4K－1

キー概念の抽出と未知語の処理に基づく情報検索方式の高度化

藤崎 博也[1] 亀田 弘之[2] 大野 澄雄[1] 阿部 賢司[1] 伊東 卓哉[1] 佐久間 聖仁[1]

[1] 東京理科大学 [2] 東京工科大学

1. はじめに

キーワードによる従来の情報検索では、キーワードの表記のみに着目して処理するため、同表記異義・異表記同義の存在が検索性能の低下をもたらす。これを避けるには、キーワードの概念(キー概念)を用いることが有効であるが[1]、キーワードがシステムの辞書に登録されていない語、すなわち未知語[2]の場合には、その処理が必要となる。また、情報検索のより一層の高度化を実現するためには、検索効率を向上させるための様々な知識を、システムが自動的に獲得する必要がある。本報告では、キー概念の抽出・未知語処理・知識獲得を組み合わせた情報検索の新しい方式を提案する。

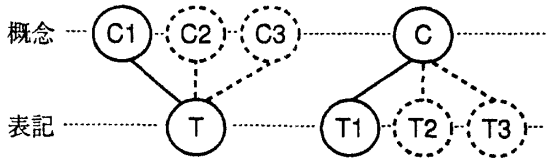
2. 情報検索の高度化

一般に語は表記(表層表現)と概念(深層表現)から構成される。語に同表記異義・異表記同義がある場合の表記と概念の関係を図1に示す。

従来のキーワード検索では、語の表記のみに着目するため、キーワードに同表記異義が存在する場合には、ユーザが呈示した表記Tにより検索し概念C1に関する情報を取り出そうとすると、概念C2, C3に関する不要な情報まで取り出してしまう(図1(a))。また、キーワードに異表記同義が存在する場合には、ユーザが呈示した表記T1では、概念Cに関わる情報のうち、表記T2, T3の形式に言語化されたものは取り出すことができない(図1(b))。すなわち、同表記異義の存在は不要な検索をもたらし、異表記同義の存在は検索洩れをもたらす。これらを避けるには、キー概念のレベルにまで遡った検索が必要である。

しかし、キーワードが未知語の場合には、キー概念を抽出することができない。したがって、未知語の処理も合わせて行う必要がある。

また、的確で効率の良い検索を行うためには、ユーザやデータベースの特徴に関する知識が必要となるが、それらをシステムに予め与えておくことは不可能であり、システムに自動獲得させる必要がある。



(a) 同表記異義　(b) 異表記同義

図1. 語に同表記異義・異表記同義がある場合の表記と概念の関係

3. 新しい情報検索システムの提案

3.1 システムの概要

このシステムは、図2に示すように、ユーザとのインターフェースを担当する1個のAgent1と、データベースとのインターフェースを担当する一般に複数のAgent2を持つ。Agent2には各々専門分野があり、インターネット上に分散する多種多様のデータベースに適切に対応する。以下では、これらのAgentの機能を具体的に説明する。

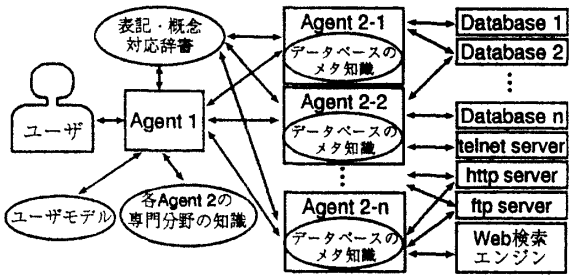


図2. 提案する情報検索システムの概要

3.2 Agent1の機能

(a) 意図推定とそれに即した検索式の生成

ユーザが必要とする情報を的確に検索するために、ユーザが呈示したキーワードとユーザとの対話に基

An Advanced Information Retrieval System Based on Extraction of Key Concepts and Processing of Unknown Words

Hiroya Fujisaki[1], Hiroyuki Kameda[2], Sumio Ohno[1], Kenji Abe[1], Takuya Ito[1], and Yoshihito Sakuma[1]

[1] Science University of Tokyo 2641 Yamazaki, Noda, Chiba, 278, [2] Tokyo Engineering University 1404-1 Katakura, Hachioji, Tokyo, 192

