しかし，現状のシステムは，ユーザのキーワード入力に対する関連文書群の提供という形で，情報の発見や理解をある程度は支援しているが，膨大な関連文書群から必要な文書を見つけ出し，その文書を読んで情報を理解し検索要求を満たす作業は全てユーザが行っているため，ユーザにとって現状の情報検索システムを使うことはかなりの負担である．

本研究では，この問題点を語の知識を蓄えた概念ベース[1]を用いた検索システムの知的化により，ユーザの理解支援を行うことで負担を軽減する．

図１のようにユーザの知らない語（未知語）の検索要求に対して，その語が属するシソーラス[5]のノードを提示し，ユーザの理解を支援することができるシステムの開発を行い，情報検索の利用者に対して負担の少ないシステムを構築する手法を本稿で提案する．

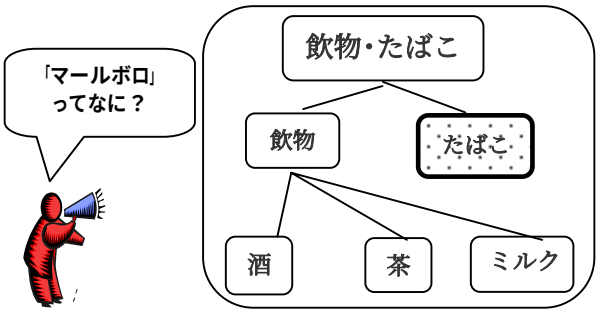

図１ シソーラスのノード提示

†同志社大学大学院　工学研究科
Graduate School of Engineering, Doshisha University

2. 概念連想メカニズム

2.1 シソーラス

シソーラスとは一般名詞の意味的用法を表す 2710 個の意味属性（ノード）の上位－下位関係，全体－部分関係が木構造で示されたものである．ノードに属する名詞として約 13 万語（リーフ）が登録されている．

本研究では，入力された未知語がシソーラスのどのノードに属するのが最も良いかを特定し，そのノードをユーザに提供する．シソーラスの 2710 ノードのうち，未知語の属すノードを探す上で必要のないノードを人手で削除している（図 2）．その結果使用するノード数は 369 個となっている．

ノード	必要・不必要
医師	必要
初心者	不必要
先輩	不必要
凡人	不必要
学者・研究者	必要
・	・
・	・

図 2 シソーラスノードの一部

2.2 概念ベース

ある語 A をその語と関連の強いと考えられる語 a_i と重み w_i の対の集合として定義する．

$$A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_m, w_m)\}$$

ここで，a_i を１次属性と呼ぶ．また便宜上，A を概念表記と呼ぶ．このような属性の定義された語（概念）を大量に集めたものを概念ベースと呼ぶ．ただし，任意の１次属性 a_i は，その概念ベース中の概念表記の集合に含まれているものとする．すなわち，属性を表す語もまた概念として定義されている．したがって，１次属性は必ずある概念表記に一致するので，さらにその１次属性を抽出することができる．これを２次属性と呼ぶ．概念ベースにおいて，「概念」は n 次までの属性の連鎖集合により定義されている．

本研究では，複数の国語辞書から，各見出し語を概念表記，その見出し語の説明文中の自立語を１次属性として抽出し，出現頻度に基づく重みを付加した約４万の概念からなる概念ベースを基に，新聞などから抽出した概念表記や属性を加え，質の向上を目的にした精錬操作（属性の追加・修正など）を施し，更に，概念間に成り立つ一般的なルールに基づく適切な重みを付加した約 10 万の概念からなる概念ベースを構築し[8]使用して，未知語を表す単語の集合（未知語属性）の作成時および，シソーラスのノードを表す語の集合（ノード属性）の作成時に利用している．

2.3 重み付き関連度計算方式

重み付き関連度計算方式[3]とは，関連の種類しか特定できない意味ネットワークのようなものとは違い，概念と概念の関連の強さを定量的に評価するものである．関連度計算方式のアルゴリズムを以下に示す．

