2D1-10

概念ベースを用いた知的検索における曖昧な質問文の意味理解

Semantic Understanding of Ambiguous Question Sentence in Intelligent DB Retrieval using Concept-Base

古川 成道*1 渡部 広一*1 河岡 司*1
Furukawa narimichi 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

In this paper, the semantic understanding technique of the question sentence using the concept-base and thesaurus required for a computer in order to make it retrieve intelligently from various databases is proposed. The conventional database retrieval was able to be searched only with notation agreement. Even if an ambiguous expression of which is not written by the database is included in the question sentence from the user, it is intelligent database retrieval that a computer understands the meaning correctly and can search from a database. In other words, a computer can search also to the question sentence of very various expression with which man is exchanged daily. The semantic understanding of a question sentence becomes important for that purpose. It is considered that intelligent database retrieval is realizable by carrying out the semantic understanding of the question sentence.

1. はじめに

近年, コンピュータという存在は, 人間の日常生活や社会活動において必要不可欠なものとなった. 人間にとってコンピュータは非常に便利な機械ではあるが, 人間と自然なコミュニケーションがとれるような『知的さ』が, コンピュータには欠けている.

この『知的さ』をコンピュータが持つことにより, 人間とコンピュータが自然な会話をすることができるようになる. そのためには, 人間の持つ感情や知識などが必要となる. その知識を得る手段の 1 つとして, 表から知識を得ることを考える. 表には様々な情報や知識が含まれているため, 表を理解することによってコンピュータはより知的になると考える.

本稿では, コンピュータに, 様々なデータベースを知的に検索させるために必要な, 概念ベース[小島 2002]やシソーラス[NTT コミュニケーション科学研究所 1997]を用いた質問文の意味理解手法を提案する. 従来のデータベース検索というのは, 表記一致のみでしか検索することができなかった. それに対し, ユーザからの質問文に, データベースには表記されていないような曖昧な表現が含まれていても, コンピュータがその意味を正しく理解し, データベースから検索できるということが知的データベース検索である. すなわち, 人間が日常的に交わされるような, 極めて多種多様な表現の質問文に対しても検索を行えるということである. そのためには質問文の意味理解が重要となる. 質問文を正しく理解できなければ, 検索することは到底できない. この質問文の意味理解をすることで, 知的データベース検索を実現できると考える.

2. 知的データベース検索

表記一致型データベース検索とは異なる, ユーザからの極めて多種多様な表現の質問文に対しても検索を行うことができるのが知的データベース検索である. 知的データベース検索全体の流れは図 1 のようになる. 黒線で囲まれた部分が本稿で述べる内容である. ここで, 質問文は二者択一文を除く5W1H 文に限る.

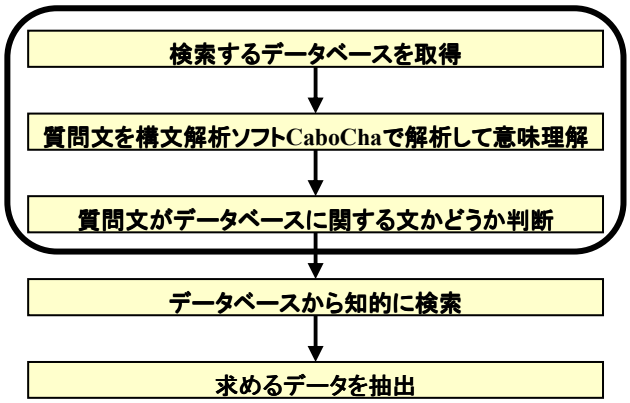

図 1 知的データベース検索の流れ

3. 概念ベース

概念ベースとは, ある単語(概念)とその意味特徴を表す属性と重みの集合で構成されたものである. 概念ベースには, 約 9 万語の概念が格納されており, 総属性数は約 254 万個である. 属性数は概念によって異なるが, 1 概念あたりの平均属性数は約 29 個である. ある概念 A に対して, その語の i 番目の属性を a_i, 重みを w_i, 概念 A の属性数を N 個とすると, 概念 A は以下のように表される.

