3C4-03

検索用知識と常識的判断システムによる知的データベース検索

Intelligent DB retrieval by the knowledge for retrieval and the commonsense judgment system

権 東旭[*1] Kwon Tongwook　京塚 浩史[*1] Kyoduka Hiroshi　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

In the conventional database, only reference from the word in the notation matching can be performed. Then, this paper proposes the intelligent system which refers to the question sentence of the natural language containing words without the notation matching. The knowledge for retrieval and various commonsense judgment system make it possible. Extraction is performed from the database of goods to the input sentence written by natural language using these.

1. はじめに

現在, 我々にとって大きな存在であるコンピュータには, 情報を知的に得るということが求められる. その手段として表というものがあり, これを利用することが出来れば, さまざまな常識的判断を用いて文章を理解することが出来る.

本稿ではユーザの要求にあったデータを正しく取り出すことが出来る知的な検索メカニズムの実現方式を提案し, その上で必要な, 自然言語による統計情報の取得方法を呈示する.

2. 知的なＤＢ検索

知的な DB 検索[1]とは入力文があいまい(DB 内には表記されていないような語を用いた文章)であっても DB 内からユーザの要求にあったデータを取り出すメカニズムである. 従来の DB 検索というのは表記一致が検索の方法となり, DB 内に存在しない語は検索できない. そこで連想メカニズム(概念ベース[2]や関連度計算[3], シソーラス[4])などを利用することで文章や単語の意味を理解しながら, 検索を行うことの出来るデータへと変換し, その取り出しを可能とする.

本稿では商品検索をするために商品 DB, 顧客 DB を対象に検索メカニズムを構築する.

それぞれの DB のフィールド名の例は以下の通りである.

商品 DB･･･商品名, 単価

顧客 DB･･･商品名, 購入月, 購入日, 購入時間, 性別, 年齢

3. 連想と常識的判断

概念ベースとは語(概念)に複数の関係のある語(属性)を持たせた知識ベースであり, 表 1 に示すようなデータが約 10 万語(概念数)格納されている.

関連度計算とはこの概念ベースを用いて二つの概念 A と B の関連性を定量的に評価するものである. 即ち, 関連度の値が概念間の繋がりの強さをあらわす.

表 1　概念ベースの構造

概念	属性
夏	肝試し, ビーチパラソル, かき氷, 冷麺･･･
医者	医者, 診察, 往診, 名医, 診る, 医師･･･
お茶	緑茶, 抹茶, 番茶, 茶道, 喫茶, 飲み物･･･

連絡先: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

常識的判断とは, 人間に近い常識をコンピュータに判断させるシステムである. 本稿では, 時間判断･感覚判断･量判断･職種判断･人物判断の5つを用いる.

4. 知的商品ＤＢ検索メカニズムの流れ

処理の流れは図1のようになる.

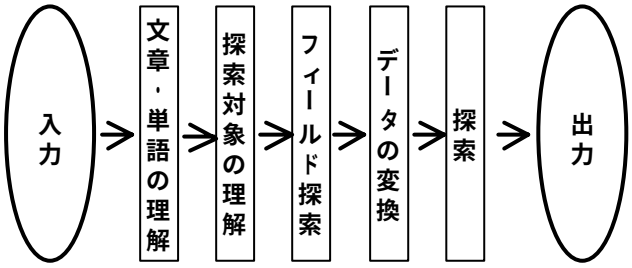

図 1　知的検索メカニズムの流れ

4.1 文章･単語･探索対象の理解

ＤＢ検索をするためには, まず質問文(入力された文章)のうち, どの単語が何を意味するのかということをそれぞれ理解しなければいけない. 例えば”夏”といえば, 『時間語』である. 文章の単語を一つずつとり, それぞれ代表区分(品詞が何であるか, 語が何を表すのかによって区分したもの)に振り分けていく. ここでの代表区分とは『副詞･動詞･商品･時間語･場所語･主体語･統計情報(平均や合計といった語)』の7種類である.

それぞれの単語が何を表すかを理解すれば, 次にその語はどういうフィールドを探せばよいかを与える必要がある. 代表区分ごとに, それぞれの探索対象というものを定義する.

例えば, その単語が『主体語』であれば, フィールドは”人”といったそのまま主体語を表すものに加えて, ”性別”や”年齢”といった人に関する語を探す.

4.2 フィールド探索

探索対象が決定すれば, 実際にそのフィールドを取り出すことが必要となる. ＤＢが入力されれば, そのＤＢ全てのフィールド名を取り出す. 決定した探索対象とそれぞれのフィールド名との最高関連度にあたるフィールドが, 求めるフィールドである. 図2にイメージを示す.

