# 全文検索技術とその応用について

プロジェクト推進事業部 営業企画部
山 本 崇 敏

## はじめに

全文検索(フリーワード検索)技術は汎用機全盛の時代から存在したが，その利用は一部ユーザーに限定されていた。それが1990年代初頭からオープン系システム上で構築され，さらにインターネットの普及により，大量に蓄積されたデータから必要に応じた情報を探し出す仕組みとして再度注目されている。

しかし多数の関連ソフトウェアが製品化されるに至っても，多種多様の全文検索について十分な理解が得られているとは思えない。そこで全文検索システムについて何を評価し，どういった問題点が潜在しているか，さらに現在の市場動向(但し，フリーソフトについての言及を除く)について，著者の経験から述べてみたいと思う。

今回の報告が，全文検索技術について理解を得る一助になれば幸いである。

## 1. 全文検索技術について

### 1.1 全文検索ロジック

全文検索技術とは，自由な文字情報を検索条件としてその文字が含まれる電子文書を探し出す技術である。これについてはWWWでYahoo！やInfoseekなどの検索サイトを利用されている方ならば，特に説明は要らないと思う。

しかし全文検索ロジックとしては，大きく分けて以下の3つがある。

(1)テキスト文から高速検索用インデックスを作成する全文検索手法。(全文検索エンジン)

(2)テキスト文の配列を力まかせに解読する全文検索手法

(3)リレーショナルデータベース(RDB)のSQL機能による検索手法

これらのロジックによる検索スピードは(1)が(2),(3)に比べて桁違いに早く，一般に全文検索という場合(1)の手法を指す。事実あるベンチマークテストで著名なRDBと全文検索スピードを比較したところ，数千倍の検索スピード差が確認できた。一方全文検索エンジンの手法にも様々なインデックスロジックがあり，代表的なものの概要を以下に紹介する。

(1)構文解析型インデックス(形態素解析型)

対象が欧文だとすると，欧文には単語間に必ず空白(スペース)がある。その空白を頼りに単語を抽出して検索インデックスを作成する方法。日本語の場合は空白がない為，TEXT文から「て，に，を，は」などを外した語句を残す事により同様の処理を行う。しかしTEXT文解読の為の辞書機能が必要となり，多言語が混在する文書などを対象とする場合には検証が必要である。うまく辞書解読ができ，日本語に限っ

た場合は，“分ち書き”の完全一致検索ができるのが特徴である。

⑵文字成分表インデックス

TEXTファイル毎に含まれる文字(1文字情報)と連接文字情報や文字位置情報などの成分表(どのファイルにどういった文字が含まれているかの一覧表)を作成し，それらをインデックスとしてどの文字がどのファイルにあるかを照合する。1文字インデックスは構造がシンプルだがシステム上の照合回数が多くなるため，ファイルアクセスが検索スピードのボトルネックになり易い。

⑶n文字インデックス

文字成分表インデックスを複数文字単位で作成する。インデックスパターンは相当に増大するため1文字インデックスに比べて検索スピードは速くなるが，一般的にインデックスサイズは大きくなる。

⑷パトリシアツリーインデックス

全ての文字始点からの文字の並び(1文字ずらした文字列全て)を検索インデックスとし，文字のビット列の0or1を判別して概念的ツリーを辿る事で検索個所にたどり着くロジックである。ロジックが簡便で非常に高速であるが，データの追加が行われる場合にインデックス全体を再作成する為に登録時間がかかる。従って頻繁にデータ更新や追加が行われる様なデータベースでの利用には疑問がある。

⑸パターン認識型インデックス

ニューラルネットワークのパターン認識技術を利用して，文字配列パターンを探し出す技術。言語に依存せずに“似た文字列”まで探すあいまい検索が可能だが，文字配列パターンのインデックス精度が荒すぎると検索誤差(ノイズ)をひらう事がある。(通常はインデックスサイズは調整できる)

### 1.2 全文検索ソフトウェア

一般に全文検索関連の製品は，一元的な比較評価が難しい場合が多い。それは利用環境や製品構成及びそれぞれの特性が異なるからである。以下にその分類例とその考察を述べる。

製品構成としての分類

⑴システム開発用の全文検索エンジン(またはセミカスタマイズ製品)

プログラミング関数または開発ツールとして提供され，入出力機能／データ管理／操作画面／システムセキュリティ／ネットワーク通信などの機能は別途構築する必要があるもの。

⑵スタンドアロン環境でのみ利用が可能なもの

ネットワーク対応の機能が用意されていないため，特定のコンピュータ単独でのみ利用が可能な製品。簡便なものでは，エクスプローラのツールメニューに有る検索機能が例としてあげられる。

⑶全文検索機能オプション

特定のシステム(グループウェアやDTPソフ

ト)にアドオンされ,そのシステムの利用を前提とした全文検索機能の追加オプション。ノーツに全文検索機能を持たせるFulcrumの『DOCS/Fulcrum』(旧名FIND!)などが例としてあげられる。

