2E−7

音声対話と全文検索を利用した電子ガイドシステム(4)
−全文検索部−*

伊藤史朗　酒井桂一　小森康弘　上田隆也　藤田稔
キヤノン(株) 情報システム研究所

1　はじめに

我々は,「旅行」に関するガイダンスを音声対話により行なう電子ガイドシステムTARSANを作成した[1]. 近年, 音声対話システムの研究が盛んであるが, 従来のシステムが対話を進めるために用いている知識は実験規模を越えていない. TARSANでは, 実世界でガイドを行なう時と同じ規模の知識を使うことを目標とした. しかし, 大規模な知識ベースを構築することは容易ではない. そこで, 旅行情報について書かれたガイドブックの文書をそのまま知識源として用いることにした. 実際に使用した文書は,「旅蔵」[2]と「JTB宿泊情報」[3]である. いずれもCD-ROM文書として計算機上での利用が可能である. これら二つの文書で, 名所・施設・宿などについての情報約8万件が記述されている.

文書データからガイドに必要な情報を取り出すために, 全文検索を利用している. 全文検索を行なうことで, CD-ROM文書を加工することなく情報検索を行なうことができる. しかし, 従来の全文検索技術[4]では, 文書自体を検索することが目的であり, 文書中に書かれている個別の情報を取り出すことはできなかった. そこで, 本システムの全文検索部では, 対話処理部からの要求に従って, ユーザが求める情報の書かれている文書を検索し, その文書中からユーザが求める情報を取り出して対話処理部に渡す枠組みを実現した. 以下, 全文検索部の特徴と構成について述べる.

2　全文検索部の特徴

TARSANの全文検索部は, 以下のような特徴を持っている.

2.1　項目単位での全文検索

ガイドブックの文書は, 形式的に区分可能な領域を持つことが多い. 図1は, TARSANで用いている文書データの例である. ここでは, 温泉の名称, 所在地, 宿泊施設の数などについての情報が記述された領域が, それぞれ行数や「所在地：」などの文字列を使って区分可能である. このように, 形式的に区分可能な領域を項目と呼ぶ. この例では, 各行がそれぞれ項目になる.

TARSANの全文検索部では, 検索の単位を項目にする. すなわち, 特定の項目に検索語が記述されているかどうかを全文検索条件とする. また, 検索された文書から特定の項目のデータを取り出すことができる.

```
湯ノ花沢温泉
所在地：神奈川県箱根町
名称ヨミ：ユノハナザワオンセン
交　通：小田原駅バス５０分
宿泊施設(軒数)：１
宿泊施設(人数)：２４０
概　要：芦ノ湯の上にあり展望絶好。硫化水素泉５５～８５度。標高９４０m。
効　能：皮膚病
利用者数(年間)：６１年(１月～１２月)２１，５０９人
```

図1: 文書データの例

2.2　属性の型に応じた検索

どのような内容の項目が文書中にあるかを示すために, 項目の内容を表す語を属性として定義する. 対話処理部との情報の受け渡しは属性を用いて行なう.

属性には型を持たせ, 型に応じて検索方法を変える. これにより, 従来の全文検索では扱うことのできない数値条件などの条件を扱うことができる. 表1にTARSANで用いられている属性と型の例を示す. 整数型では, 大小比較などを用いた数値条件を扱う. 地名型では, 地名シソーラスを用いた検索を行なう. 例えば, 条件を「湘南」とした場合に, 該当する市町村のデータを検索できる. 語や文の型では, 通常の全文検索を行なう.

表1: 属性と型の例

属性	型
名称	語
所在地	地名
施設数	整数
説明	文

項目の属性と型を定義して, これらを用いて検索を行なう方法は, いわば文書をデータベースとして用いているといえる. 型により検索手法を切り替えることにより, 全文検索に基づくデータベース検索を行なっていると考えることができる.

*An Electronic Guidance System with Speech Conversation and Full Text Retrieval (4) - Full Text Retrieval - Fumiaki ITOH, Keiichi SAKAI, Yasuhiro KOMORI, Takaya UEDA, Minoru FUJITA (Information Systems Research Center, Canon Inc.)