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

また, 概念間の関連の強さは, 関連度計算[渡部 2001]を用いて 0 から 1 の実数で表される.

4. 質問文の意味理解

質問文の意味理解とは, 質問文から質問対象語とその条件を取得することである. 質問対象語とは質問文が答えを求めている語のことで, 例えば「黒い服を着ている人は誰ですか？」の

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

場合，質問対象語は「人」，条件は「黒い服を着ている」となる．人間はこの二つを瞬時に判断し，質問文の意味を理解している．よって，質問対象語とその条件を取得出来たならば，質問文の意味を理解したとみなすことができる．すなわち，この二つの要素は知的データベース検索において非常に重要となる．質問文は次のように 3 パターンに分類され，それぞれの場合によって処理が異なる．

① 疑問詞(なぜ，誰，何，どこ，いつ，…)がない文
　例:「夏によく売れる商品は？」
② 疑問詞が文末(「?」に係る語に疑問詞がある)にくる文
　例:「夏によく売れる商品は何ですか？」
③ 疑問詞が文末以外にある文
　例:「何が夏によく売れる商品ですか？」

4.1 疑問詞がない文

「？」に係る語が質問対象語となり，条件は質問対象語に直接的・間接的に係る語となる．例えば「夏によく売れる果物は？」という質問文の場合，「？」に係る語は「果物」であるので，その語が質問対象語となる．「果物」に直接的に係っているのは「売れる」であり，間接的に係っている語は「よく」と「夏」であるので，条件は「売れる」，「よく売れる」，「夏によく売れる」となる．ここで，条件を段階的に取得しているのは，図 1 においてデータベースから知的検索させる際に必要となるからである．

4.2 疑問詞が文末にくる文

4.2.1 質問対象語の取得

まず，質問対象語の取得方法について説明する．最初に，質問文の中に「何＋単位名詞」があるか判断する．「何＋単位名詞」とは「何個」，「何人」，「何円」などを表す．文章内に「何＋単位名詞」が無ければ，次に，質問文の中に「疑問詞＋名詞」があるか判断する．例えば，「最近売り上げが伸びているのはどの店ですか？」という質問文では「どの店」がこれに相当する．「疑問詞＋名詞」が無ければ，疑問詞が「何」であるか判断し，違えば疑問詞の意味が「物」であるか判断する．疑問詞の意味というのは，疑問詞が求める答えのことである．例えば，疑問詞が「誰」なら疑問詞の意味は「人物」となる．疑問詞の意味は，疑問詞知識ベースを参照することによって理解している．疑問詞知識ベースとは表 1 のような疑問詞とその意味から成る知識ベースで，レコード数は 32 件である．以上のどの処理にも相当しなければその他の処理に移る．この全体の流れを図 2 に示す．