(4)ネットワーク環境で即運用可能なシステムパッケージ

特にプログラム開発を必要とせずに導入後すぐに運用できる全文検索システム製品。少なくとも一般的なデータ登録機能・データ管理機能・ユーザー管理機能・ネットワーク対応及び検索操作画面までが用意されている事が必要となる。

検索エンジンによる分類

全文検索の著名な海外製エンジンとして,OpenTEXT, Fulcrum, Excalibur, Verityなどがある。

(1)OpenTEXT

『OpenTEXT/LiveLinkSearch』として販売されている。パトリシアツリーインデックスを作成して高速な全文検索を実現している。構造化文書のタグを考慮した検索を処理できるのが特徴で,タグ付きデータがよく利用される“図書検索”や“CALS/EDI関連”などで実績が多い。

(2)Fulcrum

『Fulcrum Search Server』として販売されている。n文字インデックスを作成して高速全文検索を実現している。特定のシステムに組み込まれたりOEM化されての実績が多い。

(3)Excalibur

『Excalibur RetrievalWare Text』として販売されている。もともと遺伝子解析用に開発されたニューラルネットワークのパターンマッチング技術を,全文検索技術に応用したエンジン。画像パターンマッチングなどの応用製品もある。

(4)Verity

国内ではオムロンが独自の形態素解析技術を付加し,『SEARCH'97』として製品化されている。

国内で流通している全文検索製品にもこれら海外製エンジンをOEMしたものや組込商品があり,全文検索エンジンの市場はそれらと純国産エンジンとで市場が形成されている。

国内製検索エンジンの主なものを以下にあげる。(フリーソフトなどを除く)

(1)『Future/Happiness』(平和情報センター)

形態素解析型の検索エンジンとして草分け的存在。

(2)『SAVVY/TRS』(日軽情報システム)

Excalibur Technology社の基本テクノロジーを日軽情報システム㈱がライセンス取得し,独自に純国産エンジンとして製品化したもの。1980年代後半には初期バージョンが出荷

されており，全文検索サーバの国内実績は現在もっとも多いと思われる。

(3)『PanaSearch』(松下電器産業)

検索速度をアピールしており，『Yahoo！Japan』で採択された検索エンジンの一つであり，大手ガス会社のイントラネット技術情報検索でも採用された実績がある。

(4)『NSEARCH』(新日本製鉄)

『RetrievalWareText』や『SAVVY/TRS』と同じく，パターンマッチング技術を全文検索技術に応用した新日鉄の独自開発エンジン。

(5)『Bibliotheca/TS』(日立製作所)

文字分析表技術を利用した国産の草分け的エンジン

## 2. 全文検索システム構築の考察

### 2.1 システム構築の検討課題

全文検索システムの概要モデルは，利用局面に応じて要約すると図1の様なものになる。

図中に表記するシステム要素は，システムの目的と状況により変化する。しかし全文検索システム構築時に(要不要を含め)検討すべき項目である。以下，項目毎に考察する。

(1)全文検索のデータ登録

①テキスト抽出

全文検索のインデックス生成のためには事前にTEXTデータが必要である。一般にドキュメントデータはワープロファイルや紙文書さらにはPDFファイルやWWW(HTML)などのデータとして蓄積されているため，それらを全文検索対象とするには図2のようなテキスト抽出という過程が必要である。