2.3 連続領域に対するインデクス検索

検索時に全データを全文検索していると時間がかかりすぎる．そこで，TARSANでは，ジャンルと所在地を必須条件にして，全文検索にかけるデータの絞り込みを行なっている．そのために，ジャンル名と都道府県単位の所在地名をキーとしたインデクスを用意している．

一般に，文書は，ある分類に従って順番に並んでいる．TARSANで用いる文書では，ジャンルと所在地による分類に従って並んでいる．そこで，同一分類に属する文書が連続している領域を単位としてインデクスを作成する．このようにすると，1文書ずつインデクスを作成する場合に比べて，インデクス作成と検索のコストが低減できる．

このインデクス検索の目的は，全文検索にかけるデータ量を許容範囲内に抑えることであって，完全な検索を行なうことではない．言い換えれば，許容範囲内に抑えることのできるインデクスだけを用意すればよい．例えば，市町村レベルの所在地が検索条件として指定された場合，インデクス検索では都道府県レベルまでしか絞り込まない．市町村レベルでの絞り込みは全文検索により行なう．

なお，1ジャンル1都道府県を条件に指定した場合の検索時間は，平均3秒程度である．

2.4 検索条件の判定と変更要求

対話処理部から送られてくる検索条件が不適当な場合に，検索条件の変更要求を行なっている．すなわち，検索条件が不適当な場合のシステムの応答を全文検索部が決定している．このように，全文検索部の仕様に依存する対話の処理を対話処理部から切り離したことで，それぞれの独立性が高まった．

3 全文検索部の構成

TARSANの全文検索部は，CD-ROMドライブを接続したNeXTstation Turbo Color上に，Objective C言語によるソフトウェアだけで実現されている．その構成を図2に示す．

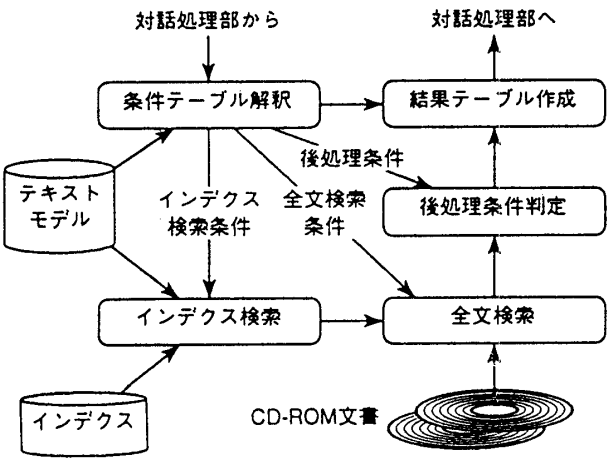

図 2: 全文検索部の構成

1:テキストモデル

各文書の属性，型，項目区分のための情報などを記述している．

2:条件テーブル解釈

対話管理部との検索条件・結果の受け渡しはテーブルを用いて行なっている．ここでは，条件テーブルの解釈を行なう．その結果，存在しない属性が用いられていたり，必須条件が指定されていないなど条件が不適切なときは，条件の変更を要求する．適切な条件の場合は，テキストモデルを参照して，属性ごとにインデクス検索条件・全文検索条件・後処理条件の内部条件を作成する．

3:インデクス検索

インデクスが付いている属性に対してインデクス検索を行ない，全文検索にかける文書のアドレスを得る．

4:全文検索

検索語の記述されている文書を検索する．項目の区切りを検索語のパターンマッチングと同時に行なうので，項目単位の検索も高速に行なえる．

5:後処理条件判定

整数型の条件判定など，全文検索では判定できない条件の判定を行なう．判定に用いる値は全文検索で用いたバッファに入っているので，判定は高速に行なわれる．

6:結果テーブル作成

検索条件の変更要求や検索結果を結果テーブルに格納し，対話処理部に送る．

4 おわりに

音声対話システムの知識源として，全文検索により文書データを用いる手法について述べた．ここで用いられている要素技術は以下の通りである．

- 項目単位での全文検索
- 属性の型に応じた検索
- 連続領域に対するインデクス検索

この手法により，既存の旅行ガイドの文書をそのまま利用した大規模データに基づく電子ガイドシステムを作成することができた．このように文書データを直接データベースとして利用する方法は，データ作成の容易さや文書データの汎用性から音声対話システムに限らず有効な方法であると考える．

参考文献

[1] 藤田他: 音声対話と全文検索を利用した電子ガイドシステム (1) −システム概要−, 本大会予稿.

[2]「旅蔵」電子ブック, 企画 製作 (株) 廣済堂, 発行 (社) 日本観光協会, 1990.

[3] JTBの「宿泊情報」(電子ブック), 編集人 中嶋隆一, 発行人 岩田光正, 発行所 JTB日本交通公社出版事業局, データ編集凸版印刷株式会社, 1992.

[4] 菊地他: 全文検索の技術動向とシステム事例, 情報学基礎 25-1, 1992.