E-002

質問に的確に回答する知的 Web 検索システム

Intelligent Web Search System to Answer Questions Exactly

源明和也† 渡部広一† 河岡司†
Kazuya Genmei Hirokazu Watabe Tsukasa Kawaoka

1. はじめに

現在，Web の急速な普及と利用者の劇的な増加に伴い，膨大な電子化された文書がオンライン上に蓄積されている．その膨大な電子化された文書は，新聞記事に留まらず，電子辞書や Blog に代表されるような多種多様な文書が電子化されている．日常会話に出現するような語句に関する情報は，ほぼ確実に Web 上に存在している．しかしながら，既存の検索システムでは，膨大な量の情報に対し，要求する情報が埋もれてしまう状況にあり，的確にユーザが必要としている情報のみを獲得するための技術が現在求められている．

この技術として，質問文応答システムが提案及び研究されている．このシステムは，ユーザの入力する質問文「2008 年のオリンピックはどこで行われるか」に対して，正しい答えである「北京」の出力を目標としている．一般的な質問文応答システムは，質問文を端的に説明する文章，すなわち文書の要約に重点を置き，文章の構文解析を用いて回答を抽出している．そのため，対象文書が多くなると処理に手間を要する．

本稿では，質問文応答システムとして，質問文に対する答え（単語）を Web から自動的に獲得する手法（知的 Web 検索システム）を提案する．提案手法では，Web から取得する自立語群（属性集合）を用いて答えを選択する．文書集合から質問文の意味的特徴を単語の集合で表すことで，質問文を一つの単語と捉える．さらに，文章ではなく，単語のみに注目しているため，既存の質問文応答システムに比べ，対象文書の量による影響が少ない．なお，質問文としては，人が Web 検索の際に主に検索する対象となる語（検索対象語）として「場所」とそれ以外の質問文を想定し，それぞれに対して別アプローチを試み，それぞれで評価を行った．この手法によりユーザの情報検索のための手間と時間を省くことが期待できる．

2．関連技術（概念ベースと関連度計算）

概念ベースと関連度計算を用いる利点として，単語間の近さを表記一致ではなく，意味を考慮して定量的に評価できることが挙げられる．

2.1 概念ベース

概念ベース[1]とは，複数の国語辞書や新聞等から機械的に構築した，語（概念）とその意味特徴を表す単語（属性）の集合からなる知識ベースである（図 1）．概念 A に付与される属性 a_n には，その重要性を表す重み w_n が付与されている．

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_n,w_n)\} \quad (2.1)$$

概念ベースには，約 9 万語の概念が収録されており，1 つの概念あたり平均 30 個の属性が付与されている．しかしながら，概念ベースにも登録されていない概念も存在しており，その概念を本稿では未定義語と定義する．

各概念に付与されている属性は，概念ベースに概念として登録されている語であるため，各属性を一つの概念表記としてみなした場合，さらにそれを表す属性を導くことができる．このように，概念は概念ベースにより n 次の属性連鎖集合として定義する．また，n 次の属性集合を n 次属性と呼ぶ．

概念	属性／重み
雪	（雪／0.61），（白い／0.30），…
白い	（雪／0.16），（白地／0.14），…

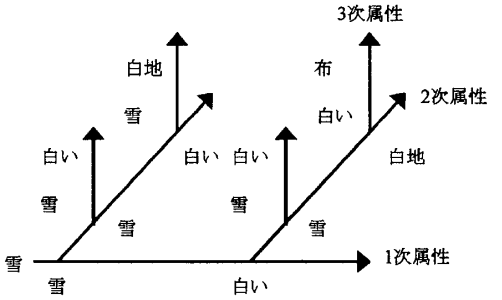

図 1 概念ベース

2.2 関連度計算方式

関連度計算方式[2]は，概念ベースに定義された語と語の関連の強さを，同義性，類似性のみに関わらず定量化する手法である．より柔軟に語と語の関連の強さを定量化するために関連度計算方式を前提としている．以下，概念間の一致度，並びに一致度に基づき関連度を求める関連度計算方式について述べる．

2.2.1 一致度

概念 A，B の属性を a_i，b_j，対応する重みを u_i，v_j とし，それぞれ属性が L 個，M 個あるとする．($L \leqq M$)．また，各概念の属性の重みを，その総和が 1.0 となるよう正規化している．