概念 A と概念 B を

$A = \{(a_i, u_i) \mid i=1\sim L\}$

$B = \{(b_j, v_j) \mid j=1\sim M\}$

と定義すると概念 A と概念 B の重み付き一致度：MatchW(A,B)は次式で表される．

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

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

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

概念 A と概念 B の関連度：ChainW(A,B)は以下のアルゴリズムで求める．

(1)属性数の少ない概念を A とし（$L \leqq M$），概念 A の属性の並びを固定する．

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

(2)概念 B の各属性に対応する概念 A の各属性との重み付一致度 MatchW(a_i,b_{xi})の合計が最大になるように並び替える．

$$B = \{(b_{x_1}, v_{x_1}),(b_{x_2}, v_{x_2}),\cdots,(b_{x_L}, v_{x_L})\}$$

ただし，対応にあふれた概念 B の属性

$$(b_{x_j} \mid j = L+1, \cdots, M)$$

は無視する．

(3)概念 A と概念 B の関連度を次式で定義する．

$$ChainW = (s_A / n_A + s_B / n_B)/2$$

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

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

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

本研究では，未知語属性とシソーラスのノード属性との関連の深さを判断するのに重み付き関連度計算方式を用いている．

2.4 概念ベース *idf*

属性の重み付け手法である概念ベース *idf*[8]とは，特定の概念で出現する属性は概念にとって重要であるという情報検索で用いられる *idf* の考え方を利用する重み付け手法である．

概念ベース *idf*(*Gidf*)は以下の式で求める．

$$Gidf(t) = \log \frac{N_{All}}{df(t)}$$

t : 概念

N_{All} : 概念ベースの全属性数

$df(t)$: 概念ベース内での概念　tの出現回数

本研究では，この重み付け手法を Web より構築する未知語属性の重み付けに用いている．

3. シソーラスノードへのマッピング

3.1　Web からの未知語属性の取得

(1)入力された未知語をキーワードとして検索エンジンである Google[6]を用いて検索を行う．その結果返答された上位１０文書の内容を取得し，HTML タグなどの不要な情報を取り除いた文書群の形態素解析[7]を行う．

(2)インターネットに頻出する語群を事前に機械的に調べ登録しておいたリスト（ストップリスト）によってフィルタをかけ, 残った語を頻度順に並べ替えたものから概念ベースに存在する語のみを, その未知語の属性とし，頻度を属性の重みとする．

(3)属性の頻度に概念ベース内の *idf* である *Gidf* の値を頻度重みに掛け合わせ, それを最終的な属性の重みとし，重み順に並び替える．

これにより入力された未知語を表す単語の集合である未知語属性が構築できる．図 3 に未知語属性の例を示す．

MRSA		アミノ式	
感染	298	燃焼	175
院内	291	系	174
病原性	235	運動	162
患者	211	曲	138
防止	175	CM	112
メチシリン	172	ドリンク	80
対策	160	ペットボトル	79
消毒	154	飲料	61
マニュアル	152	アミノ酸	45
・	・	・	・
・	・	・	・

図 3 「MRSA」「アミノ式」の未知語属性

※Google での検索結果の文書が 1 件以下の場合, 処理は打ち切る．

3.2　シソーラスのノード動詞の取得

シソーラスは作成者がある分類基準に従って単語を体系的に配置したものである．そのためシソーラスには「同一のノードに属すリーフは助詞を伴う動詞(サ変名詞を含む）の係り受けに同様の語をとる」という関係が存在する．この関係を利用して，入力された未知語がどのノードの分類基準と類似しているのかを調べることでノード推定の補助を行う．

そのため, 各ノードにあらかじめ「助詞を伴う動詞」をノード動詞として登録しておくことが必要である．以下にノード動詞の構築方法を示す．

(1)対象とするノードに属しているリーフをすべて抜き出す．

(2)そのリーフ各々をキーワードとして Google で検索し, それぞれについて 1000 件分の Google の検索結果の見出しページの文書を取得する．そしてその文書内でリーフの直後に出てきている「格助詞+動詞（サ変名詞を含む)」の部分をすべて抜き出し，足し合わせる．