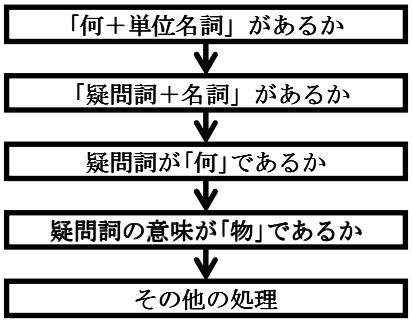

図 2 疑問詞が文末にくる文の質問対象語の取得方法

「何＋単位名詞」が文中にあれば図 3 の処理を行う．

図3において，単位の意味は表 2 の単位とその意味から成る単位知識ベースを参照する．

「疑問詞＋名詞(疑問詞付属名詞)」がある場合は，図3において，「単位の意味」→「疑問詞付属名詞」とした．

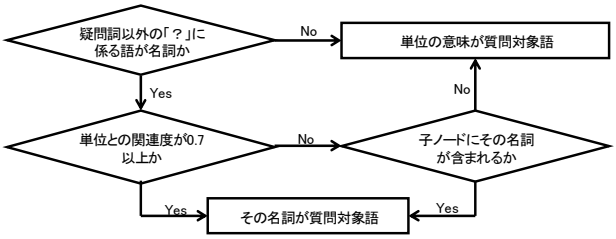

図 3「何＋単位名詞」の場合の処理の流れ

表 1　疑問詞知識ベース

疑問詞	意味
誰	人物
何処	場所
：	：

表 2　単位知識ベース

単位	意味
円	価格
枚	枚数
：	：

疑問詞が「何」である場合，「何」の前後両方に「？」に係る語があるか調べる．あれば「何」の前の語が名詞かどうか判断し，名詞であればその名詞が質問対象語となる．名詞でなければ「何」の後ろの語が名詞かどうか判断し，名詞であればその語が質問対象語となり，そうでなければ疑問詞「何」の意味である「物」が質問対象語となる．例えば「夏によく売れる商品は何という名前ですか？」の場合，「？」に係る語は「商品」，「何」，「名前」となる．「何」の前の語の「商品」が名詞なので，「商品」が質問対象語となる．

「何」の前後両方には「？」に係る語が無ければ，「何」以外の「？」に係る語が名詞かどうか判断する．名詞であればその語が質問対象語となり，そうでなければ疑問詞「何」の意味である「物」が質問対象語となる．例えば「子供に人気のあるアニメは何ですか？」の場合，「？」に係る語は「アニメ」，「何」となる．「アニメ」は名詞なので，これが質問対象語となる．この流れを図 4 に示す．

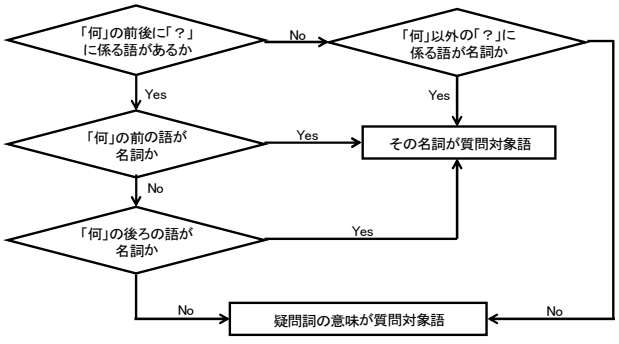

図 4　疑問詞が「何」である場合の処理の流れ

疑問詞の意味が「物」である場合，疑問詞以外の「？」に係る語が名詞ならば，それが質問対象語で，名詞でなければ「物」が質問対象語となる．例えば「女子高生が好きなお菓子はどれですか？」の場合，「？」に係る語は「お菓子」，「どれ」となる．「お菓子」は名詞なので，「お菓子」が質問対象語となる．また，「女子高生が好きなのはどれですか？」の場合，「？」に係る語は「好き」，「どれ」となり，「好き」は名詞ではないので，疑問詞「どれ」の意味である「物」が質問対象語となる．

その他の処理では，図3において「単位の意味」→「疑問詞の意味」とした．例えば「一番広い県はどこですか？」の場合，「？」に係る語は「県」，「場所」となり，疑問詞の意味は「場所」で

ある.「県」と「場所」との関連度は 0.7 未満となるが,「場所」の子ノードに「県」が含まれるので,「県」が質問対象語となる.

4.2.2 質問対象語の条件の取得

条件の取得方法は以下の 3 つの場合によって処理が異なる.

「？」に係る語の中で疑問詞ではない方の語が

1．名詞ではない
例「黒い服を着ているのは誰ですか？」
（質問対象語は「人物」）

2．名詞でその語が質問対象語
例「黒い服を着ている人は誰ですか？」
（質問対象語は「人」）

3．名詞だが疑問詞の意味が質問対象語
例「黒い服を着ている人は何歳ですか？」
（質問対象語は「年齢」）

各場合の処理方法について以下に説明する.