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_L,w_{Ln})\} \quad (2.2)$$

$$B=\{(b_1,v_1),(b_2,v_2),\cdots,(b_M,v_M)\} \quad (2.3)$$

このとき，概念 A と概念 B の属性一致度 $MatchWR(A,B)$ を以下のように定義する．ただし，$a_i=b_j$ は属性同士が一致した場合を示している．すなわち，一致した属性の重みのうち，小さい方の重みの和が一致度となる．また，一致度は 0.0~1.0 の値をとる．

$$MatchWR(A,B)=\sum_{a_i=b_j}\min(u_i,v_j) \quad (2.4)$$

2.2.2 関連度

概念関連度 MR は，対象となる二つの概念において，一次属性の組み合わせについて一致度を求め，これを基に

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

概念を構成する属性集合全体としての一致量を計算することで算出される．

具体的には，見出し語として一致する属性同士($a_i=b_j$)について，まず優先的に対応を決定する．他の属性については，全ての一次属性の組み合わせにおいて属性一致度を算出し，属性一致度の和が最大となるように組み合わせを決定する．一致度を考慮することにより，属性同士の見出し語としての一致だけではなく，一致度合いの近い属性を有効に対応づけることが可能となる．また，概念 A，B 間の見出し語として一致する属性($a_i=b_j$)については，以下の処理により別扱いとする．$a_i=b_j$ なる属性があった場合，それらの属性の重みを参照し，$u_i>v_j$ となる場合は，a_i の重み u_i を u_i-v_j とし，属性 b_j を概念 B から除外する．逆の場合は，同様に b_j の重み v_j を v_j-u_i とし，属性 b_j を概念 B から除外する．見出し語として一致する属性が T 組あった場合，概念 A，B はそれぞれ A'，B' として以下のように定義し直され，これらの属性間には見出し語として一致する属性は存在しなくなる．

$$A'=\{(a'_1,u'_1),(a'_2,u'_2),\cdots,(a'_{L-T},u'_{L-T})\} \quad (2.5)$$

$$B'=\{(b'_1,v'_1),(b'_2,v'_2),\cdots,(b'_{M-T},v'_{M-T})\} \quad (2.6)$$

見出し語として一致した属性の関連度を $MR\_com(A,B)$ とし，以下の式で定義する．

$$MR\_com(A,B)=\sum_{a_i=b_j}\min(u_i,v_j) \quad (2.7)$$

次に，見出し語として一致する属性を除外した A'，B' の関連度を $MR\_def(A',B')$ とする．$MR\_def(A',B')$ を算出するために，属性数の少ない方の概念 A' の並びを固定し，属性間の属性一致度の和が最大になるように概念 B' の属性を並べ替える．このとき，対応にあふれた属性は無視する．概念 A' の属性 a'_i と概念 B' の属性 b'_x が対応したとすると，概念 B' は以下のように並び換えられる．

$$B'=\{(b'_x,v'_x),(b'_{x+1},v'_{x+1}),\cdots,(b'_{x+L-T},v'_{x+L-T})\} \quad (2.8)$$

そして，見出し語として一致する属性を除去した属性間の関連度 $MR\_def(A',B')$ を以下の式で定義する．

$$MR\_def(A',B')=\sum_{s=1}^{x+L-T}MatchWR(a'_s,b'_s)\times\frac{\min(u'_s,v'_s)}{\max(u'_s,v'_s)}\times\frac{u'_s+v'_s}{2} \quad (2.9)$$

このように，見出し語として一致する属性間の関連度 $MR\_com(A,B)$ と，それら以外の属性間の概念関連度 $MR\_def(A',B')$ をそれぞれ算出し，合計を概念 A，B の関連度 $MR(A,B)$ とする．

$$MR(A,B)=MR\_com(A,B)+MR\_def(A',B') \quad (2.10)$$

関連度も，一致度と同様 0.0~1.0 の値をとる．

3. 関連技術（Web 関連）

本稿において，未定義語の属性獲得手法（3.3，3.4）が重要である．未定義語の属性獲得手法では，検索結果である Web のテキスト情報の単語にのみ注目する．そのため，対象文書の量が増大した場合でも影響は少ない．

3.1 TF・IDF

TF・IDF 法[3]とは，語の頻度と網羅性に基づいた重み付け手法である．TF はある文書中 d に出現する索引語 t（文書の内容を表す要素）の頻度 $tf(t,d)$ を表す尺度である．IDF はある索引語が全文書中のどれくらいの文書に出現するかを表す尺度であり，N を検索対象となる文書集合中の全文書数，$df(t)$ を索引語 t が出現する文書数とすると式 3.1 で定義される．

$$idf(t)=\log\frac{N}{df(t)}+1 \quad (3.1)$$

3.2 Web-IDF

3.1 で説明した IDF は一般的な文書（新聞や書籍など）を用いて索引語の特定性を考慮する手法である．IDF の中でも特に，Web-IDF[4]は Web にある文書のみを用いて索引語の特定性を考慮する手法である．Web-IDF では式 3.1 の N を Google[5]が保有している日本語のページ数（Google は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ため，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（1,010,000,000）を Google が保有している日本語の全ページ数としている），$df(t)$ を索引語 t を Google で検索を行ったときのヒット件数とする．