(3)それらをすべてのノードに対して行い，ノードごとに頻度順に並び替えたものをノード動詞属性とする．

(4) 後に示す手法を用いてノード動詞全体での各属性の *idf* を求め，頻度と掛け合わせて *tf・idf* 値を求めその値を属性の最終的な重みとする．

図 4 に構築したノード属性を示す．

歌手		医薬品 (薬用)	
が歌う	231	を服用する	183
が歌唱	183	が効く	175
を歌う	183	を処方する	113
を聴く	142	を飲む	77
が出演	140	を購入	75
・	・	・	・
・	・	・	・
・	・	・	・

図 4 「歌手」「医薬品（薬用)」のノード動詞

3.3　シソーラスのノード属性の取得

入力された未知語の属性は Web の検索を用いて構築したが，比較対象であるシソーラスのノードがそのままの状態では，比較が困難である．そのためシソーラスの各ノードの属性を取得する．

(1)シソーラスの各ノードにおいて，ノードに属するリーフ各々に対して概念ベース参照を行い，リーフを概念とする語の一次属性とその重みを取得し，それらを足しあわせて属性集合を取得する．

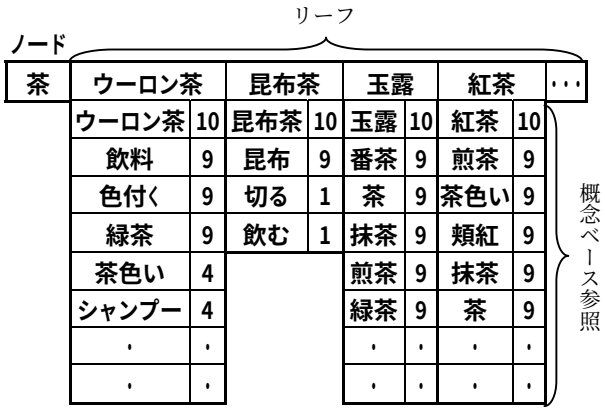

図 5 ノード「茶」の属性構築手順

(2)すべてのノードにおいてこの処理を行いシソーラスのノード属性を取得する．

(3)この取得したシソーラスの全ノード属性内で，後に示す手法を用いて各属性の *idf* を求め，重みに特定性情報を付加する．これを「*tf・idf* 重み付きノード属性」とする．

図 6 にシソーラスのノードのひとつである「茶」のノード属性を示す．

ノード属性	重み
緑茶	2371
煎茶	2250
紅茶	1896
抹茶	1744
番茶	1552
茶色	1418
'	'
'	'

図 6 ノード「茶」の属性

3.4 *tf･idf* による属性の重み付け

あるノード $node_i$ を以下のように定義する．

$$node_i = \{(d_1, w_1), \cdots, (d_L, w_L)\}$$

$node_i$ は語 d_j とその重み w_j の対の集合であり，L はその個数である．この語 d_j を属性という．

検索対象とするノード数が M で属性の異なり数が N であった場合，M 個のノードを以下のような行列ができる．

$$matrix = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ w_{M1} & w_{M2} & \cdots & w_{MN} \end{bmatrix}$$

w_{ij} はノード $node_i$ における属性 d_y の重みを表す．M 個のノードを N 個の属性で表現した $matrix$ の要素 w_{ij} はノード $node_i$ における属性 d_y の重みを表すが，その重み w_{ij} は以下の計算式によって得られる．

$$w_{ij} = tf_i(d_y) \cdot idf(d_y)$$

属性の取得時，$tf_i(d_y)$ はノード $node_i$ における属性 d_y の各重みを総和したものである．$idf(d_y)$ は属性 d_y が出現するノード数によって決まり，以下の式によって定義される．

$$idf(d_y) = 1 + \log(\frac{M}{M_{d_y}})$$

ただし，M_{d_y} は属性 d_y が出現するノード数である．頻度 $tf_i(d_y)$ と属性の網羅性 $idf(d_y)$ によって計算されることから，この重み計算方式を一般に *tf･idf* 重み付けと言う．

