情報検索システムへのキーワード自動抽出機能の適用方法について

2N-4

石黒正典　穐吉秀嗣　田中一敏
ＮＴＴ情報通信処理研究所

１．はじめに

文献情報，新聞記事情報等を提供する情報検索サービスにおいては，原情報である“原文”からキーワードを抽出し，検索用のインデックスを作成する際に，

1.原文を読みキーワードを抽出する，
2.原文そのものをデータベースに登録する，
3.登録した原文に対応するキーワード群を入力してインデックス作成をシステムに指示する，

という過程が全て人手を介して行なわれており，多大な工数を要していた．我々は，人手に頼っていた作業を可能な限り自動化し，システム全体の効率化を実現した．即ち，キーワード自動抽出用プログラムをシステムに組み込み，キーワード抽出からインデックス作成までを一貫してシステム側で自動化するようにした．本稿では，この自動化の方式について述べる．

２．情報検索サービスにおけるインデックス

文献情報，新聞記事情報データベースを用いてキーワード検索を行なう情報検索サービスでは，インデックスは“ＩＶＬ（InVerted List）”と呼ばれる構造を有する（図１）．ＩＶＬの「タグ部」の“索引語”には原文より抽出されたキーワードが登録され，「サブ部」にはＲＩＤ（原文を含むレコードのＩＤ）が登録される．

３．データベース構成

文献情報の場合を例に，サービス用データベースのテーブル構成を示す（図２）．我々のシステムではデータベース管理システムとして，マルチメディアデータベース管理システムを利用している[1),2)]．これは，マルチメディア情報をも管理することを目的に，テーブル内での長大データ（図形／画像情報や文献情報の原文等が格納対象となる）の定義を可能なようにリレーショナルモデルを拡張したものである．

また，キーワード抽出結果を格納する“抽出結果格納ＤＢ”と呼ぶ別テーブルを用意した（図３）．当該テーブルは，カラム属性が長大データである，キーワード抽出結果リスト／分析不能語リスト／同義語リストより構成され，それぞれ以下の内容が格納される．

タグ部

索引語	RID数
ｺﾝﾋﾟｭｰﾀ	520
ﾃﾞｰﾀﾍﾞｰｽ	780
⋮	⋮
ﾜｰｸｽﾃｰｼｮﾝ	215

サブ部

RID群			
#001	#003	----	#715
#003	#005	----	#935
⋮	⋮	⋮	⋮
#002	#004	----	#360

図１．ＩＶＬの構造

文献タイトル	著　者	発表年	-------	キーワード*1				原文（長大ﾃﾞｰﾀ）*2
				ｷｰﾜｰﾄﾞ1	ｷｰﾜｰﾄﾞ2	-----	ｷｰﾜｰﾄﾞn	
情報システム	Ａ氏	1989	-------	ﾃﾞｰﾀﾍﾞｰｽ	ｺﾝﾋﾟｭｰﾀ	-----	情報処理	原文Ａ
知能と知識	Ｂ氏	1988	-------	人工知能	ＡＩ	-----	知識処理	原文Ｂ
⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮
OAとﾃﾞｰﾀ処理	Ｚ氏	1983	-------	ＯＡ	帳票処理	-----	ﾃﾞｰﾀﾍﾞｰｽ	原文Ｚ

*1：「キーワード」；原文に対してキーワード抽出を行なった結果として得られたキーワードが格納されるカラムであり配列構造である．この配列の個々の値がＩＶＬのタグ部の索引語となる．
*2：「原文」　；キーワード抽出対象となる原文であり，長大データとして管理される．

図２．サービス用データベースのテーブル構成例

ＲＩＤ	キーワード抽出結果リスト	分析不能語リスト	同義語リスト	
#003	ｺﾝﾋﾟｭｰﾀ,情報処理,人工知能,ﾃﾞｰﾀﾍﾞｰｽ,知識処理,ＯＡ	ﾊﾞｲｵﾁｯﾌﾟ,ﾆｭｰﾛｺﾝﾋﾟｭｰﾀ,ﾊﾞｯｸﾌﾟﾛﾊﾟｹﾞｰｼｮﾝﾓﾃﾞﾙ,ﾊｲﾊﾟｰﾃｷｽﾄ	ｺﾝﾋﾟｭｰﾀ ﾃﾞｰﾀﾍﾞｰｽ	ｲﾝﾌｫﾒｰｼｮﾝﾌﾟﾛｾｼﾝｸﾞ ﾃﾞｰﾀ管理,ﾃﾞｨﾚｸﾄﾘ
⋮	⋮	⋮	⋮	

図３．抽出結果格納ＤＢのテーブル構成例

A Method for adapting automatic keywords extraction to I-R System
Masanori ISHIGURO, Hidetsugu AKIYOSHI, Kazutoshi TANAKA
NTT Communication and Information Processing Laboratories

・キーワード抽出結果リスト：キーワード抽出用辞書から抽出されたキーワード（名詞）の一覧が格納される．
・分析不能語リスト：原文から名詞として検出されたが，キーワード抽出用辞書に登録されておらずキーワードとは扱われなかった名詞の一覧が格納される
・同義語リスト：キーワード抽出用辞書から抽出されたキーワードに対して，同義語辞書から当該キーワードに対する同義語として抽出された名詞の一覧が格納される．

