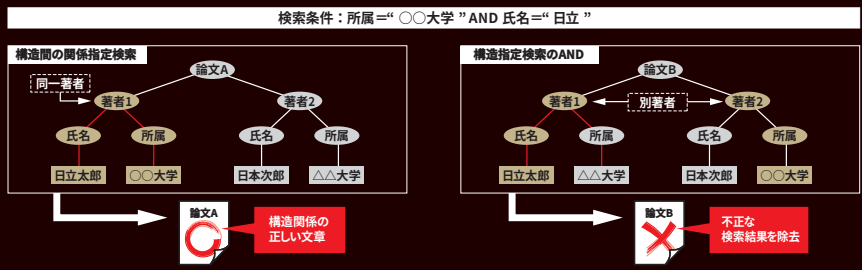
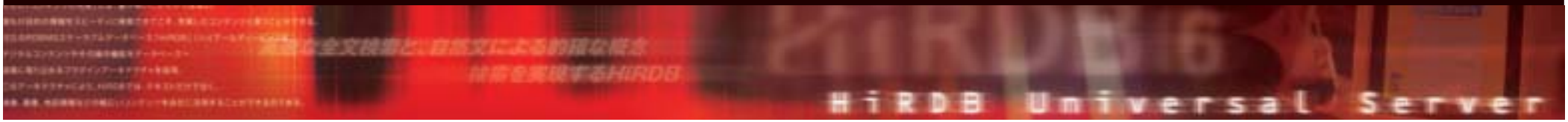
第❸回

# 高速な全文検索と、自然文による的確な概念　検索を実現するHiRDB

インターネットビジネスでは、コンテンツの充実がビジネスの成否を左右する。
ただし「コンテンツの充実」とは、量が多いことだけではない。
誰もが目的の情報をスピーディに検索できてこそ、充実したコンテンツと言うことができる。
日立のRDBMSスケーラブルデータベース「HiRDB」（ハイアールディービー）は、
デジタルコンテンツやその操作機能をデータベースへ
容易に取り込めるプラグインアーキテクチャを採用。
このアーキテクチャにより、HiRDBでは、テキストだけでなく、
映像、画像、地図情報などの幅広いコンテンツを自在に活用することができるのである。

■ XMLの構造を意識した検索




## 高速な検索が求められるコンテンツ管理

インターネットビジネスで成功を獲得するには、コンテンツを充実させるだけではなく、豊富なコンテンツをスピーディに検索し、目的に合った情報を短時間で入手できる仕組みを提供することが大切である。

特に一般ユーザーを対象にしたWebサイトを運営している場合、検索のしやすさを向上させると、確実にアクセス数の増加につながる。検索対象の主役は何と言ってもテキストである。手軽に文書が検索でき、目的にかなった情報を短時間で手に入れられると、ユーザーはそのWebサイトに満足感をおぼえ、運営している企業に対するイメージも向上する。

企業内で構築されているナレッジマネジメントシステムや企業ポータル（EIP）も同様である。

企業内に蓄積されている膨大な情報を一元的に管理し、再利用しやすくする仕組みとして企業ポータルが注目されており、ナレッジマネジメントシステムのユーザーインターフェースとしても企業ポータルを構築する企業が増えている。しかし単に、多種多様な既存情報への入り口を1画面にまとめて表示するだけでは、業務効率を向上させることはできない。複数のシステムにまたがって情報を横断的に検索し、スピーディに加工できる環境を整えてこそ、企業ポータルは威力を発揮する。

最新のCTI（Computer Telephony Integration）技術を駆使したコールセンターでも、オペレータが短時間で的確な応答をするためには、蓄積されたFAQデータベースをスピーディに検索する必要がある。百人百様の表現を受け止めて、通話している間に問題点を絞り込むことが要求されているオペレータにとって、製品名や顧客の電話番号だけでしか検索できないシステムでは役に立たない。

## インターネットビジネスではXML対応が必須

XMLは、Webサービスをはじめ、B to Bの企業間取引で多く用いられるようになり、電子政府でも主役となる言語である。また、新聞記事データや特許データなどはXML文書として、提供されるようになってきており、企業ポータルやナレッジマネジメントシステムに組み込む企業も増えている。

XMLが幅広く利用されるようになったポイントは、データを意味付けすることにより、構造化された要素を個別に取り扱うことができるからだ。特定商品の請求金額だけ取り出して集計するといったことが容易にできる。指定した項目をマークアップして次の作業に引き渡すことも容易にできるため、業務ワークフローも構築しやすい。