この重み付け手法を「シソーラスのノード動詞」および「シソーラスのノード属性」に用いている．

3.5 シソーラスのノード特定

(1)入力された未知語をキーワードとして Google で検索を行い，未知語属性とその頻度を取得する．その後，構築した属性の頻度に *Gidf* を掛け合わせたものを未知語属性の重みとする．

$$query = \{(q_1, w_1), (q_2, w_2), \cdots, (q_K, w_K)\}$$

q:属性　w:重み　K:個数

(2)シソーラスのノード集合を以下のように定義する

$$NODE = \{node_1, node_2, \cdots, node_{369}\}$$

未知語属性 *query* とシソーラスの各ノード属性 $node\_noun_i$ の間で重み付き関連度計算 *RelValue* を求める．

$$node\_noun_i = \{(n_{i1}, w_{i1}), \cdots, (n_{iX}, w_{iX})\}$$

n:属性　w:重み　X:個数

$$\mathrm{Re}lValue(node_i) = ChainW(query, node\_noun_i)$$

関連度が 0.01 以上のものを正解候補のノード集合 *NODE'* として選び出す．

$$NODE' = \{node_x \mid \mathrm{Re}lValue(node_x) > 0.01, node_x \in NODE\}$$

※関連度 0.01 以上という値は，未知語の入力に対して正解候補ノードの集合である(*NODE'*)が 30 個以下になる値を実験により求めたものである．ここでノードを 30 個以下に絞込みをしておくのは，次のノード動詞を用いた処理において多大な通信時間が掛かるのを防ぐためである．

(3)ノード i におけるノード動詞 $node\_verb_i$ は以下のようになる．

$$node\_verb_i = \{(v_{i1}, w_{i1}), \cdots, (v_{iY}, w_{iY})\}$$

v:属性（動詞）　w:重み　Y:個数

正解候補ノードにおいて，各ノードのノード動詞を参照し，重みが高いもの上位 3 つまで取得する．そして入力された未知語に取得した動詞を連結して検索キーワードをつくり Google で検索を行い，文書の HIT 数を動詞ごとに足し合わせその平均を求める．

$$VerbHit\ (node_i) = \frac{\sum_{a=1}^{3} [\text{" } query \text{ "+" } v_{ia} \text{ "}] \text{の HIT 数}}{3}$$

HIT 数が 20以上の場合は　20とする

※　入力語が「マルボロ」ノード動詞が「を吸う」のときは「マルボロを吸う」という検索キーワードとして検索して該当文書数を得る．

※　上位 3 つまでの動詞という制限は，通信時間の短縮を目的としている．本来であればすべてのノード動詞を使って平均を求めることが望ましい．

(4)以下の計算式を用いて各ノード得点を求める．

$$NodeValue\ (node_i) = \mathrm{Re}lValue\ (node_i) + 0.0025 \cdot VerbHits\ (node_i)$$

(5)*NodeValue* の値が 0.02 以上のノードを新たな正解候補ノードとする．

$$NODE'' = \{node_x \mid NodeValue\ (node_x) > 0.02, node_x \in NODE'\}$$

(6)未知語属性 *query* の (q_1からq_K) に対して，其々を概念ベースで検索し属性を取得する．そして正解候補ノード集合(*NODE*)と (q_1からq_K) の全ての組み合わせに対して関連度計算を行い，一度も 0.02 以上の値を一度も得られない未知語の属性を，不要な属性として削除し，新しい未知語属性 *query'* を求める．

$$query' = \{(q_1, w_1), (q_2, w_2), \cdots, (q_N, w_N)\} \quad (N \leq K)$$

(7)不要な属性を削除した新しい未知語属性 *query'* と不要なノードを削除した新しいノード集合 *NODE"* を用いて，(2)と(5)で使う閾値を 0.02 ずつ大きくしながら(2)から(7)の処理を繰り返す

(8)正解候補ノード集合 *NODE"* が空集合になるまで処理を行い，空集合になる直前の *NODE"* に属しているノードの中で最も *NodeValue* の値が高かったものを正解ノードとする．

4. 評価方法

評価を行うために，20 人から各 10 個ずつ自分は知っているが，自分以外には知っている人が少ないと思われる単語と，それが一言で言うと何なのかという情報を記入してもらい作成した合計 200 個の擬似的な未知語を用いてシステムの評価を行う．評価に使用した未知語の一部を表 1 に示す．