4．実現方式

システムの全体構成並びにレコードの追加～キーワード抽出～インデックス作成までの処理の流れを図４に示す．

キーワード抽出用辞書や同義語辞書においては，学術分野の追加、技術の発展等に伴って新語が逐次追加されなければならない．このため，分析不能語として扱われた名詞のうち必要な名詞（用語）をキーワード抽出用辞書／同義語辞書に適宜追加するユーティリティを用意した．これにより各辞書を最新の状態に維持でき，キーワードも最新のものを抽出させることが可能となった．

5．評価・今後の予定

本方式により，従来人手により行ってきたキーワード抽出からインデックス作成までの作業をシステム側にて自動的に行うことを可能とし，作業の効率化を実現することができた．原文の入力に当たっては今後印刷漢字ＯＣＲをシステムに組み込むことによりさらに人手の軽減を図る．

また，ＩＶＬのようなインデックス構成を有するシステムの場合は，従来ロック期間が長くなること等によりデータベース更新は夜間のバッチ処理にて行われるのが一般的であったが，我々はロック期間を極小化させオンラインサービス中におけるデータベースの更新及びＩＶＬの更新を実現した[3),4)]．これにより，キーワード抽出からインデックス作成までをオンラインサービス中に実施することも可能となる．これについては，今後検索処理への影響（応答時間等）の面から検証する予定である．

［参考文献］

1)村田，寺中他：マルチメディア・データベースに向けたデータモデルについて，データベースシステム研究会（43-1）1984.9.17
2)岸本，長浜他：マルチメディアＤＢＭＳにおける長大データ管理について，データベースシステム研究会（52-1）1986.3.10
3)坪井，川下：情報検索システムに於けるインバーテッドファイルのオンライン更新実現手法，情報処理学会第３３回全国大会（1H-2）
4)石黒，青木：情報検索システムにおける検索結果の矛盾回避手法，情報処理学会第３３回全国大会（1H-3）

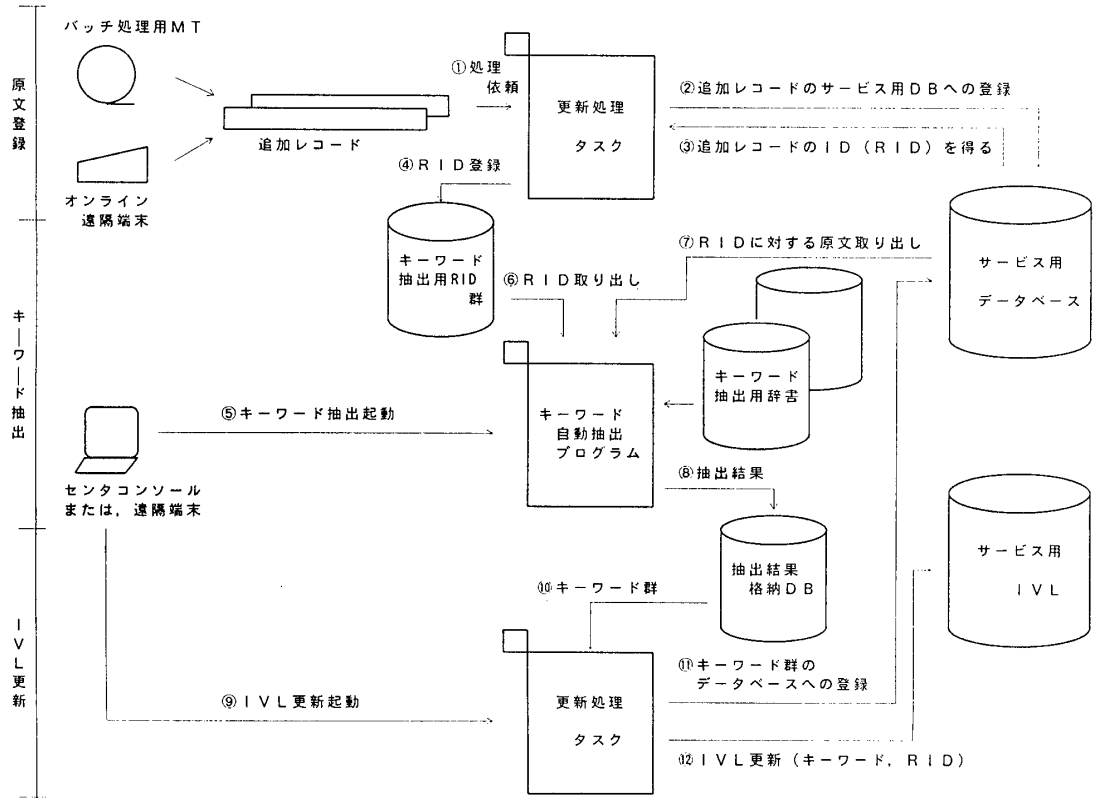

図４．システム構成／処理の流れ