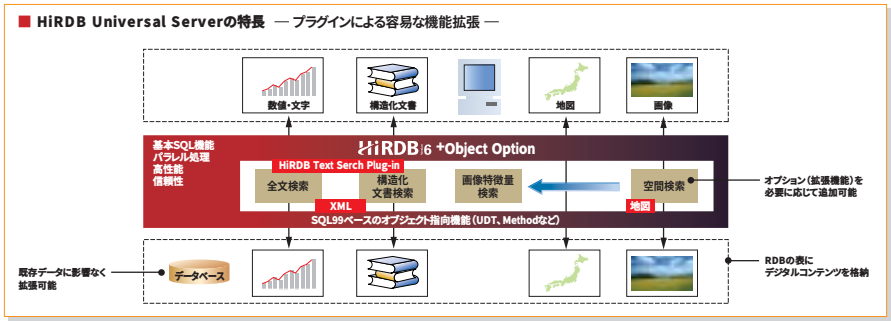
しかし、表形式で構成されるRDB（Relational DataBase）にXMLデータをそのまま格納すると、構造化言語としての機能を発揮させることができない。構造間の関係を指定した検索ができないため、例えば「氏名の項目が“ 日立太郎 ”」でかつ「所属の項目が“ ○○大学 ”」である人を探すといった作業で、誤った検索結果を返してしまう。それでは、構造情報も一緒に格納すればいいかというと、構造が複雑になるにつれてインデクス容量が膨大になってしまうため、実用的ではない。

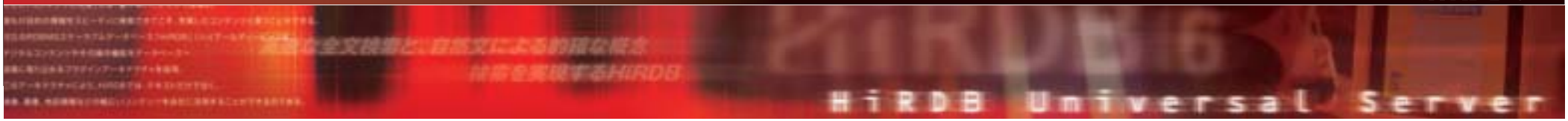
## デジタルコンテンツ管理を実現するHiRDB Universal Server

こうした問題を解決できるのが、日立のスケーラブルデータベース「HiRDB」である。HiRDBは、HiRDB Object Optionを追加することで、RDBからORDB（オブジェクト指向リレーショナル・データベース）であるHiRDB Universal Serverへと進化する。これにより、デジタルコンテンツの操作機能を提供する「プラグイン」を使用することができるようになる。

例えば、テキスト情報を検索するためのHiRDB Text Search Plug-inを装着すれば、「インクリメンタルn-gramインデクス方式」による高速全文検索機能や、自然文を元に

■ HiRDB Universal Serverの特長 ― プラグインによる容易な機能拡張 ―






検索ができる概念検索機能を利用することができる。

さらに、XMLのデータ構造を維持したままで格納しているため、XML構造の関係をきめ細かく指定した検索を実現している。「氏名の項目が“日立太郎”」で「所属の項目が“○○大学”」である人を探すといった構造指定検索を、安定かつ高速に処理できるのである。しかもインデクス容量が構造指定の階層の深さに依存しないため、データベース容量を無駄に増大させる心配がないのも大きなメリットだ。そのうえ、検索結果の得点付けやランキングの機能も利用することができる。

## RDB機能との連携が容易

HiRDB上にHiRDB Text Search Plug-inが装着されたことにより、RDBにより提供される機能と文書検索エンジンにより提供される機能とが密接に連携しているのも魅力である。

例えば、HiRDB Text Search Plug-inにより検索された文書群を、文書の作成日時で並び替えて表示したり、文書を作成した部署ごとにグルーピングして表示することができるのは、HiRDBに装着されているからこそ容易に実現される機能である。

また、種々の検索機能を組み合わせた統合検索も容易に実現される。例えば、XML/SGML文書に対する構造指定検索と概念検索、あるいは、文書作成日時などの属性検索と概念検索を組み合わせることもできる。

## 高速な全文検索を実現するインクリメンタルn-gramインデクス方式

n-gramインデクス方式は、辞書メンテナンスにコストをかけることなく、漏れのない検索を実現する全文検索方式のひとつである。

従来の日本語検索エンジンは、形態素解析という文法的な分析処理を行って文章を単語に区切り、区切った単語を単位として検索するものが多かった。この方法は、単語の情報が格納された辞書を使うため、辞書に入っていない単語は容易には検索できない。たとえば「企業ポータル」や「Webサービス」など、新しい単語が登場するたびに、辞書をメンテナンスしなければならない。

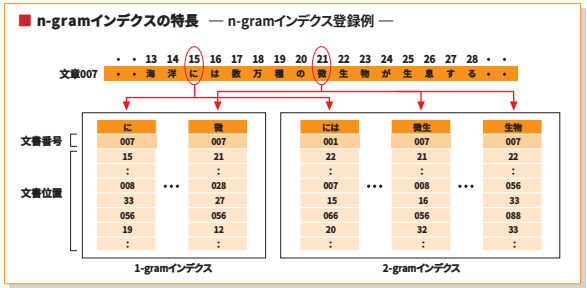
これに対してn-gram方式では、辞書を使わずに、文書中に存在するn文字の文字列そのものをインデクスにするため、辞書のメンテナンスが不要だ。