（1） 名詞ではない

条件は質問対象語に係る語となるが, この場合は質問対象語が質問文の中には含まれていないため係り語が分からない. 例えば「黒い服を着ているのは誰ですか？」の場合, 質問対象語は「人物」だが,「人物」という語が文章中に無いので「人物」に係る語が分からない. そこで, 質問文のうち, 意味的に名詞となる「の」という言葉を質問対象語に置換し, もう一度 CaboCha［奈良先端科学技術大学院大学 2003］で解析し, 係り語を条件とする. この例では,「黒い服を着ている人物は誰ですか？」となり, 条件は「着ている」,「服を着ている」,「黒い服を着ている」となる.

（2） 名詞でその語が質問対象語

条件は質問対象語に係る語となる. 例えば「黒い服を着ている人は誰ですか？」の場合, 質問対象語である「人」に係る語が条件となり, 条件は「着ている」,「服を着ている」,「黒い服を着ている」となる.

（3） 名詞だが疑問詞の意味が質問対象語

(1)と同じくこの場合も質問対象語は質問文の中には含まれていない. しかし, 名詞が質問対象語の条件の一つ目と思われる. よって, 条件は名詞とその名詞に係る語となる. 例えば「黒い服を着ている人は何歳ですか？」の場合, 条件は「人」,「着ている人」,「服を着ている人」,「黒い服を着ている人」となる.

4.3 疑問詞が文末以外にくる文

図 2 の『疑問詞が「何」であるか』を除いた処理を行い, 各処理に対して 4.2 と同様の処理を行う. ただし, 図 3 の『疑問詞以外の「？」に係る語が名詞か』という条件部が『「？」に係る語が名詞か』となる. 質問対象語の条件は, 疑問詞の文節を削除して再度 CaboCha で解析し, 質問対象語に係る語となる. 質問対象語が文中に含まれていない場合, 4.2 と同様の処理を行う.

4.4 質問文の意味理解の評価

307 個の質問文に対して, それぞれ出力された結果を人目で判断し, 質問対象語と条件の両方が正しく出力されていれば○, 質問対象語のみ正しく出力されていれば△, どちらも間違っていれば×とした. 評価結果を図 5 に示す.

正しく質問対象語と条件を取得できたのは全体の約 66%だった. 質問対象語に関しては, 約 90%正しく取得できた.

質問対象語が取得できなかった原因としては, 疑問詞の意味が一つとは限らないということが挙げられる. 例えば「料亭にはどれくらい行きましたか？」の場合, 質問対象語は「回数」が適切

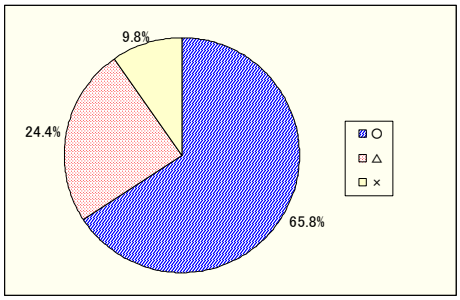

図 5 評価結果

だが, 疑問詞知識ベースには「どれくらい」の意味は「量」としているので, 質問対象語は「量」となってしまう. このように, 疑問詞の意味は全てが一つとは限らないため, How の疑問詞の意味をもう少し細かく定義する必要があると考えられる.

条件が正しく取得できなかった原因として, Cabocha の係り受けの解析ミスが挙げられる. 例えば「どこにパン屋はありますか？」の場合, 質問対象語は「場所」で, 条件は「ある」のみとなってしまう. なぜなら, CaboCha で解析すると「パン屋」は「？」に係っているからである. 我々人間は,「パン屋」は「ある」に係っていると判断する. よって, CaboCha の解析ミスが原因であると分かる.

以上のことから, CaboCha の精度向上や, 疑問詞の意味の詳細な定義をすることによって, さらに精度が上がると考えられる.