表 1　テストセットの一部

クラウンマジェスタ	ザ･マッカラン	いぬがらし
クロウシノシタ	クイニーアマン	MRSA
マールボロライト	コスモクロア	プロゲスチン
カウス･ボレアリス	アロンアルファ	横山大観

テストセットの各未知語の入力に対して，システムが返答したシソーラスのノードを人手で吟味し，正解か不正解かを評価する．

5. 結果

まず未知語の属すノードを特定するのに不必要なシソーラスのノードを削除した効果を図 7 に示す．未知語属性の重みは頻度情報である *tf* のみであり，未知語の属すノードの決定手法は重み付き関連度方式を一度だけ用いて最高関連度のノードを選んでおり，ノード動詞は用いていない．

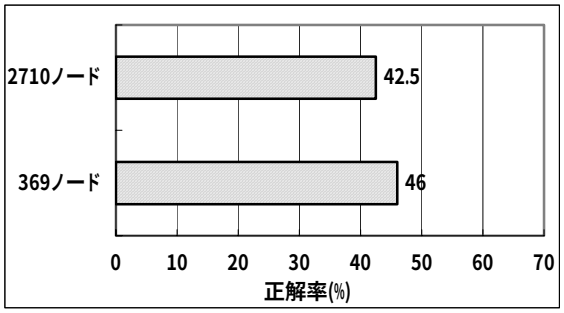

図 7 ノード数の変化による評価

ノード数が約８分の１になっているが，元々不必要と思われるノードを人手で削ったため， 42.5%から 46.0%とあまり精度の向上にはつながらなかった．しかし計算時間の面では大幅な速度向上につながった．

次に上と同様の条件で未知語属性の重みを頻度のみから, 頻度に *Gidf* を掛けたものに変更したときの評価を示す．シソーラスのノード数は削減した後の 369 ノードのシソーラスを用いている（図 8）．

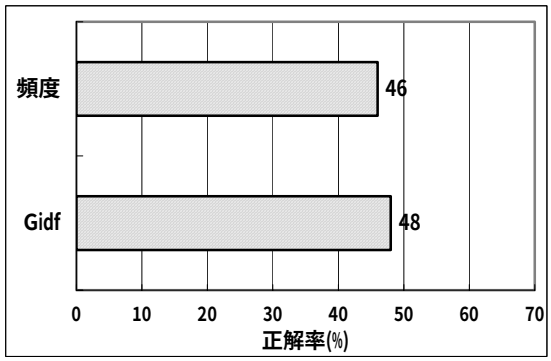

図 8 *Gidf* による重み付けの評価

Gidf 重み付けにより未知語属性の重みを変化させたが，精度の向上は 2%であった．根本的に属性の雑音が多いため, その雑音の重みを下げるというよりは, 雑音を取り除く手法がさらに必要であると思われる．

次にノード動詞の効果および繰り返し処理の効果について評価する（図 9）．シソーラスのノード数は削減した後の 369 ノードのシソーラスを用いており，繰り返し処理に関しては，ノードの絞り込みと未知語の絞込みを繰り返して行ってノードを特定する．

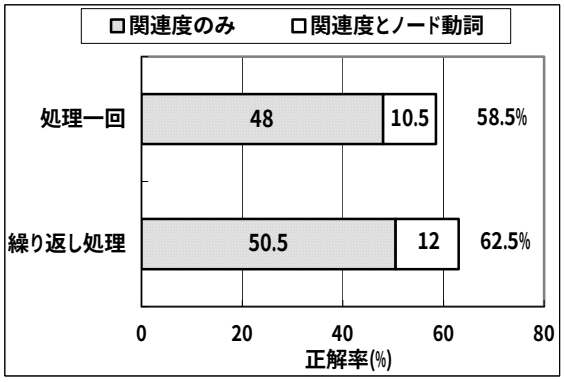

図 9 ノード動詞および繰り返し処理の評価

重み付き関連時計算に加えてノード動詞を用いた手法では, 重み付き関連時計算単独のときに比べ約 10%ほどの精度の向上が見られた．

ノード動詞はノード「医薬品」であれば「を服用する」のように，そのノードに属す語のみにほぼ固有に接続できるノード動詞をいかに上手く登録できているかがポイントとなる．ノード動詞間で *idf* を行い固有に使われる動詞に大きい重みが振られたため良い結果が出たと考えられる．