全文検索はTEXTデータに対する処理であり，どれだけ多様なマルチデータに対処できるかはシ

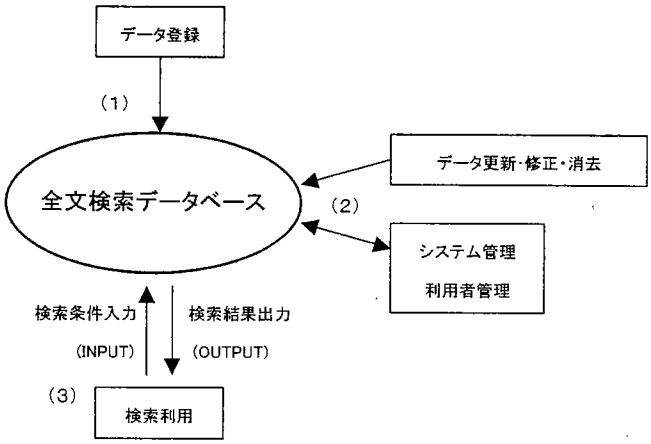



図1

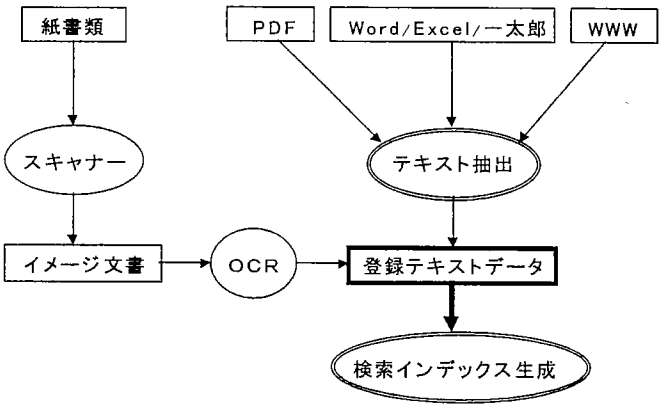



図2

ステム選択の重要なポイントである。そしてそのシステムがどういったデータまでを全文検索対象とできるかは，テキスト抽出の変換ソフトや仕組みがどこまで用意できるかにかかっている。ただしその登録システムの操作性は別途確認する必要があり，目的に応じた，「フィルタープログラムを用意しての一括登録」や「1ファイル毎にDorp & Drugしての登録(自動抽出)」などの手法が考えられている。

②バッチ登録と随時登録

文書管理の運用目的と内容によって登録方法は異なる。しかしいずれにしても以下の方法を組み合わせた登録方法を選択する事になる。

＊事前バッチ(batch)処理による一括登録

(長所)データ管理が容易。更新が必要でない大量データの登録に向いている。

(短所)随時更新したい情報の登録処理には向いていない。

＊登録権限を設定する事により，特定者による随時登録・更新

(長所)勝手な登録・更新を許さない為，情報の信頼性が保持できる。

(短所)権限管理の仕組みが必要になる。

＊利用者全員による自由な登録

(長所)全員による随時更新により，最新の情報が登録される。

(短所)情報の信頼性が薄く，データのメンテナンスも難しくなる。

③登録スピード(検索インデックスの作成時間)

全文検索システムへの登録データ件数が数十万件～数百万件に上ることはさほど珍しくなく，仮に書誌データ1000000件を一括登録するとしたら，数時間から数日間のインデックス作成時間が必要となる。

インデックス作成に時間が必要な点は検索エンジンによっても大差は無いが，しかし問題は一件の追加データにもインデックス全体を再構築するものがあり注意が必要である。この様な検索エンジンは大量かつ頻繁にデータ登録が行われるシステムでは利用できないと判断すべきである。

④ロボット型データ登録

WWWのHomePageから自動でのTEXTデータを取得し，全文検索対象として登録する仕組み。データ登録の手間が省けるため多くのWWW検索サイトで採用されており，膨大なインターネット上からのデータ収集を実現している。全世界のWWWをサポートするロボット型検索サイトでは新たなHomePageを更新しても即座に検索対象となる事はない(インデックス更新には通常数週間かかる事が多い)が，情報連携したい指定サイトからのデータ自動取得には有効な手段である。

⑵データ蓄積と管理

①検索インデックス

全文検索のための検索インデックスサイズは，登録される本体TEXT容量の数倍に及ぶ事が珍しくない。

特に2バイト系の日本語などを“n文字分析方式”でインデックス生成した場合などは，文字列の切り取り情報を綿密に取れば取るほどインデックスサイズは大きなものとなる。

例えば本体テキストが1GBありインデックスがその300％のサイズが予想される場合，実データと同等の作業容量を考慮すれば10GB近いディスクスペースが必要となる。著名な全文検索エンジンの場合，インデックスサイズは対象テキストの約70％～300％程度になることが多い。

②元データ管理

全文検索対象として登録されるデータは，いつかは更新される事が予想される。また全文検索の結果表示／閲覧した後にその元々のドキュメントファイルを入手・流用・配布したい場合がある。従って実用システムの構築時にそれらの要望に対処する為には，登録時に一方向的なテキスト抽出を行うだけでなく，抽出テキストの元となるデータファイルの管理機能についてもあわせて検討する必要がある。

③大量件数データの対応

全文検索システムに対し，膨大な登録データ件数を要求される事が増えている。数百万件となる事も珍しくなく，一千万件オーダーの全文検索システムを構築されている事例もある。インデックス作成の方式によっては文字数よりデータ件数にシステムの限界がある場合も多く，著名な検索エンジンでも百万件を超えるデータ量の場合は，その検索スピードを含めて対応可能か検証する事が必要である。

④データマネージメント機能

一般にデータベースというとリレーショナルデータベース(RDB)を指す事が多いが，Oracleなど著名なデータベース構築ソフトウェアはリレーショナルデータベースマネージメントシステム(RDBMS)と呼ばれるのが正しい。それは単に連携データの表計算／検索をするだけでなく，多人数が利用する際の排他制御や利用権限などのマネージメント機能が重要だからである。

当然の事ながら全文検索システム構築の際にも，同様の観点が必要である。検索機能を論ずるだけでなく，誰がどういったデータを登録し，どういうタイミングで誰が検索要求を出すのか，などを必要に応じて検討すべきである。またシステム管理者のためのユーティリティ(システムログ出力など)機能なども，実運用では必要である。

