情 報 学 基 礎 25－6
（１９９２．５．１２）

ＳＡＶＶＹ／ＴＲＳによるドキュメントデータベース

荒井　敏博

日軽情報システム株式会社

ＳＡＶＶＹ／ＴＲＳは、米国エクスカリバーテクノロジー社により開発されたパターン認識技術に基づく高速な文字列検索ソフトウエアーである。ＳＡＶＶＹ／ＴＲＳは、テキストデータを学習することによりパターン索引ファイルを自動的に作成することが可能であり、そのファイルにより文字列パターンを高速に検索することが実現できる。また、検索された各テキストはパターン認識の結果として類似度が与えられ、これにより完全に一致しない文字パターンの検索も可能となる。本報告では、ＳＡＶＶＹ／ＴＲＳを用いたドキュメントデータベースシステムを紹介する。

Document Database using SAVVY Text Retrieval System

Toshihiro Arai

Nikkei Information Systems Co.,Ltd.

Isuzu Shiba Bldg., 4-2-3 Shiba, Minato-Ku, Tokyo 108, Japan

Savvy/TRS developped by Excalibur Technologies is high-speed text retrieval software based on pattern recognition technology. Savvy can automatically generate Pattern Index Network file for later text retrieval using machine learning method. In addition, similarity measure will be given for each found record as a result of pattern recognition. Because of the measure, TRS can find a record which does not have perfect much string but has similar one. In this report, I will introduce document database system which has SAVVY/TRS as text retrieval engine.

1．はじめに

近年のコンピュータの処理速度の向上に伴いデータベースシステムの機能も向上している。しかしながら、文字列の検索機能は、コンピュータの処理資源を大量に必要とするため多くの場合大型コンピュータまたは、スーパーコンピュータ等にて実用システムが稼動している。また、文献検索等の場合各文献データにキーワードを設定することにより、データベースのサーチ構造を実現し高速な検索機能を提供しているものがある。この場合、キーワードのメンテナンスの手間及びフリーワードの検索でない点に不満が残る。

ＳＡＶＶＹ／ＴＲＳは、パターン認識技術の応用により柔軟でしかも高速な文字列の検索機能を安価なワークステーション上に実現している。また、特別なハードウエアー等を必要としないため、様々なシステムと統合して使用することができる。

ここでは、ＳＡＶＶＹ／ＴＲＳの概要を説明し、その検索機能を採用したドキュメントデータベースシステムを紹介する。

2．SAVVY／TRS

ＳＡＶＶＹ／ＴＲＳは、テキストデータの学習機能により、ＰａｔｔｅｒＩｎｄｅｘＮｅｔｗｏｒｋと呼ばれるファイルを自動的に作成する。このファイルにより学習したテキストデータの高速検索を実行する。概要を図１に示す。検索結果には、与えられた検索文字列を含むテキストデータまたは似たテキストを持つデータが含まれている。検索されたデータには、それぞれ類似度が付与される。図２に学習したデータとその検索結果及び類似度をサンプルプログラムの結果から示す。図２にあるようにＴＲＳは、完全に一致する文字列を検索するだけでなく、パターンの検索を実行している。これにより、柔軟な検索を実現することができる。

ＳＡＶＶＹ／ＴＲＳの最も重要な特徴は、その検索スピードにある。少なくとも順番に文字列検索を

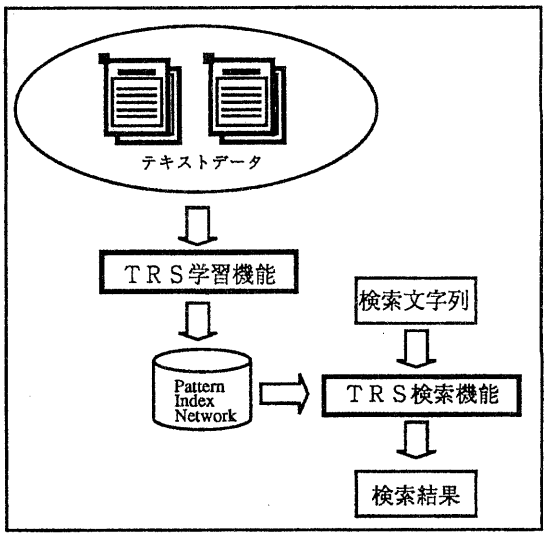

図１．ＴＲＳの動作概要

```
※学習を開始します。

<   1> <シンボル，文字を認識する画像処理>
<   2> <画像処理システムによるロボット制御>
<   3> <物体の確認をニューラルネットにて実現>
<   4> <右脳と左脳の役割分担についての最新情報>
<   5> <コンピュータ処理によるパターン認識>

※ネットワークファイルが作成されました。


検索文字列を入力> 画像処理

 検索文字列は <画像処理> です。
 検索数は <3> 件です。

Ｉｄ 類似度   検索された文字列
 1    828    シンボル，文字を認識する画像処理
 2    828    画像処理システムによるロボット制御
 5    348    コンピュータ処理によるパターン認識
```