5. 質問文がデータベースに関する文かの判断

入力された質問文がデータベースに関係ない質問文の場合は, 検索する必要がないので, データベースに関する文かどうか判断する必要がある. この判断に, 4 章で取得した質問対象語とその条件を用いる.

まず, 質問対象語とデータベースの各フィールド名と関連度計算をする. この際, 質問対象語が具体であれば, 具体であるフィールド名のみと関連度計算をし, 質問対象語が抽象であれば, 抽象であるフィールド名のみと関連度計算をするようにする. そして, 質問対象語との関連度が同義語とみなされる閾値 0.7 以上のフィールドがなければ, 関連度が最も高いフィールドのレコードに, 意味的に関連が強いと考えられる閾値 0.23 以上の関連度のものがあるか, もしくはシソーラスの親ノードに質問対象語が含まれているか判断する. この条件にも満たされなかった場合は, この質問文はデータベースに関する文ではないと判断する. ただし, 質問対象語が「合計」や「平均」といった統計情報の単語の場合は, 計算を必要とする語であり, 直接検索する語ではないので, データベースに関する文と判断する. 以上の条件を満たした場合と質問対象語との関連度が 0.7 以上のフィールドがある場合は, 質問対象語の条件から名詞を抜き出す. 抜き出した各名詞に対して質問対象語と同様の処理をする. ただし, 名詞であっても統計情報の単語である場合と数字である場合は対象としない. 前述の条件を満たした名詞の数が, 抜き出した名詞の総数の半分以上であればその質問文はデータベースに関する文であると判断し, 半分以下であればデータベースに関する文ではないと判断する. 質問対象語に対する処理の流れを図 6 に, 条件に対する処理の流れを図 7 に示す.

図 7 において, 質問対象語の処理というのは図 6 の点線で囲まれた部分の処理である.

例えば, フィールド名が(商品, 値段, 売り上げ)のデータベースに対して,「果物の価格は？」という質問文が入力された場合, 質問対象語の「価格」は抽象語なので, 抽象語のフィールド