⑶データ検索／表示

①検索スピード

昨今では検索インデックスロジックの工夫とコンピュータ性能の高度化により，数十億文字／秒の単位で検索が可能な検索エンジンがある。ただし多くの場合それは純粋に検索するだけ(見つけるだけ)のスピードであり，実際には大量に検索結果が生じた時は，リストアップ処理に相当の時間がかかる場合がある。インターネットの多くの全文検索サイトではこのリストアップ処理を簡便化(上位数十件だけの表示だけの処理にとどめる)することにより高速化を実現しているが，それにより高度な絞込み検索や一括リスト出力などには対応できなくなっている。従って，システム化する時には目的と規模に応じた考慮が必要である。

②検索辞書機能

“アメリカ”という言葉の同義語として“米国”，

“USA”, “合衆国”などがあげられる。こういった関連語を登録した辞書データをシソーラス辞書というが，その機能を全文検索システムに付加して辞書機能をアピールしている製品もある。昨今では検索条件に文書を入力しその文書自体に対し形態素解析とシソーラス展開する製品や，言葉の意味体系を統計解析して情報検索する製品まで出てきている。これらは全文検索技術と多様な辞書機能を組み合わせた製品である。

シソーラス辞書はJICST(科学技術振興事業団)が一般辞書と各業界用辞書の販売を行っているほか，全文検索システムのメーカーがシステムのオプションとして販売している。しかしシソーラス辞書の精度は単純に登録言語数が豊富だからといって高機能とは限らない。利用者が何らかの専門分野のプロフェッショナルであればあるほど，業界毎の辞書をさらに調整(チューニング)する必要がある。従って汎用のシソーラス辞書を使って拡大解釈できるキーワードで検索した場合，とても絞り込みができないほどの検索結果数が検出される事が予想される。

③あいまい検索

一般にあいまい検索を２種類の意味で紹介されている。

一つは“シソーラス”などの辞書を組込んだ検索エンジンでのあいまい検索を指し，本来ならば“辞書機能”と言い換えても差支え無いものである。

もう一つは，主にパターンマッチング系の検索エンジンでマッチング度を調整する事で類似語句を探し出す機能である。検索結果数が少ない時や英文における多少のスペルミスをひらう時なども，語句の一致度合いを調整する事により可能である。しかしむやみに一致度合いを下げると，検索誤差(ノイズ)となってしまうので注意が必要である。

④絞り込み検索(and/or/sub)

全文検索処理の前後に集合演算機能を持たせる事で実現される。

追記方式で検索キーワードの“積(and)”のみを条件とできるものから，“積(and)”, “和(or)”, “差(sub)”の自由な演算を行うものまである。

但し集合演算のためには検索リストアップの内部処理が必要であり，膨大なデータを対象とした全文検索においてはそのレスポンスに注意が必要である。

⑤全文検索技術と異なる情報検索手法との併用

また目的によっては“日付”, “著者”, “情報分類”などのドキュメント付加情報(ドキュメント属性)も検索条件にしたいことがある。ドキュメント内に必ずしも記載の無い属性情報を検索条件とするには別途登録作業も必要であるが，組み合わせることで効率良く目的のドキュメントを探し出す事ができる。

こういった複合的な検索のイメージを図３に示す。

⑥階層化検索

『Yahoo！』は全文検索サイトとして著名であるが，一方で分類された階層化検索を同時に実現している。それは探したい情報によっては目次を

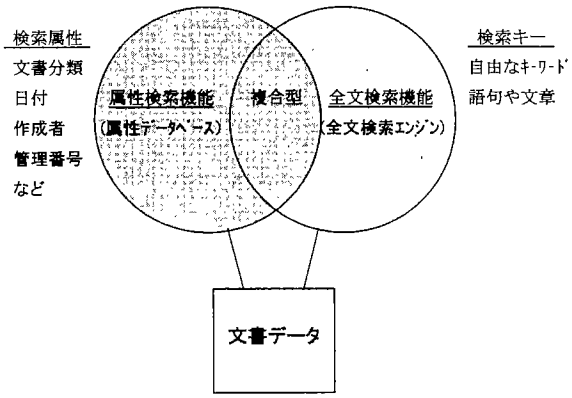



図3

たどるような階層化検索の方が効率的な場合もあるからであり，この機能もできれば全文検索機能と同居させたい機能である。

⑦検索結果表示／出力

検索結果として得られるものは先ず検索結果リスト(一覧表)である。絞り込み途中の検索結果リストでは簡略表示されている場合が多いが，そのリスト自体が重要な意味を持つことも多くある。リスト項目には以下の様なものが想定される。

＊表題
＊属性情報
＊先頭文書の書出し
＊ファイル名(TEXT抽出された元ファイル名)

これらの表示は上位数十件をページング／スクロールするものが多いが，もっとグラフィカルに簡易画像(サムネイル)で表現し，本をめくるような操作性を実現した製品もある。(ただし，サムネイル画像の送信がネットワークに負荷を掛けない事の検証は状況により必要である。)