図２．ＴＲＳの検索結果例

行うアルゴリズムと比較した場合数百倍の検索スピードとなる。２番目の特徴は、言語の構造等に依存しない一般的なパターンプロセッサーとして開発されている点である。このため、学習したあらゆる文字列パターンを検索対象とすることができ、日本語、英語、ドイツ語等また、化学式のような文字列

パターンの検索も可能である。学習は、可変長のテキストレコード毎に行うことができ、新しいデータの学習は、単なる追加として処理される。また、学習したテキスト毎に学習内容を消去させることができる。ただし、ガベージコレクションは動的に実行されない。

3．SAVVY／TRS　vs　キーワード

ここで、文献検索のシステムにてよく使用されたキーワードによる検索とＴＲＳによる検索を比較する。

キーワードによる方法の場合、あらかじめ各文献に検索のためのキーワードを設定しておく必要がある。通常以下の２通りの方法にておこなわれている。

１）人手にて文献にキーワードを付与

２）文献に含まれる文章よりキーワードを抽出

後者の場合も、キーワード抽出のための辞書をメンテナンスする必要がある。また、人名がキーワードとして考えられた場合、検索漏れを少なくするために［沢田］、［研二］、［沢田研二］、［沢田　研二］などの重複したキーワードを設定する場合が多い。これは、キーワードにより検索構造を設定する場合避けるのが困難な問題である。これに対して、ＳＡＶＶＹ／ＴＲＳでは、プログラムが効率の良い検索構造を自動的に作成することになる。ただし、ＴＲＳの場合文字列を純粋なコード列として扱っているため、［人工知能］と［ＡＩ］が同じ意味であることを理解できない。このため、シソーラス辞書の展開機能と合わせて使用するとより良い検索が実現できる。この場合は、シソーラス辞書のメンテナンスが必要となる。最近では、各領域毎にシソーラスデータを開発／販売している会社も見うけられる。

4．ドキュメントデータベース

先に概説した全文フリーワードの文字列検索エンジンを採用したドキュメントデータベースを紹介する。これまでのドキュメントデータベースは、電子ファイルシステムに見られるように、いかに情報をファイルするかという点を考えて開発されていた。これに対して今回開発したシステムは、より積極的に情報の再利用が可能なように考慮されている。その基本データ構造を図３に示す。

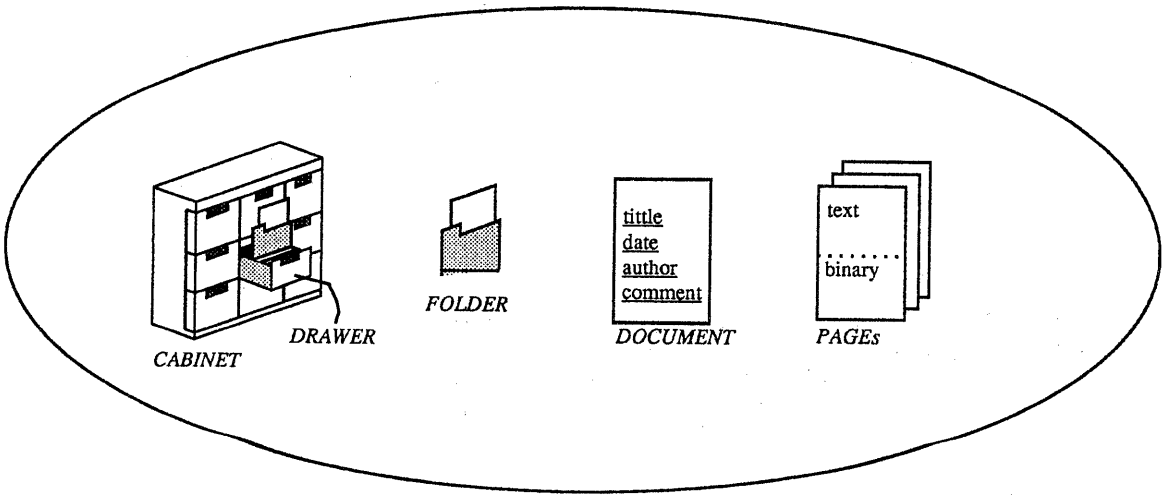

図３．ドキュメントデータベースの構造

ＦｉｌＲｏｏｍは、最も基本的なオブジェクトであり、このシステムは、ＦｉｌｅＲｏｏｍ単位に分散処理が可能なサーバ／クライアント方式にてインプリメントされている。これにより以下の利点を受ける事が出来る。