づいてユーザの意図を推定し、これに即してキー概念からなる検索式を生成する。さらに、この対話およびユーザの検索歴に基づいて、ユーザが要求するキー概念の重みづけを行う。なお、この際の対話は、ユーザの思考の流れを妨げないように、主として音声により行う。

(b) 表記・概念対応辞書の管理・拡張と知識獲得

対話は自然言語を介して行うため、意図を推定するには、表層表現 (表記) と深層表現 (概念) との対応づけが必要となる。Agent1 は、表記と概念の双方から参照できる表記・概念対応辞書を用いて、この対応づけを行い、さらに、辞書の管理と拡張を行う。最初は、システム設計者から与えられた小規模な辞書から始まり、以後、辞書拡張の知識を自動的に獲得する。

(c) 未知語の処理

ユーザが呈示したキーワードが未知語の場合には、対話によりその概念を明確化し、システムにおける既知概念との対応づけを行う。もし明確化された概念に対応するものがシステムになければ、新概念としてシステムの辞書に登録する。

(d) ユーザのシステム利用法に関する知識の獲得

検索精度と検索速度を向上させるために、ユーザのシステム利用法に関する特徴をデータ (ユーザモデル) として蓄積し、ユーザに適した検索ルートを自動的に獲得する。

(e) 検索依頼

抽出したキー概念の分野を判断し、適切な Agent2 (一般に複数) に検索を依頼する。

(f) 検索結果の検討

Agent2 から受けた検索結果とユーザの意図との整合性を上記 (a) で求めた概念の重みづけに基づいてチェックし、検索結果の順位づけを行う。さらに、その結果をユーザに呈示し、検索結果の妥当性をユーザとの対話によって確認する。結果が妥当であれば処理を終了する。ユーザの意図と合致しない場合には、キー概念の重みづけを変更し、既に求めた検索結果の順位を変えて再度ユーザに呈示する。また、対話により新たなキー概念が呈示された場合には、検索式を修正し、再び検索を Agent2 に依頼する。以後、満足な結果が得られるまでこの処理を繰り返す。

3.3 Agent2 の機能

(a) データベースからの情報検索

Agent1 から検索要求を受け、データベースの内容とその記述形態に関する知識 (データベースのメタ知識) を利用しながらキー概念検索を行う。

(b) 未知語の処理

データベース上のキーワードにも未知語が存在するが、対話処理による概念推定は行えない。したがって、キーワードの表層上の特徴から概念推定を行う。例えば、キーワードが複合語である場合には、各構成要素に着目し、その要素がシステムの表記・概念対応辞書に登録されていれば、要素の概念から全体の概念を推定する。[3]

(c) データベースに関するメタ知識の獲得

各データベースを定期的にチェックし、データベースの内容に変更がある場合には、それに応じてデータベースのメタ知識も変更する。また、検索時のヒット件数を記録することにより、データベースの有用性に関する知識も、メタ知識として自動的に獲得する。

4. おわりに

本稿では、キー概念の抽出・未知語処理・知識獲得を組み合わせた新しい情報検索システムを提案した。このシステムでは、検索にキー概念を用いることにより、ユーザの意図により即した情報検索を行い、また、未知語処理による検索精度の向上、知識獲得による検索効率の向上を図っている。さらに、ユーザの意図推定と情報検索とを分離して取り扱うことにより、検索精度・検索効率の一層の向上が期待される。

参考文献

[1] 藤崎博也, 亀田弘之, 河井恒 : " 新聞記事情報の階層構造に基づく記事分類・検索システム, " 情報処理学会「自然言語処理」研究会資料 44 - 4 (1984).

[2] 亀田弘之・藤崎博也・森田敏生・倉島顕尚 : " 未知語の分類とその処理に関する考察, " 情報処理学会第 36 回全国大会講演論文集, 5T-5, pp. 1195-1196 (1988).

[3] 亀田 弘之: " 日本語文章理解における未知語とその処理, " 知識科学の最前線シンポジウム論文集別添資料, pp.1-11(1993).