また結果表示順を何らかのロジックで順位付けして表示させたり，必要な一覧表データを指定フォーマットでファイル出力(又はダウンロード)できるような製品もある。

⑷システム構築における検討課題

①更新／検索／利用権限(セキュリティ管理)

利用者権限を管理する必要のある文書には，以下のようなものが考えられる。

＊検索・閲覧利用が制限される文書データ
特定者以外の閲覧が許可されない機密文書や個人所有文書などが想定される。

＊登録・更新の制限される文書データ
論文や各種マニュアル，顧客情報文書など，検索閲覧や流用も一般に可能であるが利用者に勝手に更新されては困る文書データが想定される。また日常の一般文書を部署内で共有し検索管理する場合なども，文書作成者以外の編集・更新を禁止する必要がある。

利用権限の設定例を表1に示す。

表1　利用権限の設定例

| | 検索・閲覧 | データ流用 | 登　録 | 更　新 | データ削除 | 権限設定 |
|---|---|---|---|---|---|---|
| 一般社員 | ○ | × | × | × | × | × |
| 営業系社員 | ○ | ○ | × | × | × | × |
| 技術系社員 | ○ | ○ | ○ | × | × | × |
| 管理職社員 | ○ | ○ | ○ | ○ | × | × |
| システム職員 | ○ | ○ | ○ | ○ | ○ | × |
| システム管理者 | ○ | ○ | ○ | ○ | ○ | ○ |

○：利用可能　×：利用不可

②多言語対応

全文検索のロジック(検索インデックス手法)は,時としてその対象言語を限定する。「て,に,を,は」を外して分かち書きするなど日本語でしか通用しないインデックス作成手法であれば当然登録対象は日本語のみとなり,欧米語の空白(スペース)を頼りに単語切り出ししたインデックスでは日本語の全文検索では支障がでる。一方,言語依存せずにインデックスを作成する検索エンジンは理論上は多言語検索が可能であるが,実際には言語体系間のコンピュータコード問題や,各言語独自の処理などもあり,一概に対応できるとは考えられない。

ただ日本語と英語の混在文書が日常的な昨今では,日英語のバイリンガル検索に十分対処する必要がある。海外の全文検索エンジンには半角カタカナを考慮していないものも多いため特に注意が必要である。(例として,インターネット上の全文検索サイトで半角カタカナ文字を使用して検索すると,半分以上は文字化けしたり検索ヒット数に誤差が生じたりする。)

③文字コード

日本語Windowsパソコンで通常使用するコンピュータ文字コードはS-JISコードであり,多くのUNIXコンピュータではEUCコードなどが使用されている。しかし全文検索処理を実行させるコンピュータ内では,対象文書の文字コード体系を検索エンジンがサポートするコードに統一しておく必要がある。もしこの処理が徹底されていなければ,検索処理は“文字化け”状態で処理される事になり,当然望む検索結果は得られない。こういった現象は,C/S環境,WEB環境に関わらず事情は変わらず,注意が必要である。

図4のように,検索クライアントがWindows,MAC,UNIX(EUC),登録端末がWindows,全文検索処理を行うサーバがUNIX(EUC)というプラットフォーム環境を想定してみる。

これらの端末環境では,そのOS(オペレーションシステム)事情や使用する言語体系事情及び端末表示事情などにより,利用できるコンピュータ

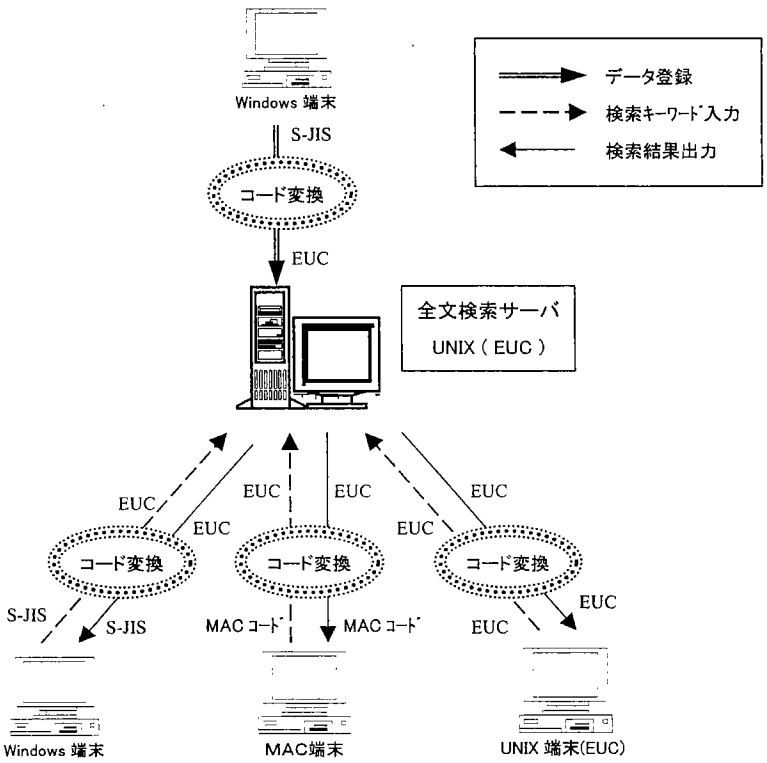



図4

文字コードが制限される。これはマルチプラットフォーム対応のWWWブラウザを利用した場合も同様である。検索キーワードのキーボード入力コード，全文検索処理の統一コード，さらには各端末での結果表示を実現するためのコードについて何らかのストーリを作成しての対処が必要である。

