

WWW 情報と推論システムを用いたテレビ番組情報検索システム

1 A C - 1

Somnuk Sanguantrakul 塚本昌彦 西尾章治郎

大阪大学大学院 工学研究科 情報システム工学専攻

1 はじめに

近年、計算機の高性能化および記憶媒体の低価格化によって、電子的にデータを蓄えることが多くなった。また、急激に拡大しつつあるインターネットによって、電子的に提供されているデータ量が増加している。この膨大なデータの中から必要とする情報を探し出すことは時間と労力のかかる作業である。このため、柔軟性の高い検索システムの重要性が高まってきていると考えられる。しかし、従来の WWW サーチエンジンのような検索システムでは、問合せとして文字列を受け取り、その文字列との単純な比較により、問合せの答を求めている。このような手法にはいくつかの問題点がある。

まずひとつ目の問題点として、問合せに抽象的な概念を表す語を用いることができない点が挙げられる。このため、ユーザはその概念に含まれる具体的な要素を羅列しなければならないので、問合せ文が必要以上に長くなり、入力し損うなどの間違いが生じやすくなる。また、ユーザは検索システムの扱っているジャンルに関する背景知識を常に正確に把握している必要があるため、検索システムの利用者が限定されてしまうことになる。

ふたつ目の問題点として、検索データのさまざまな性質から柔軟に問合せを行うことができないという点があげられる。例えば、ユーザは「主題歌の歌手が関西のグループであるような番組」の情報を調べようとしても、文字列の比較で検索するようなシステムではこのような条件を直接指定できない。

さらに、検索システムの扱うジャンルによっては、数字の比較などの処理を伴う問合せ記述の実現が要求される場合があるが、文字列の比較だけではこのような問合せに対処できないという問題点もある。

そこで本稿では、知識ベースを用いることで柔軟に問合せを処理でき、扱うジャンルに応じて幅広く問合せできるような検索システムを提案し、それをテレビ放送番組検索に適用したシステムの設計と実装について述べる。

以下では、まず、2章で検索システムの構築方針について述べる。次に、3章でテレビ番組情報検索システムの構成について、4章でその実装について述べる。最後に、5章で結論と今後の課題を示す。

2 検索システムの構築方針

本稿では対象としてテレビ番組情報を考える。近年のチャンネル数の急増によりテレビ番組の数が増えている。そのため、全ての番組の情報を把握することが困難になり、検索システムが必要になってくる。また、番組情報は放送時間、出演者、対象など多様な属性をもっているため、柔軟に問合せを処理できる検索システムが必要である。

本稿では、検索システムを構築する際に知識ベースを導入することを考える。管理者があらかじめ検索システムの扱っているジャンルに関する知識を知識ベースに蓄えておき、検索システムはその知識と推論システムを用いることで、ユーザから受け取った問合せを柔軟に処理することができる。

実際に構築したテレビ番組情報検索システムでは、知識ベースとして DOT を採用している。DOT とは、ドット記法と

IS-A 関係を用いて知識を表現し、IS-A 関係の順序性と継承関係をもとに推論を行う知識ベースシステムである [2]。DOT を採用することにより、さまざまな観点からの情報を統一的なモデルで扱うことができ、知識を簡潔に記述することが可能になる。

しかし、一般に知識ベースシステムは、知識量が増えるにつれて推論機構における推論時間が長くなるという特性があるため、背景知識が複雑な分野において検索システムに適用することは難しい。この問題に対して、データベースシステムの導入および知識の分割と階層化で解決を図った。

データベースシステムの導入：一般に検索データの中には、推論に寄与するデータとそうでないデータがある。これらのデータを全て知識ベースに蓄えると、知識量が多くなり、推論機構の性能の低下につながる。これを避けるためには、検索システムにデータベースシステムを導入し、推論に寄与しないデータをデータベース内に蓄えておくことで知識量を減らすことができる。