1）システム規模の拡張が容易

2）部門／地域間の分散処理が可能

3）異機種マシンの接続が容易

また、通信プロトコルには、ＵＮＩＸ上にて標準となっているＴＣＰ／ＩＰを採用している。

４．１　各オブジェクト

ＦｉｌｅＲｏｏｍ内には、通常オフィスにてファイリングのために使用されている各種オブジェクトが用意されており、それぞれの名称にて管理される。それらは、ＣＡＢＩＮＥＴ、ＤＲＡＷＥＲ、ＦＯＬＤＥＲである。ＤＯＣＵＭＥＮＴは、名前及び表題、日付等の属性データを保持しており、さらに複数のＰＡＧＥにて構成されている。ＰＡＧＥ内は、２つの部分に別れておりそれぞれテキストデータの領域とバイナリデータの領域である。テキストデータは、各ページの説明文、要約等を保持するためのもので後の説明にあるとおりＳＡＶＶＹ／ＴＲＳによる全文フリーワード検索を実行することができる。バイナリーの領域は、イメージデータ、グラフ、ＣＡＤ図面等の各種データを保持する部分である。

４．２　検索の方法

当システムでは、現在３つ方法にて各種データへアクセスすることが可能である。第１の方法は、図４にあるように各オブジェクトの構造に従いアクセスする方法である。これは、通常のオフィスでの操作をコンピュータ端末上にて行うことである。また、ドキュメントの属性データによる検索も可能である。ＰＡＧＥ内のテキストデータにたいしては、ＳＡＶＶＹ／ＴＲＳによる高速パターン検索を実行することができる。ＴＲＳの学習は、ドキュメントの登録時にオンラインで実行されるので、当システムの使用者は、データの学習等を意識する必要はない。これらの情報検索機能により、情報資産の活用が促進されると思われる。

また各種データからテキストデータさえ抽出できれば、簡単にテキスト検索の構造を実現できる。つまりデータフォーマットの公開されているデータであれば、本システムにて一元管理することが可能である。例えば、ＣＡＤ図面、ＤＴＰドキュメントなども、当システムに登録し検索／再利用することができる。図５の様式にて検索構造の設定がかなり容易になる。

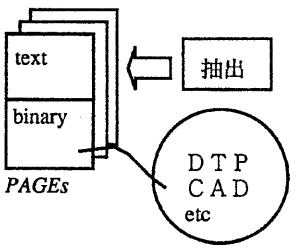

図５．テキスト検索機能の活用

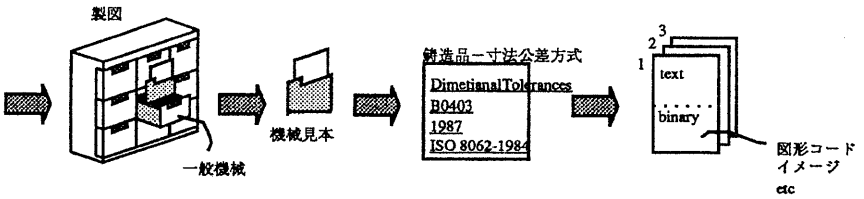

図４．キャビネット構造によるアクセス

4．3　その他の主な機能

本システムは、検索機能の他にもいくつかの機能、特徴があり、それを簡単に解説する。

4．3．1　機密保護

本システムには、パスワードによるシステムへのアクセス制限を設定することが可能である。また、ＦｉｌｅＲｏｏｍ内の各オブジェクトに対して読み込み／書き込みの制限を設けることができる。これにより、自分専用のキャビネットを作成したり、部門毎にアクセス可能な環境を設定することができる。

4．3．2　マルチサーバー

クライアントは、同時に複数のサーバーと処理が可能なため、遠隔地にある複数のサーバーを同時にアクセスすることが可能であり、組織体の持つ情報を横断的に検索することができる。

4．3．3　オープン指向

現在、ＵＮＩＸにて稼動するソフトウエアー及びハードウエアーにおいてオープン指向が一つの潮流になっている。これは、各種のソフトウエアー／ハードウエアーを統合するためのコンセプトである。本ソフトウエアーでも、サーバーのインターフェース規約を公開することにより、他システムとの柔軟な接続を指向した。

4．3．4　シソーラスサーバー

先に説明したＳＡＶＶＹ／ＴＲＳの欠点を補うために、同義語、狭義語、広義語を展開するためのシソーラス展開機能を提供している。

4．3．5　ユーザーインターフェース

クライアントプログラムの操作は、文字列の入力以外は全てマウスにより簡単に実行できる。またオンラインヘルプ機能により誰もが簡単に各機能を使用することが可能である。

5．まとめ

昨今のコンピュータの普及に伴い、ワードプロセッサー、ＤＴＰ、ＣＡＤ等のソフトウエアーにより多くの電子メディアが作成されている。そのような情報資産を有効活用するために、今後電子メディア特に不定形なデジタル情報のデータベースが必要となることが予想される。

本報告では、パターン認識技術に基礎を置く高速文字列検索サフトウエアーＳＡＶＶＹ／ＴＲＳの概要を説明した。また、ＳＡＶＶＹ／ＴＲＳを検索機能に採用したドキュメントデータベースを紹介した。当ソフトウエアーは、不定形のデジタル情報の管理に的していると考えている。

参考文献

なし