次に繰り返して処理を行ったときの属性と正解候補ノードの結果を図 10 と図 11 に示す．

残った属性		削除した属性	
目薬	47	ソング	51
目	38	アイリス	29
クール	29	場	25
健康	28	ネオ	21
疲れ	21	スマイル	19
コンタクト	15	アール	19
眼	14	モード	17
・	・	・	・
・	・	・	・

図 10 「サンテ FX」（目薬）の未知語属性

始めの正解候補ノード		最後の正解候補ノード	
ノード	重み	ノード	重み
歌手	0.03376	薬品類（医用）	0.06167
草花・野草	0.03191	眼鏡	0.04610
薬品類（医用）	0.02747	身体障害	0.05098
身体障害	0.04374		
眼鏡	0.04147		
和歌	0.03943		
楽曲	0.03414		
・	・		
・	・		

図 11 「サンテ FX」の正解候補ノード

繰り返し処理を行い，未知語属性と正解候補ノードの不要な要素を削除していった結果，精度がさらに約 2%向上した．

Web から取得する属性は多くの雑音を含んでいるが，その雑音が処理によって削減されたことが精度の向上に寄与した．図 10 を見て解るように未知語属性として相応しくない「アール」や「モード」などが削除されている．

またそれに伴って正解候補属性においても正解ノードとして相応しいノードである「医薬品（薬用）」や正解に近いノードである「眼鏡」などのノードに大きい値が与えられている．（図 11）

最終的に精度は 62.5%となったが，間違ってしまったものの大部分は，入力された語の属性を上手く取得できなかったことが原因である．どんなに様々な処理を行っても，元々Web から取得した属性にまったく正しい属性が含まれていないような状況では正解のノードに導くことが不可能である．

雑音の多い検索対象である Web から如何に入力語に相応しい属性を取得するかが精度向上の鍵である．

ノード動詞に関しては，通信時間の都合上，重みが大きいもの上位 3 件までの検索を行い，その平均を値としているが，本来はもっと数多くの動詞を使用するべきである．精度と実行時間を考えながら，更なる考慮が必要である．

また現在一つのノードしかユーザに提示を行っていない．例えば「田中康夫」さんが入力されたとき，正解ノードは「作家」「政治家」「長」などが考えられる．

現在は正解とされるノードの中の一つが提示されれば正解としているが，これらのノードをすべてユーザに提供することができれば最も良い．しかし閾値で提示するノードを複数選ぶと間違ったノードが数多く提示されてしまう．このような問題点を解決するのが今後の大きな課題である．

6. おわりに

本論文では，ユーザが未知の単語に出会い，その単語が一体何なのかを知りたいという知的欲求を持ったとき，ユーザにその語が大局的に見て何なのかを提示することができる手法を提案した．これにより，知的要求を持ったユーザが未知の語が何なのかを理解し，その時点で知的要求を十分満足させられたケースに対しても，またその時点でまたもっと詳しく知りたいという要求を持ったユーザに対しても，その語が何なのかを理解するために文書を読む手間と時間を削減することができる．これによりユーザに使いやすい情報検索システムの実現につながるであろう．

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における研究の一環として行ったものである．

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,No7,pp.1272-1283,1997
[2]入江毅,渡部広一,河岡司,松澤和光:”知的判断メカニズムのための概念間の類似度定量化方式”, 信学技報, pp.37-44,2000
[3]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, pp.39-54, 2001
[4]徳永健伸:情報検索と言語処理,東京大学出版会, 1999.
[5]NTT コミュニケーション科学研究所監修,日本語語彙体系,岩波書店,1997
[6] Google http://www.google.co.jp
[7] chasen http://chasen.aist-nara.ac.jp/
[8] 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, NLC2001-93, pp.109-116,2002