情報検索システムの動向

東京芝浦電気株式会社 半 沢 久 生
齋 藤 孝

コンピュータによる情報処理の一分野として発達している、主として言語情報の蓄積と検索を目的とする情報検索 (IR) システムの現状について、アプリケーションの理論とプログラム言語およびハードウェアシステムの観点で展望した。

1. ま え が き

情報化社会は雑多な情報の氾濫の時代ともなる。特に、言語情報を表現する出版物の増加は著しい。この多量の言語情報の管理をいかにするかが、今後の情報処理の課題ともなっている。言語情報の処理に限っても、ドキュメンテーション活動に見られるように、必要な情報を含む文献の収集、情報化を目的とするファイル編成、適合情報を含む文献の検索、などのシステムの設計と開発が要求されている。今日、このシステムを総称して情報の蓄積と検索 (Information Storage and Retrieval, 略して IR) システムと呼ぶようになった。

本稿は IR システムの理論とその応用によるソフトウェアシステムの例を論じ、IR を目的とする周辺装置および専用機器について展望するものである。

2. IR システムの理論

2.1 IR システムのタイプ

一般に、IR システムは取扱う情報単位によって次のタイプが考えられる。

- (1) 事柄検索 (Fact Retrieval) システム
- (2) 文献検索 (Document Retrieval) システム

事柄検索は在庫管理、人事管理、予約システムに見られるように、データそのものの値を情報の単位とするものであって、「アルミニウムの融点は？」という問合せに対して「約 659°C」と回答する。また、「動物園にトラがいる」と「東京に動物園がある」という前提を与え「東京にはトラがいるか？」という問合せに対し、推論により「はい」と回答するなどの例がある。

一方、文献検索は「絶対零度とは？」に関して「理化学辞典の 745 ページを参照せよ」といった文献の出典事

項を回答する。

事柄検索は概して回答が一意的に決定するので、確定的なシステムモデルを準備し、それに対し文献検索は適合する情報を含む文献をできるだけ網羅的に取り出すとするので、確率的なモデルを準備することになる。

2.2 IR システムの機能

システムの基本的な機能を抽象的に示せば、図1のように入力として大量のデータを与え、それをファイルしておき、必要なときに情報として出力することである。

これは具体的には、ある主題の情報を集め、これをファイルして貯わえておき、問合せにおいてファイルから適合情報を検索するという、ファイルとその要素であるレコードと、さらにその要素であるアイテムの加工機能といえよう。

2.3 システムの手順

IR システムの手順は、理想的には人間の記憶の方法

図1 システムの抽象的機能

図2 IR システムの機能手順

"Latest Trends of Information Storage and Retrieval System" by Hisao Hanzawa and Takashi Saito (Tokyo Shibaura Electric Co., Ltd., Tokyo)

と記憶の取出し照合における思考の演算、つまり知識活動に似せ図2のようになる。ここでの特色は、入力となるデータそのものが言語による意味情報であることである。意味情報は文献のレコードでも明らかなように、日常のことば、いわゆる自然語 (natural language) によるものである。当然、この点で数値データの処理手順と違ってくる。

2.4 自然語の処理

意味情報を表現する自然語を対象とする IR システムは、情報の主題内容を指示する手段として指標 (index) を用いる。指標とは単行本の索引ページにある見出し語とその文脈の所在の関係の例のように、キーワードとか索引語と呼ぶもののことである。これらの索引語は、自然語かまたは、それにもとづく記号を採用している。そこで、主題を抽出し指標に変換することを主題分析 (subject analysis) という。主題分析は人事記録などで社内歴、学歴といったアイテムの抽出や、また学術論文のように何を目的とし、いかなる方法で、どんな結果を得たかを著者の専門領域のことば (terminology) により記述したものから主題に関係するキーワードを抽出するといった操作である。

2.5 情報主題の構造

記録された情報は、内容的にも形式的にも一見種々雑多であるが、よく見ると部分的に一定の形式化を試みることができる。化学情報などでは、原料、反応、生成物、触媒、測定、操作といったアイテムが抽出され、これが情報主題の構造を示す要素となる。図3のUDC (国際10進分類表)などは概念の階層性による表現である。

図3 UDC による概念の階層性による構造

図4 キーワードの関連地図による構造

自然語キーワードにより情報を示す EJC (Engineer Joint Council) シソーラスや関連地図 (図4) なども主題の構造を表現しようとする試みである。

3. IR システムの自動化の手法

自動化の手法として文献検索システムの全体図を示し、機能とその構成を説明する。図5は TOSBAC コンピューター用に開発された ALISS¹⁾ の機能を示している。図5において、情報の蓄積段階で1次情報とは文献そのものであって、2次情報とは主題構造の抽出による索引語への変換後の蓄積ファイルのことである。2次情報を作成する技術として次の手法がこれまで研究されている。