3.3 未定義語の属性獲得手法

未定義語の属性獲得手法[4]とは，未定義語 X（概念ベースに定義されていない概念）の意味的特徴を表す属性（単語）とその重要性を表す重みの組を Web を用いて自動的に構成する手法である．まず，ロボット型検索エンジン[5]を用いて検索を行って獲得したテキスト情報から形態素解析[6]を行い自立語を出現単語として抽出する．その後，獲得したテキスト情報空間内での出現単語の出現頻度と Web-IDF の算出を行い，TF・Web-IDF 重み付けを行う．重み順に上位から自立語とその重みの対の集合を X の属性とする．この手法を用いて未定義語 X の属性とその重みの組を構成する．未定義語 X の属性は式 3.2 のように構成される．

$$X=\{(x_1,w_1),(x_2,w_2)\cdots,(x_n,w_n)\} \quad (3.2)$$

本稿では，この未定義語の属性獲得手法をオートフィードバック（Auto Feedback：AF）と呼ぶことにする．

なお，AF は概念ベース内に存在している語のみを属性として獲得する．すなわち，未定義語に対して AF を用いることで，概念ベースで定義されている語とみなせるので関連度計算が可能となる．

3.4 拡張 AF

拡張 AF は 3.3 を拡張した手法である．Web のテキスト情報から形態素解析を行い獲得した自立語に対して以下のルールを適用した上で出現単語として抽出する点が AF と異なる．

- 括弧（「」）の中の語を複合する
- 名詞の連続は複合する
- アルファベットの連続は複合する

このルールを用いて，拡張 AF は概念ベースに存在しない語（複合語など）も属性として獲得できる．本稿では，この属性を固有属性と呼ぶことにする．

AF の属性が概念ベースに定義されている語のみなのに対し，固有属性では制限がないため，幅広い語を属性として獲得できる．これにより，固有属性の中から答えを見出すことが可能となっている．

3.5 シソーラスマッピング

シソーラスマッピング[7]とは，未定義語が大局的にどのような意味を持つのかを，その語の所属するべきシソーラス[8]のノードを提示する手法である．未定義語のシソーラスマッピングは未定義語属性の獲得，未定義語と各ノードの比較に分かれる．未定義語の属性は 3.3 の手法を用いて獲得する．未定義語と各ノードの比較に 2.2 の関連度計算を用いる．なお，各ノードの属性は概念ベースを利用してあらかじめ定義している．

4. 知的 Web 検索システム

知的 Web 検索システムの流れは図 2 の通りである．

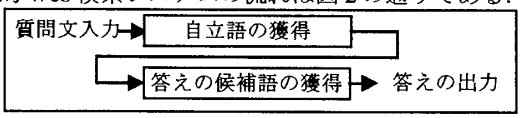

図 2　知的 Web 検索システムの流れ

まず，入力された質問文から自立語を獲得する．次に，自立語を用いて質問文の答えの候補語の獲得を行う．なお，現状のシステムとしては獲得した答えの候補語を答えとして出力する．

なお，答えの候補語の獲得に関して，検索対象語が「場所」である場合とそれ以外の場合で別処理を行っている．検索対象語が「場所」の場合に，提案した答えの特定手法の精度が低くなったためである（5.1）．

4.1 自立語の獲得

質問文の自立語は，質問文に対して形態素解析を行い，以下のルールを適用した上で獲得する．

- 括弧内の文字は一つの語とする
- 名詞，数字，アルファベットの連続は連結する
- 動詞，形容詞は基本形に変換する

このルールにより，質問文「2008 年のオリンピックはどこで行われるか」に対して「2008 年」，「オリンピック」，「行う」を獲得する．

本稿では，質問文から獲得した自立語を And 検索で用いることで質問文を未定義語とみなす．例えば，上記の質問文を未定義語とみなした場合「2008 年&オリンピック&行う」が該当する．