知識の分割と階層化：通常、ユーザからの問合せを処理するために要する知識は知識ベース内のごく一部だけである。よって、問合せを処理するために知識ベース内の知識を全部推論に用いるのは効率的ではない。また、DOT 推論機構における推論時間は、知識ベース内の知識量に依存する。すなわち、関連知識量が多くなるにつれて、DOT 推論機構の推論時間が長くなる [1]。この問題を解決するための手法として本稿では、推論に用いる知識をいくつかに分割し、問合せを評価するときは、メタ知識を用いることで検索の際に必要な最小限の知識を用いて推論するという手法を用いた。

また、検索データを蓄える機構に依存することなく、応用分野に応じた問合せに柔軟に処理できるようにするために、本システムでは、検索システムに関数ライブラリを導入している。これにより、文字列の比較やデータベースシステムの問合せ言語などでは対処できないような問合せに対して、関数ライブラリ内に蓄えている関数を用いて対処できる。

本稿で構築するテレビ番組情報検索システムでは、番組の開始時間および終了時間をデータベースに蓄えることで知識量を減らしている。また、番組に関する知識を放送日とチャンネルによって分割することで検索時に用いる知識量を減らしている。さらに、関数ライブラリを用いることで、DOT で扱うことが困難である時間の比較などができる。

3 システムの構成

本システムは、特定のページから番組表のデータを定期的 to 取得し、それに基づいてデータベースと一部の知識を自動的に生成する。そして、生成されたデータベースとあらかじめ用意してある知識を用いて、ユーザから与えられた問合せを満たす番組の諸情報を列挙するページを作成してユーザに提示する。図 1 にシステムの構成を示し、システムの各要素について簡単に述べる。以下に、システムの各要素について簡単に説明する。

インタフェース部：ユーザから問合せを受け取って構文解析を行い、いくつかの副問合せに分割して実行部に渡す。また、副問合せの答として実行部から返された番組の集合から問合せの最終的な答を求めてユーザに提示する。

実行部：副問合せをいくつかの条件式に分割し、条件式内の語句に応じて知識処理機構が関数処理機構に条件式を渡す。

TV Program Information Search Engine Using WWW Information and Deductive System

Somnuk SANGUANTRAKUL, Masahiko TSUKAMOTO, and Shojiro NISHIO.

Department of Information Systems Engineering, Graduate School of Engineering, Osaka University.

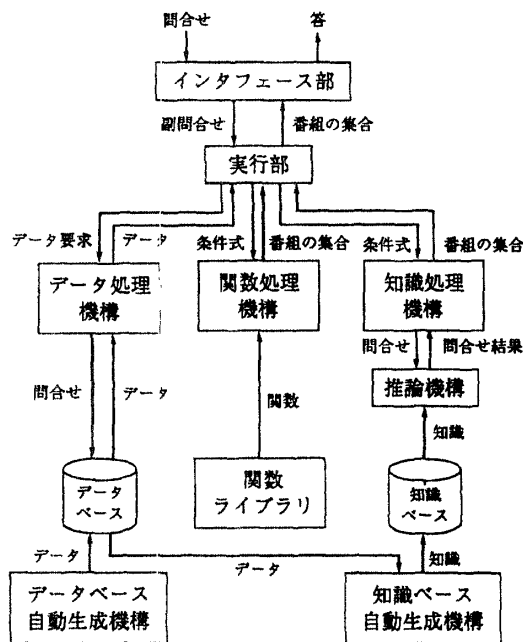


図1: システムの構成

また、条件式の答として返された番組の集合から副問い合わせの答を求めてインタフェース部に返す。さらに、データベース内に蓄えている番組のデータの要求があれば、データ処理機構に対してそれを要求する。

知識処理機構: 実行部からの条件式を推論機構の問い合わせ言語に変換して推論機構に問い合わせる。また、知識ベースシステムから返された答に基づいてその条件式を満たす番組を選択する。

データ処理機構: 実行部が要求した番組のデータをデータベースシステムに問い合わせ、その結果を実行部で用いる形式に変換して実行部に返す。

関数処理機構: 実行部からの条件式に応じて、適切な関数を選択し、条件式のパターンから関数に渡す引数を求めて関数を実行する。

関数ライブラリ: 本システムにおいて、知識ベースシステムの推論機構で処理できないような条件式は、perl 言語で管理者が作成する関数を用いて処理する。関数ライブラリはこのような関数の集まりである。