- (1) 自動索引法 (automatic indexing)
- (2) 自動抄録法 (telegraphic abstracting)
- (3) 自動分類法 (automatic classification)

3.1 自動索引法

自動索引法は H. P. Luhn²⁾ の提案による KWIC (Key Word-In-Context) 索引が有名である。KWIC 索引は図6の例のように、キーワードを文脈中に保存し、一定の位置に配列してリストする。これによって自然語の欠点となることばの同意、同音異義、多義性などによる次の相違を前後関係によって解決しようとする。

- (1) 概念の相違…OIL (COOKING), OIL (MOTOR)
- (2) 結合の相違…BELT CONVEYER, CONVEYER BELT

このアイディアの索引は、シソーラスなど、ことばの定義辞書を必要としないので2次情報化が速く、しかも

図5 ALISS システムの機能構成

TOSBAC-3400 IR SYSTEM * A L I S S * IR SYS
** TOSHIBA INFORMATION CENTER

[illegible]

9

表 1 2×2 分割表による関連係数の例
2×2 分割表

	<i>a</i>	$\bar{a}$	総 計
<i>b</i>	f_{ab}	$f_{\bar{b}-f_{ab}}$	f_b
$\bar{b}$	$f_{a-f_{ab}}$	$N-f_{a-f_{ab}}+f_{ab}$	$N-f_b$
総 計	f_a	$N-f_a$	N

f_a : キーワード *a* の単純生起頻度

f_b : キーワード *b* の単純生起頻度

f_{ab} : キーワード *a* と *b* の共出生起頻度

N : 文献数

A_{ab} : キーワードの関連係数

1) Stiles による関連係数

$$A_{ab} = \log_{10} \frac{(|f_{ab}N - f_a f_b| - N/2) \cdot N}{f_a f_b (N - f_a) (N - f_b)}$$

2) キーワードの共出生起頻度と期待値の差による関連係数

$$A_{ab} = f_{ab} - \frac{f_a f_b}{N}$$

3) 標準偏差による関連係数

$$A_{ab} = \frac{(f_{ab} - \frac{f_a f_b}{N})}{\sqrt{\frac{f_a f_b}{N}}}$$

簡単であるという理由で広く普及している。

3.2 自動抄録法と自動分類法

これは自動索引法が実用化されているのに対し、まだ実験段階の手法である。自動抄録法の手法は、仮説として「文献の主題に関係の深いキーワードは、その文献中に出現する頻度も高い。そこでこのようなキーワードを一文の中に数多く含む文は、主題を表現する可能性が高い。このような文を配列すれば一種の抄録となろう」ということを利用する。この統計的手法に変形文法³⁾などの文法理論、さらに意味論的手法を加えて改良した実験を Edmundson⁴⁾などが試みた。

自動分類法⁵⁾も統計的手法を利用し、あらかじめ設定したカテゴリーとキーワードの相関係数により文献を分類しようとする。これらの手法を実際のシステムに応用した例では、関連係数 (association factor) によるキーワードファイルの管理システムがある。関連係数とは Stiles⁶⁾などが定義したものであって、文献とキーワードの関係づけを次のようにとらえて考え出している。

つまり、相互に関連するキーワードを中心において文献を通して見たとき、その文献中に同時に生起するキーワードがあれば、その間に何らかの意味的な関連があると仮定できるとする。表 1 は 2×2 分割表による関連係数の例である。この関連係数の利用による検索効率の向上を試みたものとして SMART⁷⁾ システムが有名である。

4. IR プログラミングシステムの動向

先の理論、手法の展望により明らかのように、IR システムのプログラミング技術は、次のものが主体となる。

(1) 2 次情報ファイルの作成に関して

主にファイルの編成方式に関する技術であり、次のようなものがある。

- (a) 直接ファイル (direct file)
- (b) 逐次ファイル (sequential file)
- (c) 木構造ファイル (tree structure file)
- (d) インバーテッドファイル (inverted file)

(2) 問合せ処理

これは論理演算や問合せ言語の分析の技術である。

(3) ファイルの検索

蓄積ファイルを検索条件に従って探索する技術として次のものがある。

- (a) 逐次探索法 (sequential search)
- (b) 2 分探索法 (binary search)
- (c) 索引探索法 (index term search)
- (d) 自然語探索法 (natural language search)

これらの技術を目的とするプログラムシステムの動向を見てみよう。

4.1 ファイル処理のプログラムシステム

ファイルとは普通、磁気テープ、ディスクなど外部記憶装置に蓄積する。そこで、これらの媒体を能率よくコントロールし、構造などの定義を簡単に与え、目的とするファイルを準備してくれるようなソフトウェアが要求される。一般に、磁気テープを利用するファイルは、情報の順序付けがきまってしまうので操作は単純であるが、磁気ディスク、ドラムなど乱処理 (random access) が可能なファイルは、その管理方法によりシステムの効率が変わる。