4.2 答えの候補語の獲得

答えの候補語の獲得の流れを図 3 に示す．

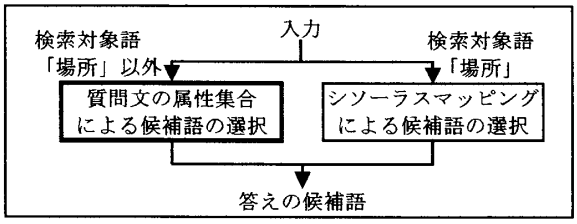

図 3　答えの候補語の獲得の流れ

図 3 に示したように答えの候補語は，検索対象語が「場所」以外の場合と「場所」の場合で別処理を行っている．

4.2.1 質問文の属性集合による候補語の選択

質問文の属性集合による候補語の選択の流れを図 4 に示す．まず，未定義語と見なした質問文を入力として拡張 AF を用いて固有属性（概念ベース内に存在しない単語）を獲得する（①）．続いて，質問文と各固有属性に対して AF を用いて概念ベース内に存在する単語のみを属性として獲得する（②）．質問文と各固有属性の関連度計算を行い，関連度が上位 5 個の固有属性を候補語として獲得する（③）．

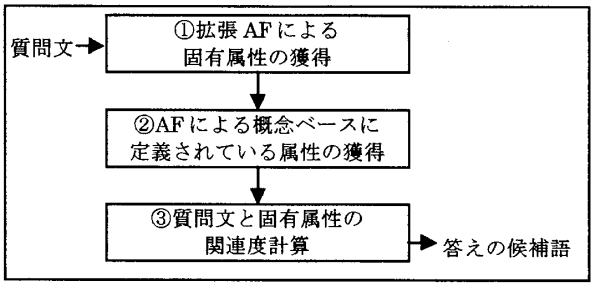

図 4　質問文の属性集合による候補語の選択の流れ

具体例として質問文「2008 年のオリンピックはどこで行われるか」を用いて説明する．このとき，入力として用いるのは「2008 年&オリンピック&行う」である．

① 固有属性として{北京オリンピック，五輪，中国，…}を獲得する．
② 質問文の属性（AF）として{(オリンピック，1398)，(北京，767)，…}，各固有属性の属性（AF）を獲得（例：「北京オリンピック」={(北京，1782)，(オリンピック，1706)，…}）する．この流れにより，質問文と固有属性の関連度計算を行うことが可能となる．
③ 質問文と各固有属性{北京オリンピック，五輪，中国，…}の関連度計算を行い，関連度上位 5 個{五輪(関連度：0.55)，北京オリンピック(関連度：0.48)，…}を候補語として獲得する．

4.2.2 シソーラスマッピングによる候補語の選択

未定義語とみなした質問文を入力として，拡張 AF を用いて固有属性（15 個）を獲得する．さらに，獲得した固有属性に対してシソーラスマッピングを行いノードに分類する．ノードは場所と場所ではない一部のノードを用いており，場所のノード（表 1）に分類された固有属性を候補語として獲得する．

表 1　場所のノード

場所	施設	公共施設
文化施設	地域（範囲）	土地
都市	村落	郷里

5. 評価

本稿では「場所」とそれ以外の語について別アプローチによる検索手法を提案しているため，それぞれに人手で作成した評価セットを準備し，結果を評価する．

5.1，5.2 では答えの候補語の獲得を行い，答えの候補語の中に質問文に対する答えが含まれる割合を評価する．

5.1 「場所」に関する評価

検索対象語が「場所」の質問文合計 30 文（表 2）を用いて，シソーラスマッピングを用いた場合と質問文の属性集合を用いた場合の候補語の選択を行い，それぞれの候補語に含まれる答えの割合を比較した．

表 2　テストセット（場所）

質問文	答え
2008 年のオリンピックはどこで行われるか	北京
本州最北端の町はどこか	大間町

シソーラスマッピングを利用した場合の精度は 76.7%（23 文/30 文），質問文の属性集合を利用した場合の精度は 50%（15 文/30 文）であった（図 5）．特に，質問文の