またこれらのプラットフォーム依存のコード体系を統一し，多言語をサポートするコードとしてUnicodeが考えられている。最近のWindowsパソコンやMachintoshは内部コードとして，またJavaやXMLでも標準コードとしてUnicodeが採用されているが，現状普及しているアプリケーションの対応はほとんどされていない。

Unicodeの抱える問題として，以下のような事が考えられる。

＊全世界の言語(特に漢字圏)を統一するには，まだ表現できるコード数が足りない。(UTF-8/UTF-16)

＊日本語と中国語の場合など，言語間で漢字の異体字文字を区別できない。

＊対応する表示フォントが用意されていない。

＊既存ソフトウェアの大半は改造する必要がある。

＊デファクトスタンダードとなるまでに，今後も規格が変わるおそれがある。

④マルチサーバ

利用者が格納場所を意識せずに複数の全文検索サーバを一元検索(串刺し検索)するには分散サーバ(マルチサーバ)機能が必要である。全文検索システム製品の中にはこういった機能を備えたものもあり，離れた部門や遠隔地の支店または別会社で運用されるサーバの一元検索を実現している。ただしUNIXサーバとNTサーバが混在する時は，前に述べたとおり文字コード体系の統一処理も必須である。システムのイメージ例を図5に示す。

## 3. 全文検索関連製品の市場動向

当初全文検索製品は，“整理分類できていない電子情報から自由なキーワードで検索する技術”をセールスポイントとしていた。しかし今日の全文検索技術は，様々なデータ事情・インフラ事情・利用目的などを吸収して変化して来ている。またインターネットの急速な普及をバックボーンに，テキストデータとして集まってくる情報量はより急増している。今後もこの技術が活躍する場は，その利用の指向とその形態を様々に変えてますます広がる事は確実である。

現在の全文検索関連の各製品において，著者が観察できる主な技術指向を図6に示す。

以下，全文検索製品が技術指向する目的と概要について述べる。

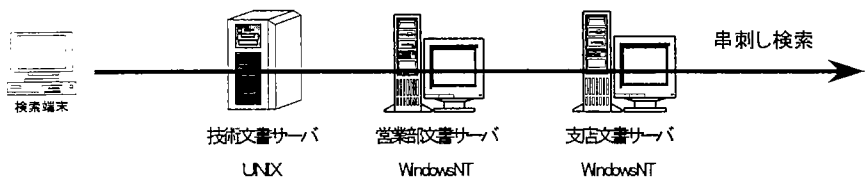



図5

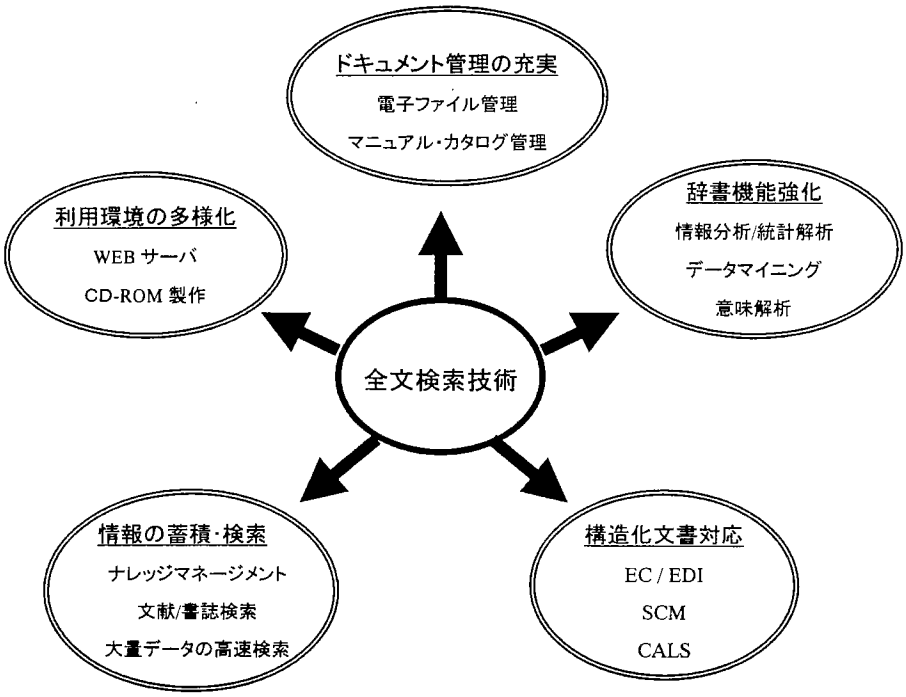



図6

(1)情報の蓄積・検索

研究支援，調査支援，営業支援などの目的で文献や資料を電子的に蓄積し，全文検索で必要な情報を検索するといったシステムが考えられる。現在もっとも一般に利用されている全文検索の利用形態である。全文検索技術を取り入れた事で多彩な検索利用を考慮せずに情報を蓄積できるため，情報を登録する側の負担も軽減している事も一つの特徴である。企業毎の知識情報管理(ナレッジマネージメント)として応用する事も考えられる。

