インターネット多角的検索システムOTROS

3L－3

－データベースを用いたクエリ展開方式の提案－

竹元義美　山田洋志　福島俊一
NECヒューマンメディア研究所

1. はじめに

インターネットが爆発的に普及し、膨大な量の電子化データを対象とした情報検索技術の高精度化に対する要求が高まっている。

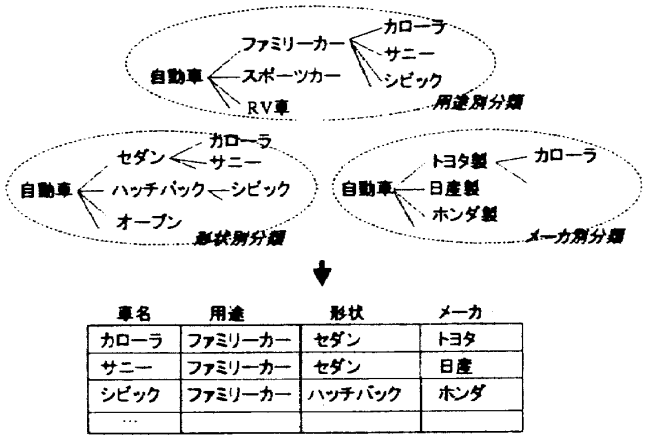
テキスト検索において、辞書やシソーラスなど語彙知識情報を用いて、入力された検索キーと同義・類義・階層関係などにある語を検索キーとして追加することで再現率を向上させる技術（クエリ展開）がある[1-4]。例えば、「自動車」に関する文書を検索したいとき、シソーラスを用いて「自動車」という検索キーを下位語の「乗用車」「トラック」などに展開して検索することにより、検索もれを少なくすることができる。インターネット上には、製品名だけが検索の手がかりとなる文書も多いため、クエリ展開は、一般語レベルだけでなく製品名などのインスタンスレベルまで行えることが望ましい。また、インスタンスは一般語に比べて多義語が少ないので、インスタンスへの展開は検索結果のゴミを増やさずに再現率の向上が望める。

本稿では、検索キーをインスタンスに展開するために、RDBのような表形式のデータベースをクエリ展開の知識として用いるクエリ展開方式を提案する。RDBは、シソーラスのような木構造知識に比べ、インスタンスの持つ複数の分類情報や数値情報を一元的に管理しやすいという利点がある。今回は、自動車関連文書を検索する場面を例に、自動車製品情報RDBを用いたクエリ展開システムを試作した。

2. 従来のクエリ展開方式の問題点

検索キーをインスタンスに展開するために、従来のようにシソーラスなどの木構造知識を用いることには、次のような問題点がある。

第1に、複数視点からの展開情報を管理しにくいという問題がある。自動車の例では、用途別（「ファミリーカー」「スポーツカー」等）、形状別（「セダン」「クーペ」等）、メーカ別などの多様な視点での問い合わせ（クエリ）が想定され、各視点について木構造で分類が必要である。これらの分類を示す語は互いに明確な上位・下位関係が必ずしもなく、複数の木構造



車名	用途	形状	メーカ
カローラ	ファミリーカー	セダン	トヨタ
サニー	ファミリーカー	セダン	日産
シビック	ファミリーカー	ハッチバック	ホンダ
…			

図1：インスタンスの分類視点と表現形式

知識を構築しなければならず、管理コストがかかる。

第2に、木構造知識で数値情報が扱いにくいという問題がある。自動車に関するクエリでは、価格や排気量などの数値条件も想定され、これらの数値条件を用いたクエリ展開が必要となる。

また、構造上の問題以外に、インスタンスレベルの木構造知識がほとんどないという問題がある。シソーラスは一般語を中心に構成されており、固有名詞情報まで含めることは少ない。

3. RDBを用いたクエリ展開方式

検索キーをインスタンスに展開するために、RDBをクエリ展開の知識として用いるクエリ展開方式を提案する。図1にインスタンスの複数分類視点を木構造で表現した場合と表形式で表現した場合との比較を示す。RDBは、複数の分類視点を単一の表で管理できるため、シソーラスのような木構造知識に比べ、複数視点からの展開情報を一元的に管理しやすい。また、RDBは、製品情報などのDBに広く利用されているため、インスタンス展開に流用しやすい。なお、2節で述べた問題点はインスタンス展開で発生するものなので、一般語展開は従来通りシソーラス・同義語辞書を用いる考えである。

自動車関連文書を検索する場面を例に、自動車製品情報RDBを用いたクエリ展開システムを試作した。試作システムのUIについて、ユーザが「セダン」の車について検索する例を説明する。ユーザはセダンの車名を列挙したいがそれをよく知らないので、RDBを検索して車名群を得ることが目的である。ここで、RDB検索

Internet Diversity Retrieval System OTROS
-Query Expansion Using Relational Databases-
Yoshikazu Takemoto, Hiroshi Yamada, and Toshikazu Fukushima
Human Media Research Labs, NEC Corp.