属性集合を利用した場合の精度はシソーラスマッピングを利用した場合に比べて低い．

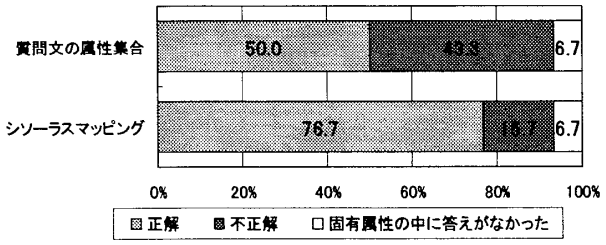

図 5　評価結果

5.2「場所」以外に関する評価

検索対象語が「場所」以外の質問文合計 136 文（表 3）を用いて質問文の属性集合を用いた候補語の選択を行い，候補語に答えが存在する割合の精度を評価した．

表 3　テストセット（場所以外）

質問文	答え
同志社大学の前身である同志社英学校を創立したのは誰か	新島襄
東北三大祭のうち青森で開催されるのは何か	青森ねぶた祭り

精度は 64.7%（88 文/136 文）であった（図 6）．しかし，固有属性の中に答えがない場合が 21.3%（29 文/136 文）であったため，固有属性上位 15 個に答えが存在していた場合の候補語の正解率は 82.2%（88 文/107 文）と非常に高いことが分かった．

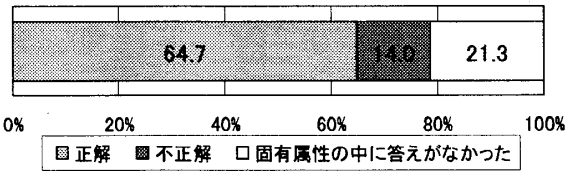

図 6　評価結果

4.3 考察

4.1 に関して，実験結果から 4.2.1 の手法よりも 4.2.2 のシソーラスマッピングの方が有効であると考えられる．

4.2 に関して，成功例として表 4 に「同志社大学の前身である同志社英学校を創立したのは誰か」と固有属性「新島襄」（答え）の獲得した属性（AF）を示す．

表 4　成功例

「質問文」の属性		「新島襄」の属性	
創立	大学	創立	旧宅
学校	設置	邸	キリスト教
前身	女学校	明治	青年
明治	…	渡航	…

表 4 のように AF により獲得する属性が非常に近い単語で構成されている．このため，質問文と答えの関連が強くなったと考えられる．固有属性として獲得できた場合に候補語として獲得できる割合が非常に高かったため，表 4 の例と同じく属性が近い場合が多いと考えられる．以上から 4.2.1 の候補語の獲得は有効であると考えられる．

固有属性として獲得できた場合でも候補語として獲得できなかった場合があった．表 5 に示したのは「総理大臣の経験者である田中角栄が逮捕された事件は何か」と固有属性「ロッキード」（答え）の属性（AF）の比較である．

表 5　失敗例

「質問文」の属性		「ロッキード」の属性	
総理	経験	事件	立花
大臣	内閣	戦闘	航空機
逮捕	首相	ブレーキ	言論
事件	…	真実	…

この例の場合，固有属性「ロッキード」が求める答えの表記の一部であり，Web から飛行機に関する単語を多く獲得してしまったと考えられる．そのため，意味的に近い属性が少なくなったと考えられる．

問題点として固有属性に答えが存在しない場合も挙げられる．今後の研究課題として固有属性の獲得方法も変更する必要があると考えられる．

5. おわりに

本稿では，質問文の意味（属性集合）に着目し，それに対する答えを Web から検索する知的 Web 検索システムを提案した．提案手法では，質問文を概念として捉え，その属性を獲得し利用することで，質問文の意味に着目した．本稿では，質問文に対する答えは単語として出力したが，質問文の意味を考慮することで，質問文に対する答えを提供することができ，従来の Web ページを検索結果として提供する Web 検索よりも，ユーザが求める情報を提供することができると考えられる．

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「人間と生物の賢さの解明とその応用」における研究の一環として行った．

参考文献

[1] 奥村紀之，北川晋也，渡部広一，河岡司，“概念ベースの分析と精錬”，同志社大学理工学研究報告，Vol.46，No.3，pp.133-141，2005.

[2] 渡部広一，奥村紀之，河岡司，“概念の意味属性と共起情報を用いた関連度計算方式”，自然言語処理，Vol.13，No.1，pp.53-74，2006.

[3] 徳永健伸，“言語処理と計算5情報検索と言語処理”，東京大学出版会，1999.

[4] 辻泰希，渡部広一，河岡司，“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”，人工知能学会全国大会，2D1-01，2003.

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

[6] 松本裕治，北内啓，山下達雄，平野善隆，今一修，今村友明，“日本語形態素解析システム『茶筌』version1.0 使用説明書”，NAIST TechnicalReport，NAIST-IS-TR97007，1997.

[7] 後藤和人，渡部広一，河岡司，“Webを用いた未知語検索キーワードのシソーラスノードへの割付け手法”，情報処理学会第68回全国大会，4N-3，2006.

[8] 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.