文献の宝庫である図書館においても同様である。現状では図書の全文検索は無理でも書誌の全文検索を可能とするシステムが徐々に構築されている。今後はインターネット／イントラネット環境を利用した電子図書館が国内外で計画されている事もあり，書誌データだけでなく内容に踏み込んだ全文検索の実現も考えられる。しかし多言語文献の電子化を考える場合にはバイリンガル／マルチリンガルの全文検索も必要となり，今後の課題である。

(2)電子ファイル管理

各種電子マニュアルやISO9000ドキュメントなど，電子的に管理したいドキュメントは多種多様に存在する。しかしそれらがMS-Word，PDF，画像文書と統一できない場合がある。そこでこういった場合にそれぞれのファイルに含まれる文字情報を抽出し，全文検索での検索を実現すれば，一元的なファイル情報管理が実現できる。これにより操作上も文字情報に対する全文検索で統一され，必要であればでオリジナルデータ(バイナリデータ)を入手できるようなシステムも考えられる。データの修正時も一連の処理を同様に施せ

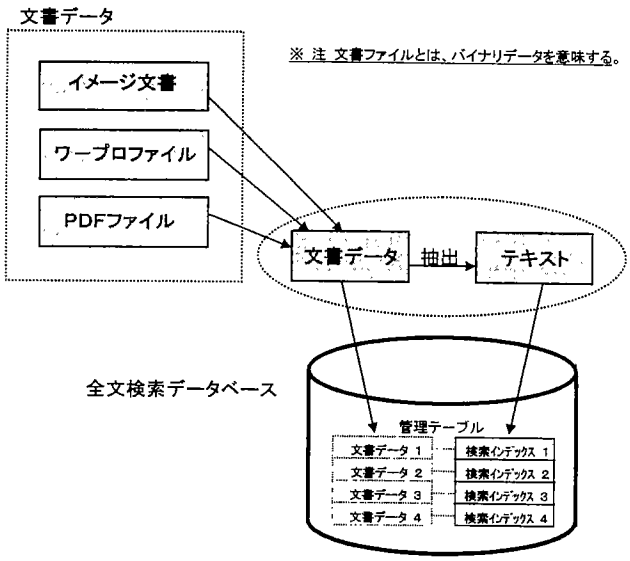



図7

が絞り込めない事が容易に想像できるからである。しかし新しいロジックでもあるため，一定の使用によって類推傾向を学習するような機能を有しているならば，今後も注目したい技術である。

また意味ベクトルから解析する辞書機能を，データマイニング／情報分析／統計解析に応用したシステムも製品化されている。例えば，メール，WWWページ，検索条件キーワードなどのテキスト情報を蓄積し，それらに含まれている言葉の意味統計を行うという“逆転の発想”である。EC/EDIをはじめインターネットで飛び交うランダムな文字情報分析など，マーケティング的な応用が考えられる。

ば，バイナリデータと検索データの同期をとることも可能である。登録のイメージを図7に示す。

⑶辞書機能強化

全文検索に関連する辞書機能として，以下の2種類が考えられる。

①シソーラス辞書(関連語句の登録データ)

②語句の意味統計を解読し，関連語を類推する辞書(意味ベクトルによる解析)

①のシソーラス辞書データは全文検索の利用目的により調整が必要な事は以前述べたが，その調整品質をアピールする製品も多い。

また②は近年製品化された新しい概念である。製品例として，『VextSearch』(コマツソフト)や『ConceptBASE』(ジャストシステム)などがあり，話題を呼んでいる。ただ全文検索の関連語類推には，シソーラス辞書がそうであるように“汎用”はありえない。検索語から汎用的に類推された検索条件ではその和集合が大きくなりすぎ，検索結果

⑷構造化文書対応

構造化文書とはドキュメント構造が定義できるデータ構造を持つ。具体的には昨今注目を浴びているSGML/XMLデータを意味する事が多く，特に今後普及しそうなXMLには注目したい。

XMLデータは，タグ付きのテキストデータで表現されており，意味構造が定義できる事からデータ構造がいかなるルールに則って記述されたかが解読可能である。またXMLは文書だけでなく，CAD図面や地図さらには音声・画像に至るまで表現可能なマルチメディア対応規格である。これらの事からXMLで記述されたドキュメントはその構成要素毎にコンピュータ処理が可能であり，EC(エレクトリックコマース)やSCM(サプライチェーンマネージメント)の電子帳票などでの採用が期待されている。