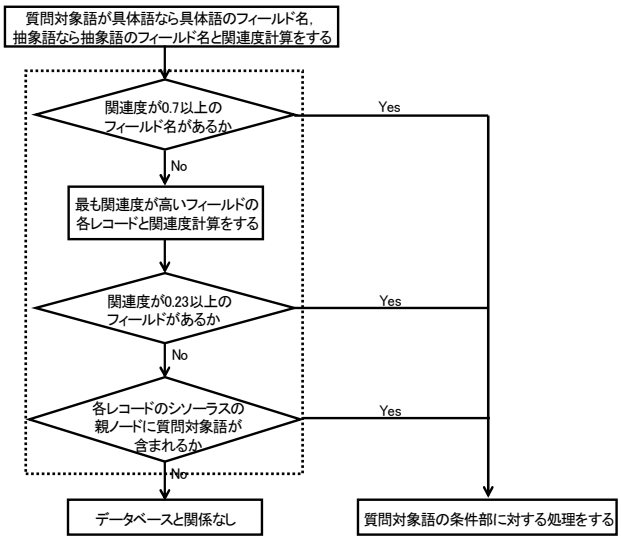

図 6　質問対象語の処理

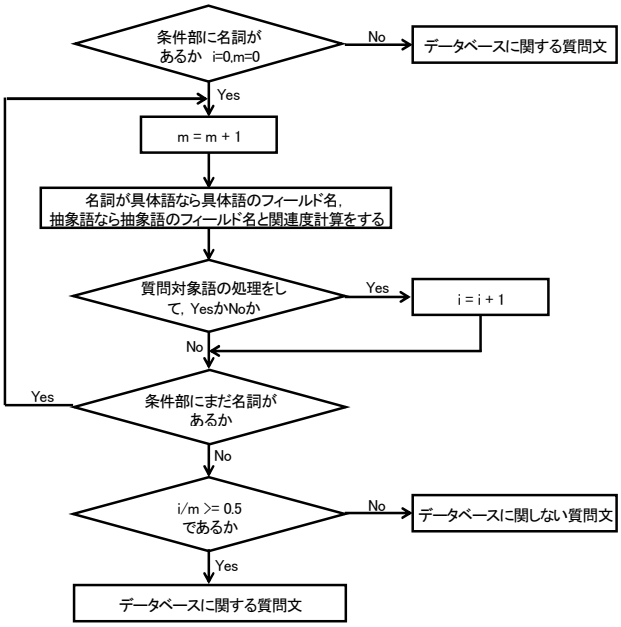

図 7　質問対象語の条件の処理の流れ

名である「値段」・「売り上げ」とそれぞれ関連度計算をして，「値段」との関連度が 0.7 以上であるので，質問対象語の条件の処理に移る．質問対象語の条件は「果物の」である．名詞は「果物」で具体語なので，具体語のフィールド名である「商品」と関連度計算をする．関連度は 0.7 未満であるので，「商品」の各レコードと関連度計算をする．レコードに「蜜柑」という語があった場合，「果物」と「蜜柑」との関連度は 0.23 以上なので，「果物」は条件を満たす．名詞は「果物」のみなので，図 7 における i/m=1 となり，この質問文はデータベースに関する質問文であると判断できる．

この処理に対して評価を行った．3 つのデータベースを作成し，それぞれに対してデータベースに関する質問文と関しない質問文を 60 個ずつ人手で作成した．データベースに関係あり・なしの判断が正しければ○，正しくなければ×とした．評価結果を図 8 に示す．

データベースに関する文では，約 67%正しく判断でき，データベースに関しない文では，約 82%正しく判断できた．

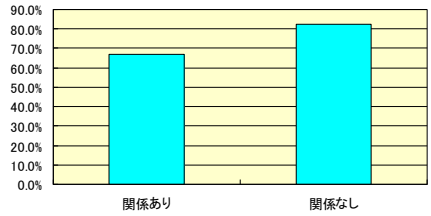

図 8　評価結果

データベースに関する文の間違いの原因としては，関連度がうまく取れないことが挙げられる．例えば，データベースのフィールド名に「店」，「面積」があり，「敷地が一番広い店は？」という質問文の場合，「敷地」と「面積」の関連度が低いため，この質問文はデータベースに関しない文と判断されてしまう．

データベースに関しない文の間違いの原因としては，動詞の意味を考慮していないということが挙げられる．例えば，データベースのフィールド名が「店」・「座席数」・「予算」・「月 日」で「店」の中に「料亭」があり，「料亭は何処にありますか?」という質問文の場合，店の所在地を聞いているので，データベースには関係ないが，名詞だけを対象として判断しているため，『関係あり』と出力されてしまう．よって，動詞も考慮する必要がある．

6.　おわりに

本稿では，ユーザからの多様で曖昧な質問文に対してもデータベースから検索できるために重要となる，質問文の意味理解手法を提案した．概念ベースやシソーラスを用いることによって，質問文から質問対象語とその条件を約 66%の精度で正しく取得できた．サ変接続に対する処理と，疑問詞の意味の詳細な定義をすることによって，さらに精度が上がると考えられる．

また，概念ベースやシソーラスを用いることによって，データベースに関する文を約 67%，関しない文を約 82%の精度で正しく取得できた．新たな知識ベースの作成・動詞の考慮などをすることによって，さらに精度が上がると考えられる．

これら二つの処理を利用することによって，知的データベース検索を実現できると考える．そして，この知的データベース検索を実現できることによって，膨大な量のデータが格納されたデータベースからでも我々人間が必要となる情報をすぐに得ることが出来るようになると考える．

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

[小島　2002]　小島一秀，渡部広一，河岡司：連想システムのための概念ベース構成法－国語辞書から抽出した概念間論理関係の利用，自然言語処理，Vol.9，No.5，pp.93-110,2002

[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，pp.39-54,2001

[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，岩波書店，1997.

[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