たとえば、「海洋には数万種の微生物が生息する」という文章を1-gramインデクスで登録すると、16個の文字に対して、1文字ずつ文書番号と文書内での位置情報が登録される。2-gramインデクスで登録した場合には、「海洋」「洋に」「には」のように、隣り合う2文字ずつのそれぞれについて、文書番号と文書位置を登録する。「微生物」という文字検索を行う場合には、「微生」と「生物」というインデクスが隣り合っているものを探すため、事前に「微生物」という単語を辞書に登録しておく必要がない。

一般的にn-gram方式では、インデクス付けする文字列の長さが長くなるほど、検索スピードは速くなるが、最初に文書を登録するときに時間がかかるため、用途と利用環境に応じて設定を調整する必要がある。しかし、HiRDB Text Search Plug-inでは、検索スピードが遅い文字列に対してだけ、インデクスの長さを拡張するインクリメンタルn-gram方式を採用しているため、煩わしい設定は不要だ。

また、通常のシステムでの検索では「インタネット」と「インターネット」のように異表記関係にある単語群の、表記の揺らぎを吸収した検索は苦手だ。これらのカタカナや英字は新語が多い上、文字の組み合せのバリエーションを生みやすいため、同義語辞書による展開が適用しにくいためだ。HiRDB Text Search Plug-inは、同義語辞書による展開に加えて、異表記展開ルールにより異表記語を展開する。このため、異表記語群を漏れなく検索することができる。さらに、異表記展開、同義語展開、異表記展開と3段階の展開を行うため、同義語辞書の見出し語における表記の揺らぎも吸収することができる。

また、「“最新”と“技術”の文字の間に2文字入った文字列」といった近傍条件検索や、「“政治”と“経済”を含む文書を検索したいが、“政治”よりも“経済”を重視して検索したい」といった重みづけを指定しての検索など、きめ細かい検索条件を指定することができるのである。




## 概念検索までサポートするデータベースHiRDB

たとえば「近年、環境保護に力を入れている自治体が増えている」という意味の文書を検索するというように、自然文で指定した概念に沿って、類似した概念を持つ文章を検索するのが概念検索である。

HiRDB Text Search Plug-inにおける概念検索では、類似文書を判定する際に単語辞書を使用しないため、辞書のメンテナンスは不要である。しかも、適切なキーワードが思いつかない場合でも関係深いと思われる文書が検索できるため、ユーザーがホームページを使ってセルフサービスで問題解決をする場合や、コールセンターでのFAQデータベース検索などに適している。また、ナレッジデータベースから、新しいビジネスのヒントやアイデアを得る場合にも有効だ。

さらに、「近年、環境保護に力を入れている自治体が増えている」という例示文書から、「環境保護」「自治体」というポイントとなるキーワードを自動的に抽出する機能もある。

また、概念検索の結果に対して、検索者が適不適の評価をすることで、検索精度を高めていく学習機能もある。この機能を使えば、検索結果を見ながら欲しい情報を集めていくことができる。

HiRDB Text Search Plug-inでは、全文検索と概念検索を一本化したアーキテクチャで実現している。全文検索用および概念検索用といった形で検索用インデクスを個別に作成する必要がないため、運用が極めて簡単である。また、全文検索結果として得られた文書群を、所定の概念で並べ替えて参照したり、概念検索結果として得られた文書群を全文検索条件で絞り込んだりすることもできる。

## 新しいデジタルコンテンツ管理ソリューション

HiRDBは、優れた日本語全文検索機能を持つデータベースである。インクリメンタルn-gram方式で高速な検索ができるうえ、概念検索など誰でもわかりやすい方法も提供している。XML/SGMLにも対応し、複合検索で絞り込みを行う機能まで備えている。

また、HiRDB Text Search Plug-inだけでなく、映像、画像、地図情報などを扱うプラグインが豊富に用意されており、幅広いマルチメディアコンテンツを自在に活用することができる。プラグインはHiRDBの内部ルーチンとして組み込むため、デジタルコンテンツを高速に処理できるのもHiRDB Universal Serverの特長である。

もうひとつ忘れてならないのは、HiRDBは、ミッションクリティカルな業務にも対応できるように開発された信頼性の高いデータベースであり、障害発生時のリカバリや、サービスを停止せずにメンテナンスしたりバックアップする機能などが網羅されていることだ。また、シェアードナッシング型アーキテクチャを採用した並列データベースであるため、サーバ台数にほぼ比例してスケーラブルな処理性能を発揮でき、データ量の増加やアクセス負荷の増加に応じて柔軟に拡張することができる。

HiRDB Universal Serverを活用すれば、インターネットを通じて24時間利用されるシステムに、コンテンツの柔軟な利用環境を提供することができる。インターネットビジネスを成功に導く新しいデジタルコンテンツ管理を実現するのが、日立のスケーラブルデータベース「HiRDB」なのである。



お問い合わせ
株式会社 日立製作所
ソフトウェア事業部 販売推進部
〒140-8573 東京都品川区南大井6-26-2 大森ベルポートB館
TEL.03-5471-2592 FAX.03-5471-2395
www.hitachi.co.jp/soft/hirdb/
e-mail:hirdb@itg.hitachi.co.jp