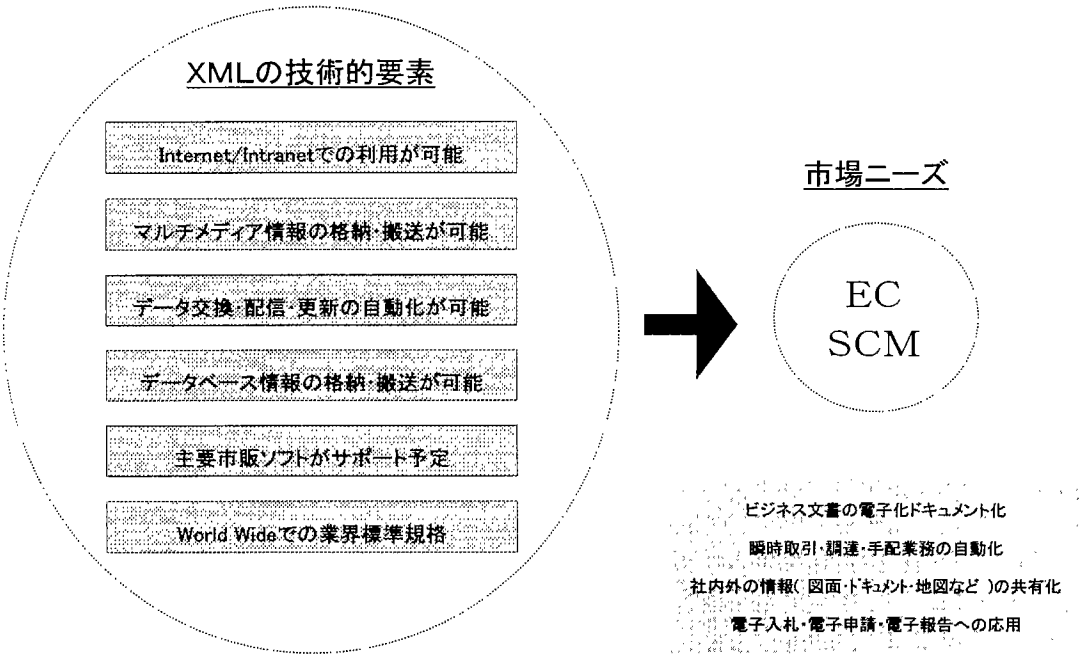



図8

XMLの特徴を図8に示す。

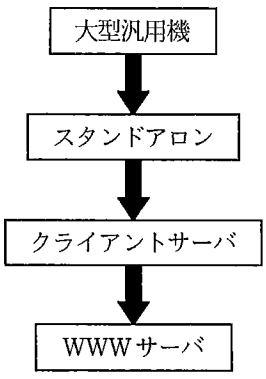
ここで注目すべきは，XMLのドキュメント要素(ELEMENT)は一般に不定長データである点である。固定長の数値情報などであればRDBでも高速検索は可能である。しかし大量の不定長データに対する検索においては，全文検索エンジンはRDBに真似のできない超高速検索を実現できる。

またXMLのタグ構造で表現されているドキュメント要素毎に全文検索処理ができれば，“表題”とか“著者名”，“企業名”といった要素属性に応じた検索も可能である。これによりドキュメントの意味合いをさらに考慮しての全文検索を行い，目的のドキュメントをより効率良く探し出せる事になる。

著者はXML関連システム構築において，RDB/OODBなどの利用以外にこういった全文検索技術を取り入れたサーバの構築手法が今後増えると考えている。現在ではドキュメントのタグ要素に直接全文検索できる製品として『OpenTEXT』が上げられる。しかしDTD(文書構造定義)に柔軟に対応できるシステムを構築するにはまだ課題は多い。XMLの特性を活かせる全文検索エンジンが今後開発される事を期待したい。

⑸利用環境の多様化

全文検索の利用環境は，右のように変遷してきた。




しかし近年は印刷物のCD-ROM化などが推進され，コンテンツを電子化するに伴い全文検索エンジンを組込

む事例も増えている。対象として以下のようなものが考えられる。

①総合カタログ

②製品カタログ

③各種マニュアル

④Q＆A集

⑤各種辞典・学習書

⑥特許情報・法令事例集

⑦論文集・学術専門書・報告書

⑧新聞・雑誌記事

これらのデータはCD-ROMなどの電子媒体で配布されるが，ユーザーは必要により全文検索の機能を享受できる事になる。WWWサーバとの併用を含め，このようにデータメディアに検索エンジンも組込む事例が今後増えるのではないだろうか？

## 参考文献

『日経バイト/No.156』(1996.10月号) p.142～p.167　日経BP社刊

『マルチリンガルWEBガイド』三上吉彦，関根謙司，小原信利　共著　1997.8　㈱オライリー・ジャパン刊