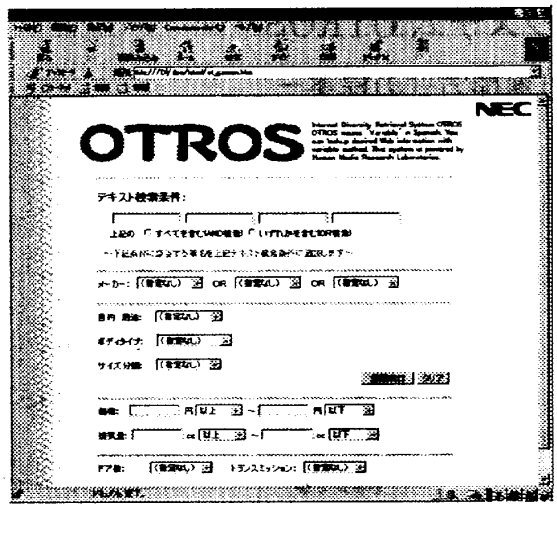


図２：クエリ展開用の条件入力画面

が主目的ではなく、RDB 検索によりテキスト検索のためのクエリ展開を行っていることに注意したい。

検索システムの初期画面で、検索キーの展開に用いる RDB を選択するためのメニューを用意し、ここで「自動車 RDB」を選択する。RDB を選択後、クエリ展開用の条件を入力する画面（図 2）が提示される。この画面で、選択した RDB の属性項目を一覧表示し、各属性項目に対する値をメニューから選択できる。例えば図 2 で「ボディタイプ（形状）」の項目から「セダン」を選択する。

図 2 で展開実行ボタンを押下すると、自動車 RDB から「形状＝“セダン”」を満たすレコードを選択し、そのレコードからインスタンスとして車名“カローラ”“サニー”等を取得し、インスタンスを検索用の演算子で結合した検索式（インスタンス検索式）を作成する。提案方式の目的が検索キーの拡張にあるので、インスタンス間の結合は、「“カローラ” OR“サニー” OR...」のように OR 結合とする。こうして作成したインスタンス検索式を検索システムに入力する。試作システムの構成を図３に示す。

4. 提案方式の効果および今後の課題

次の 2 つの観点から提案方式の効果と今後の課題を述べる。

● クエリ展開における RDB の利用

インスタンスは下位レベルの語なので複数の視点から分類され得るため、木構造知識を用いるよりも、提案方式はインスタンスへの展開情報の管理に有効であると考える。試作システムでは、RDB を限定してクエリ展開に利用したが、今後、既存の RDB を流用する汎用的な仕組みを検討することで様々な分野のインスタンス展開が可能になると考えている。

また、提案方式の実用化のためには、UI 設計が重要で、ユーザが RDB 検索をできるだけ意識することなくクエリ展開を利用できることが望まれる。現状は、ユーザが DB 検索式を SQL 文で記述するのは難しいことを考慮し、図２に示す UI を用意した。より望ましい UI としては、ユーザが入力した初期検索式を解析して適切な RDB を選択し、適当な DB 検索式を作成して RDB を検索し、自動的にインスタンスを取得することが考えられる。

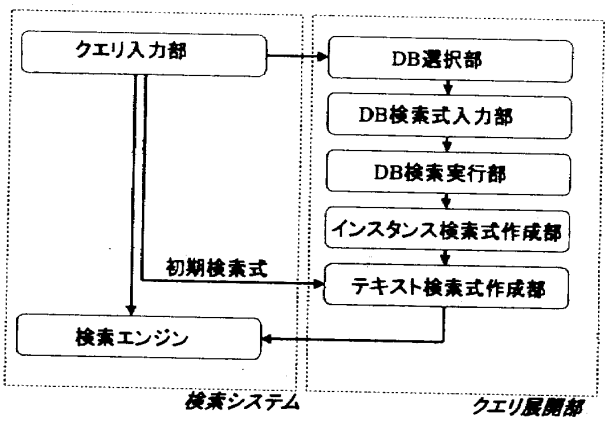


図３：システム構成

● クエリ展開におけるインスタンスの利用

提案方式により再現率は向上する。例えば「セダン」に関する文書を検索したいとき、“セダン”という語を含まなくても、そのインスタンスを含む文書を検索できる。一方で、クエリによってはインスタンスの数が多くなり、とくにインスタンスに曖昧性（多義・部分一致）があるときに検索結果のゴミが増えてしまう。曖昧性があるインスタンスは、検索式を作成する際に、RDB の他の属性値を AND 結合した式や近接演算式を作成することで適合率を高めることができる。

5. おわりに

RDB を用いて検索キーをインスタンスに展開するクエリ展開方式を提案し、インターネット多角的検索システム OTROS[5]に実装した。提案方式はインスタンスへの展開情報管理に有効である。インスタンス展開により、検索結果のゴミを抑えつつ再現率を向上させることができる。

参考文献

[1] Ellen M. Voorhees, “Query Expansion using Lexical-Semantic Relations”, ACM SIGIR '94

[2] 下畑他、“多様分類情報による検索語拡張”、NL 研、115-19、1996

[3] 斉藤他、“概念に基づく検索要求文の拡張”、NL 研 121-18、1997

[4] 太田他、“EDR 電子化辞書を用いたクエリー拡張による検索支援”、言語処理学会第３回年次大会、1997

[5] 山田他、“インターネット多角的検索システム OTROS－全体の概要と構成－”、情処 57 全大、3L-01