オンライン IR システムに必要な直接ファイルなどはディスク上に展開する。そこで、最近では高水準のファイル管理プログラム言語として IDS⁸⁾ や GIS⁹⁾ などが開発されるようになった。これらの言語の主機能は次の 4 つとされる。

- (1) ファイルの生成 (file creation)
- (2) ファイルの更新 (file maintenance)
- (3) ファイルの検索 (file search)
- (4) レポートの作成 (report generation)

4.2 問合せ処理プログラムシステム

検索とは基本的には問合せ言語を準備しておいて質問文を作文し、ファイルから必要なアイテムを探索することである。図 7 は事柄検索システムである QUERY¹⁰⁾ の問合せ言語と構文分析の手順を示したものである。

4.3 ファイル検索システムの構成

これまでのプログラム技術を応用してファイル検索のシステムは図 8 のように構成される。まず、問合せ言語による質問文を読取った後、図のような構文分析手順に

問合せ質問文例：Count all planse on bases within 10 miles of Tokyo.

図 7 問合せ言語の構文分析の手順図

図 8 ファイル検索の概略フローチャート

より条件パラメーターを解釈する。次に照合演算が容易なような記号列にキーワードと論理演算子の記号式を変換する。これを逆ポーランド記法への変換と呼ぶ。それからファイルの検索と内容の探索に入り、適合したレコードを出力する。

これらの IR システムのプログラム技術については、斉藤の論文¹¹⁾が考察しているので参照されたい。

5. 最近のIRシステムのためのハードウェア

これまでの説明のように、IR システムの基本はファイルの準備と管理にあるといえよう。そこで、ハードウェアについてもファイルの蓄積媒体となる各種の2次記憶装置の進歩に注目する。最近はいくつかの磁気によるファイル化装置から、フィルムなどオプティカルな媒体を使用する傾向が強い。特にマイクロフィルムの利用は著しいものがある。

5.1 フィルムファイル

マイクロフィルムを IR の蓄積媒体として利用した例は、コンピューターの出現以前にもかなり古い歴史がある。しかし、これまでのフィルムファイルの作成は、文庫をカメラで撮影していき、現像処理というエレクトロニクスと異質のオペレーションがあること、また縮小化

図 9 COM のハードウェアシステムの概要

しても検索手段が貧弱な場合は保管に苦勞するなどという理由により、その用途は限定されがちであった。1969 年代に実用化した COM (Computer Output Microfilm) 装置の出現は、このような障害を取り除き、コンピューター周辺機器としてのスタイルをとり、フィルムファイルのオンライン、オフラインによる作成を可能にした。

COM 装置を簡単に示すと図 9 のようなシステムとなる。その機能は磁気テープの内容により制御部を働かし必要なデータを CRT 上に描かせ、フィルムに焼きつける。この過程は CRT 装置に光学系装置を付加したものと単純に考えてもよい。COM の詳細な機能およびソフトウェアについては文献¹²⁾を参照されたい。

COM 装置により索引ファイル、文献ファイルなどを直接 16mm サイズのロール状のフィルムにしたり、105 mm サイズのフィッシュフィルム (写真 1) にして保管し、オフライン検索が可能になった。

フィルムファイルと磁気ファイルを組合せて情報の蓄積媒体に利用し、オンラインリアルタイム文献検索システムを実験開発しているのに MIT における INTREX

(Information Transfer Experiments)¹³⁾ が知られている。INTREX システムは図 10 に示したように、図書館の遠隔地からの分散利用を目的とする将来の図書館の自動化のモデルである。

図書など文献のオリジナル情報は、すべて磁気ファイルにすることなくマイクロフィッシュを利用してフィルムファイルとして蓄積しておく。そして、問合せにおいて端末より会話モードで磁気ファイル中の 2 次情報の主題検索を行なう。問合せに適合した文献の書誌的事項や標題などをキャラクターディスプレイに表示し、必要とあらば、フィッシュ上のオリジナルを Microfacsimile¹⁴⁾ により画像を送信して端末のリーダーに表示する。

このように、INTREX は、フィルムと磁気ファイルの利点をうまく結びつけ、リアルタイム文献検索を試みているわけである。

今後、フィルムファイルは CIM (Computer Input Microfilm)¹⁵⁾ の開発により、フィルム上のアナログ情報を再びデジタル化してコンピューターに入力することが可能になる日も近いであろう。

5.2 画像情報のハードウェアシステム

これまでの磁気ファイルは、写真、設計図などパターンイメージの蓄積には不向きである。しかし情報は文字による言語情報ばかり対象とするわけにいかず、これらのイメージのファイル媒体と検索装置が要求された。