知識ベース: 番組やチャンネルに関する知識など、推論に用いる知識を格納する。

データベース: 推論に寄与しない番組のデータを格納する。

推論機構: 知識ベース内に蓄えられている知識を用いて推論を行い、知識処理機構からの問い合わせに答える。

データベース自動生成機構: 番組の情報を提供するページのソースファイルを定期的に取得して解析してデータベースを生成する。

知識ベースの自動生成機構: 日付に関する知識や、不規則に現れる番組に関する知識を自動的に生成する。

実行部と推論機構、関数ライブラリおよびデータベースシステムとの独立性を図るため、その間に知識処理機構、関数処理機構およびデータ処理機構を設けている。これにより、検索システムの扱うジャンルに適する推論機構、データベースシステム及び関数ライブラリを用いることができる。

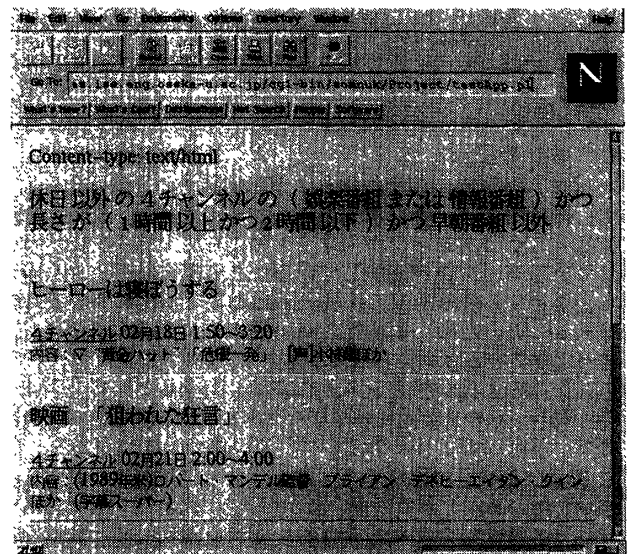


図2: 問い合わせ結果の例

4 システムの実装

本システムは、Sun SPARC station 上で、データベースシステムとしてmSQLを用い、知識ベースシステムとして、筆者らが開発を行っているDOTを用いて構築した。

本システムにおいて、問い合わせを満たす番組の集合は番組名をキーとするハッシュによって実現する。検索を効率に行うために番組の集合を以下に示す二通りの方法で実現している。

実行部より下位の階層で用いる集合: この集合は主に副問い合わせを評価するとき用いる。副問い合わせを評価する際に検索を効率的に実行するため、番組を識別できるための最小限の情報である開始時間だけを集合に付加する。

実行部より上位の階層で用いる集合: この集合は主に問い合わせに対する答のページを作成する際に参照されるもので、放送日や放送チャンネルなど番組の諸情報が含まれている。

本システムの問い合わせ結果の例を図2に示す。

5 おわりに

本稿では、柔軟に問い合わせを処理できる検索システムの実現例として放送番組検索システムについて述べた。検索システムを構築する際、システムの扱っているジャンルに関する知識と推論システムを用いることで問い合わせを柔軟に処理できた。また、データベースシステムの導入及び知識の分割によって知識ベースシステムにおける推論時間を短縮でき、関数ライブラリの導入によって推論システムの能力に依存せず、応用分野に応じて幅広く問い合わせを処理できた。

今後の課題としては、管理者の負担を軽減するための知識ベース構築の自動化と、主観的な知識を扱うためのユーザ依存の知識の作成などが挙げられる。

謝辞

末筆ながら、本研究を進めるにあたり貴重な助言を頂いた、柳沢豊氏を始めとする西尾研究室の諸氏に感謝の意を表す。

参考文献

- [1] 清一隆, 塚本昌彦, 西尾章治郎: 「ドット記法とIS-A 関係を用いた推論システムの高速化に関する研究」, 第8回人工知能学会全国大会論文集, pp.517-520(1994).
- [2] 塚本昌彦, 西尾章治郎: 「ドット記法とIS-A 関係を用いた知識表現システムDOT」, 人工知能学会誌, vol. 10, no. 2, pp.288-297(1995).