コモンHIサービス環境を利用した文書の構造化

7Q－1

芝崎 靖代　長谷川 保　真鍋 俊彦　中山 康子
(株)東芝　研究開発センター

1　はじめに

筆者等は、組織の知識情報共有を促進するシステム (Advice/Help on Demand) を開発し、実際にストック情報のコンテンツを入れて組織内で実践評価を行なっている [1]。本システムは、オフィスの人・組織情報、文書や個人のノウハウ等を知識データベースにストックし、自然言語対話により検索するものである。

このようなシステムが有効に利用されるためには、コンテンツの充実と、知識情報を検索・活用しやすいように構造化して獲得することが重要である。そこで、音声認識、文字認識、文書理解等の各種メディア変換処理を種々の応用から利用できるＨＩウェア（コモンＨＩサービス環境)[2] を活用し、紙の文書を含んだマルチメディア情報を構造化してデータベースに登録するインタフェースを開発している。本稿では、文書の構造データの獲得方法について述べる。

2　文書の構造データの獲得

文書登録インタフェースのシステム構成を、図１に示す。

ユーザの自由な検索文から対象文書を検索できるようにするためには、文書の内容を表す構造情報を抽出しなければならない。そこで、オフィスの知識を記述したオフィス知識ベースを参照して、文書の内容理解を行なう。

オフィス知識ベースは、オフィスの人・組織、業務、文書、相互関係や機能、業務手順等を記述したものである。例えば、会議資料を作成するプロジェクトの業務構造として、プロジェクト名、リーダー、参加組織、ミッション、会議スケジュール、キーワードを記述している。これを参照して、構造情報を持たない会議資料などの非定型文書から、タイトル、作成者、作成日、会議日、会議名、参加者、プロジェクト名、キーワード、関連文書の構造情報を抽出する。

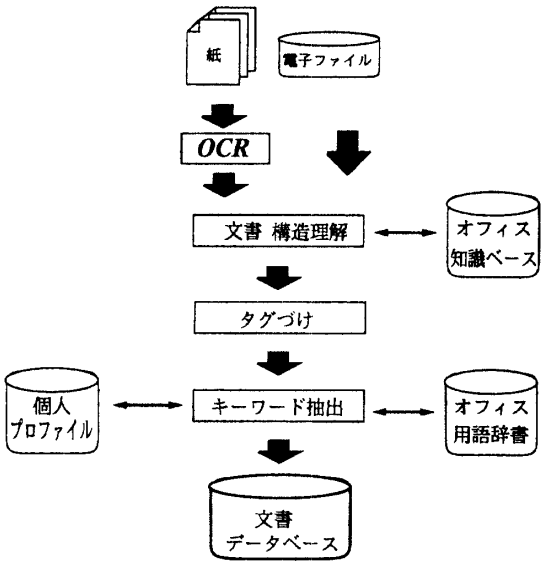

図 1: 文書登録インタフェース

このような知識を参照して、さらに文書を共有するメンバーの限定、会議内容に関連する他の文書の参照、プロジェクト管理等が可能になる。

3　キーワード自動抽出

オフィス用語と個人プロファイルデータを用いて、文書の内容からキーワードを抽出する手法を、図２に示す。

3.1　オフィス用語辞書

文書の内容理解とキーワード抽出率を向上させるため、オフィス業務で使われているあらゆる用語を網羅した、オフィス用語辞書を開発した。この辞書は、３００人規模のオフィスを対象に、事務手続き関連用語、技術専門用語など約１０，７８０語を収集したもので、各分野の専門家が同義語 (同じ意

Keyword Extraction method for Documents
Yasuyo Shibazaki,
Tamotsu Hasegawa and Toshihiko Manabe
and Yasuko Nakayama
Toshiba Corporation

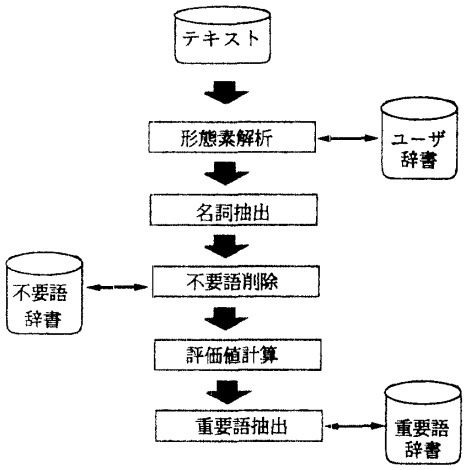

図 2: キーワード自動抽出

味をもつ様々な表現や言い回し) を登録した。これにより、キーワード抽出率を向上させ、またユーザが自由な検索文を記述できるようにした。

3.2 プロファイルデータ

オフィス用語辞書から、作成者の専門分野、担当分野、個人的表現などを収集した、個人プロファイルデータを開発した。個人プロファイルデータと文書の種類を組み合わせることで、進捗報告や会議資料などの文書から、より専門性の高いキーワードを抽出することができる。

3.3 キーワード抽出方法

文書のテキスト文に対して形態素解析を行ない、キーワードの候補となる名詞の切り出しを行なう。この際、形態素解析のユーザ辞書に、キーワードの候補となる重要語としてオフィス用語辞書を登録する。これにより、「知識」「情報」「共有」という3つの単語を1つの複合語「知識情報共有」として切り出すことができる。

次に、一般的な名詞など、キーワードとしてふさわしくない単語を登録した不要語辞書を用い、検出された名詞から不要語を取り除く。そして、重要な単語は出現頻度も高いという仮説に基づき、名詞の出現頻度をカウントし、それぞれの評価値とする。最後に、キーワードとしてふさわしい用語を登録した重要語辞書を参照し、評価値を重みづけし、総合評価値の高い用語からキーワードとして抽出する。

不要語辞書と重要語辞書について説明する。オフィス用語辞書は、【個人プロファイルデータ】【一般研究用語】【社外組織名】【社内組織名】【製品名】に予め分類してある。一方、進捗報告、出張レポートなどの文書の種類と作成者の組合せで、用語分類の優先ルールを設定する。例えば、文書が個人の進捗報告であれば、【個人プロファイルデータ】を重要語辞書に、【社内組織名】を不要語辞書にそれぞれ登録すれば、作成者の専門研究分野に限定されたキーワードが抽出できる。

4 文書データベースと検索

以上の手法で、進捗報告、会議資料、出張レポート、物品発注などの文書 1,500 件を構造化してデータベース化し、検索できるようにした。

例えば、「先期、鈴木さんが書いた知識情報共有に関するレジュメ」という自由な表現の検索文を、オフィス知識ベースを用いて、

- 作成日：９６年４月～９月
- 作成者：鈴木
- 文書の種類：個人進捗報告、会議資料
- キーワード：
  「知識情報共有」「Advice on Demand」

のような検索条件に展開し、文書データベースの構造情報とマッチングを行ない、該当する文書を提示する。

5 まとめ

メディア変換処理を備えたコモンＨＩサービス環境を応用し、組織や個人で所有する様々な形式の文書から、検索に有効な構造化データを自動的に獲得する機能を開発した。オフィス用語辞書、個人プロファイルデータ、オフィス知識ベースを用いることで、文書の内容に即した構造情報とキーワードを獲得することができた。

今後は、映像、イメージなども含めたマルチメディア情報の構造化を検討していく。

参考文献

[1] 中山他. 知識情報共有システムの開発と実践－オフィス知識ベースとノウハウベースの構築－ インタラクション'97, 1997.

[2] 杉山他. コモン HI サービス環境の開発 第 52 回情報処理学会全国大会, 6-193, 1996.