Ampex 社が開発した VIDEOFILE システム¹⁶⁾ がその要求に答えるハードウェアシステムである。このシステムはファイルとして磁気ビデオテープを利用する。図 11 はビデオテープの記録フォーマットを示している。テープの幅は 2in で横 1/3in 幅に A4 判サイズの画像を収容できるようにしてある。情報の蓄積は走査線が 1280 本のテレビカメラにより撮影し、ファイル化し、検索に

写真 1 マイクロフィッシュフィルムファイルの例

図 10 INTREX システムの概略図

図 11 VIDEOFILE 磁気テープのファイルフォーマット

図 12 VIDEOFILE システムの構成

図 13 GESCAN Model 6J101 の系統図

においては、コマの上部にある最大 19 文字のアドレス情報の内容を探索しディスプレイ上に写しだす。ファイルの容量は 4,800 ft に最大 166,600 ページ分の画像を収容できる。検索スピードは約 4 万ページ/min となる。

図 12 は VIDEOFILE システムの構成を示している。

5.3 その他のハードウェアシステム

その他のハードウェアシステムの動向として、IR 専用の磁気ファイル検索装置である GESCAN¹⁷⁾ (図13) などが開発されている。また、漢字処理を目的とする漢字高速度プリンターの開発も 1970 年以降著しい進歩があった。

IBM 社、富士通、高千穂交易、さらに日本電子産業などの漢字処理装置の発表は、日本語 IR システムの実用化を促進しそうである。これについては高橋の論文¹⁸⁾が詳細に論じている。

漢字 IR システムに関しては、現在入力オペレーションに関係する省力化、簡易性、経済性の解決を試みる傾向が強い。

6. む す び

今後、IR システムは新しい概念で、しかも安価で容量の多いファイル媒体の開発と、連想記憶 (associative memory) など、検索に有利なハードウェアの開発が要求されると同時に、常に情報の中味となる主題の構造、ことばの研究など、論理学、言語学、心理学に及ぶ広い領域の応用を必要としている。特に、日本語による IR システムなどは漢字の入出力装置の開発以上に、分ち書きの問題、漢字内部コードの規格化などの問題の解決が

熱望されるであろう。

最後に、本稿で論じなかったが、すでに実用化された大規模な IR システムとして、日本では JICST の IR システム¹⁹⁾ があり、米国において医学の MEDLARS²⁰⁾ などが有名であることを参考までに示しておきたい。

(昭和 46 年 11 月 11 日受付)

〔参 考 文 献〕

- 1) 斎藤孝: TOSBAC による IR システム ALISS, 東芝レビュー, 26, 4(1971) 485-492
- 2) H. P. Luhn: Keyword-In-Context Index for technical literature, Amer. Documentation, 11 (1960) 288-295
- 3) N. Chomsky: Syntactic Structure, Mouton, the Hague (1957)
- 4) H. P. Edmundson: Automatic abstracting and indexing survey and recommendation, Comm. ACM, 4 (1961) 226-234
- 5) H. Borko: Automatic document classification, J. ACM, 11, 2 (1963) 151-162
- 6) H. E. Stiles: The association factor in information retrieval, J. ACM, 8, 2 (1961) 271-279
- 7) G. Salton: The SMART system retrieval result and future plane, Cornell Univ., Scientific Report No. ISR-11(1966)
- 8) 東京芝浦電気: TOSBAC-3400 IDS 解説書, TOSBAC アプリケーションライブラリー No. 34AP410C (1970)
- 9) IBM system 360—Generalized information system application description manual (1968)
- 10) T. E. Cheatham: Translation of retrieval requests couched in a semi formal English-like language, Comm. ACM, 3 (1962) 34-39
- 11) 斎藤 孝: 自然言語の処理を目的とする IR システムのプログラムの考察, Lib. and Infor. Science, 9 (1971) 439-458
- 12) 斎藤 孝: COM 専用ソフトウェアの機能と実際, 事務と経営, 23, 279 (1971) 43-47
- 13) D. R. Knudson: Remote text access in a computerized library information retrieval system, Spring Joint Computer Conf. (1969) 475-481
- 14) D. M. Costigan: Microfasmile: a status report, J. Micrographics, 4, 4 (1971) 189-199
- 15) 川田 陽一: インプットにおけるマイクロフィルムの利用—CIM への期待, Computer Report, 10 (1970) 36-41
- 16) 原科 茂: VIDEOFILE Information system とその応用, 情報科学, 7, 9 (1971) 20-26
- 17) General Electric Co.: GESCAN—The rapid search machine—operator's manual
- 18) 高橋 達郎: 日本語の機械処理—コンピュータによる印刷革命—, 東洋経済新報社 (1970)
- 19) 日本科学技術情報センター: JICST 総合 IR システム—漢字モード IR システム説明書 (1969)
- 20) C. J. Austin: MEDLARS 1963-1967, National Library of Medicine (1968)