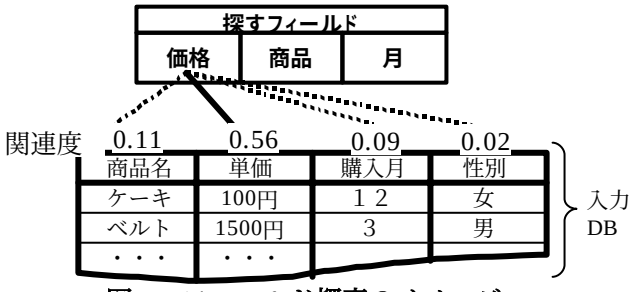

図 2　フィールド探索のイメージ

5. データの変換と取り出し

質問した対象が，ＤＢ内にそのままの表記で存在するわけではないため，探索するフィールドが分かれば，その単語を求めるべきデータへと変換してからフィールド内から取り出さなければならない．それにはさまざまな連想システムや常識的判断を用いる．

5.1 商品区分の判断

対象は商品ＤＢからの検索であるため，商品というものが第一の理解対象となる．一般的に商品一つ一つには商品区分というものがあると考えられる．商品区分というのは，その商品が属する種類のことである．これらの判断にはシソーラスと概念ベースを用いる．入力語から決定したシソーラスのノードより下位に入っているかどうかによって，商品区分を選び出す．

評価は商品を表す語を入力語としたデータ 100 個に対し，与えられた商品ＤＢからどれだけ正しいものが取り出せるかを人目で判定した．その結果を表2に示す．

表 2　商品区分の判断結果

手法	再現率	精度
関連度のみ	37.3%	29.0%
シソーラスのみ	59.2%	51.1%
関連度＋シソーラス	64.4%	54.2%

5.2 さまざまな常識的判断を用いた取り出し

データの取出しにはさまざまな常識的判断メカニズムも利用する．これらを組み合わせることで，あいまいな言葉に対してもその言葉から意味のある単語や数字に変換して，ＤＢ内から正しいデータを検索し，そのレコードを取り出す．これらは商品の判断にも，直接レコードの検索にも用いられる．あいまいな質問文から，検索できるデータへと変換するのに常識的判断を利用する．

5.3 知識による自然言語情報の取得

扱う自然言語の情報は，統計情報，状態語，副詞の3つである．統計情報とは数字自体を表す語，あるいは数字を限定する語のことであり，状態語は数の変化などを表す語のことである．これらは知識ベースを用いて，その単語ごとに定義を与える．

統計情報を DB 化，またプログラム化することによって，これらの言葉からそれぞれの処理を行うことを可能にする．例えば，『およそ』という言葉には『最大値と最小値の差の３％を目安として，基準値±目安がおよその範囲である』という意味のことを DB に知識として格納しておく．このようにそれぞれの言葉にその意味を，知識として格納する．

数字に関する自然文には『伸びる』や『一定』，『多い』といったような状態語が存在する．これらにおいても同様にその状態を満たしている定義･条件を知識として格納することで，対象ＤＢの中から条件を満たすもののみを抽出することが出来る．

述語には副詞がかかる場合が多いが，本稿ではその副詞のうち程度に関するものを定義した．副詞はそれ単独では定義することが難しいので，述語や状態語と組み合わせたもの（『よく』＋『伸びる』など）としてそれぞれを定義した．

これら自然言語情報は，知識ベースを用いてそれぞれを定義し，もし入力語が未知語（知識ベースに存在しない語）の場合は知識ベース内にある単語と入力語で関連度計算を行う．

6. 評価と考察

2つのＤＢに対する質問文 100 文に対する出力結果の正誤判定を人目で評価した．その結果を図3に示す．

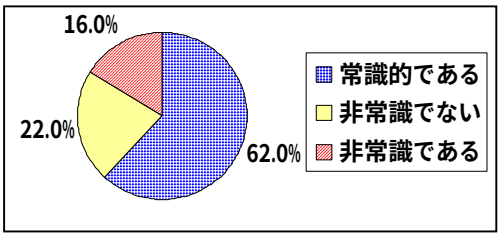

図 3 知的検索メカニズム全体評価

評価の非常識の多くは，自然言語情報の定義の時点での人の感覚からの遠さである．”およそ”や”伸びる”といった言葉に対しての定義はもっとファジーにしなくてはいけないことが理解できた．今回，表の理解ということを想定して知的ＤＢ検索メカニズムを構築したが，そのためにはもっと複雑な表現を理解させる必要がある．そしてそれを実現するには知識ベースをより細かく，より正確に定義することが必要である．

7. おわりに

本研究では，従来のデータベース検索では出来ない，自然言語からのデータベース検索について取り扱った．その材料として，さまざまな常識的判断，そしてそれを補うような定義や知識を与えることでその知的検索を可能にすることが出来た．

目標は全てのＤＢ，全ての表に対して自然言語による検索を行えるようにすることである．また，それらからコンピュータが知識や情報を得るということである．それの第一歩としての研究は，『語→探索対象→検索フィールド→検索』という形を提案することで成功したといえる．

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

[宮本　1999] 宮本敬介：「概念処理を用いたデータベースの知的検索」 同志社大学大学院工学研究科知識工学専攻修士論文，1999.

[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.