5P－10

知識指向文書管理基盤の開発（4）
ORDB向け構造化文書全文検索プラグイン

後地陽介† 川下靖司† 山本伸也† 多田勝己‡ 河村信男‡
(株)日立製作所 †ソフトウェア事業部 ‡システム開発本部

1. はじめに

弊社では、SGML などの構造化文書の構造を指定した全文検索、検索でヒットした文書に得点付けを行うスコアリング機能およびスコア順にデータ表示をするランキング機能を実現するため、インクリメンタル n-gram インデクス方式[1]を採用した全文検索システム Bibliotheca2 TextSearch を開発した[2]。

しかし、近年のインターネット、イントラネットの急速な普及と、特許、新聞記事といったデータ量の大規模化に伴い、全文検索システムの運用性や信頼性が問題になってきた。特にデータ規模が大きくなると、DBMS と同様なデータ登録やバックアップといった運用性の容易化や信頼性の確保などが全文検索システムにも求められるようになった。

一方、DBMS においては、多様なメディアのデータを活用してサービスを提供することが不可欠になってきており、文書に関しては全文検索、並びに、構造化文書のデータベースとの統合が、データの一元管理、検索機能の統一性、電子化文書登録等において、特に効果的として実現が求められている。

これらの要求に応えるため、柔軟なスケーラビリティと高い性能・信頼性を備え、ORDB（Object Relational Database）として拡張可能である日立スケーラブルデータベースサーバ HiRDB Version 5.0（以降、HiRDB と略す）のプラグインアーキテクチャ（図1）に対応した全文検索プラグイン HiRDB Text Search Plug-in を開発した。なお、ここで言うプラグインとは、DBMS 本体のプログラムを変更せずに、後から処理モジュールを追加して、DBMS の機能を拡張できる機構のことである。

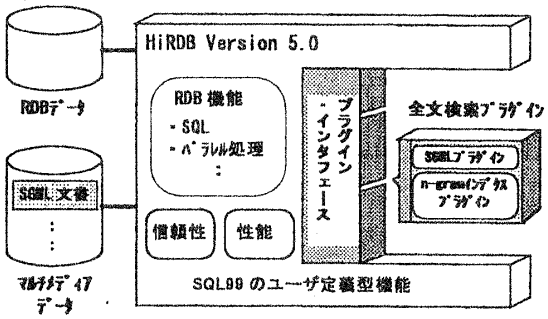

図1 ORDBとプラグインアーキテクチャ

2. システムの概要

2.1. システムの構成

シングル構成からパラレル構成まで幅広く対応し、かつ、大規模並列処理に強い Shared Nothing 方式を採用した HiRDB に、全文検索機能のプラグインである HiRDB Text Search Plug-in（以降、Text Search Plug-in と略す）を組込むことにより、HiRDB に全文検索の機能を拡張することができる。

Text Search Plug-in は、図1に示すように、SGML プラグインと n-gram インデクスプラグインで構成する。SGML プラグインはデータ型を追加するプラグインで構造化文書の検索、登録等の操作機能を実現するための抽象データ型を追加する。また、n-gram インデクスプラグインは、データ型に対する検索インデクスを追加するプラグインで、インクリメンタル n-gram インデクス方式による検索手段を提供する。

2.2. 開発方針

Text Search Plug-in を開発するにあたり、全文検索システム Bibliotheca2 TextSearch で実現した検索機能を継承し、かつ、HiRDB に組込むことにより、以下の項目を目標とした。

Full Text Search Plug-in for Large Strctured Document on ORDB
Yousuke USHIROJI†, Yasushi KAWASHIMO†, Shinya YAMAMOTO†, Katsumi TADA‡, Nobuo KAWAMURA‡
†Software Division, Hitachi, Ltd. ‡Business & Information Systems Development Division, Hitachi, Ltd.

①1,000万件レベルの構造化文書の高速な全文検索
②ORDBにおける属性検索と全文検索を融合した複合検索
③運用性の容易化と信頼性の確保
④複雑な検索条件式を作成せずに所望の文書を検索できる機能の実現

3. Text Search Plug-inの開発

3.1. 大量文書の高速ノイズレスな構造全文検索

Text Search Plug-inでは、大量文書を対象に、高速な全文検索と構造化文書の検索を実現するため、Bibliotheca2 TextSearchでのインクリメンタルn-gramインデクス方式を採用した。この方式では、総インデクス容量を抑えながら、出現頻度の高いn-gramを含む検索語が指定された場合でも高速な全文検索が可能である（図2）。

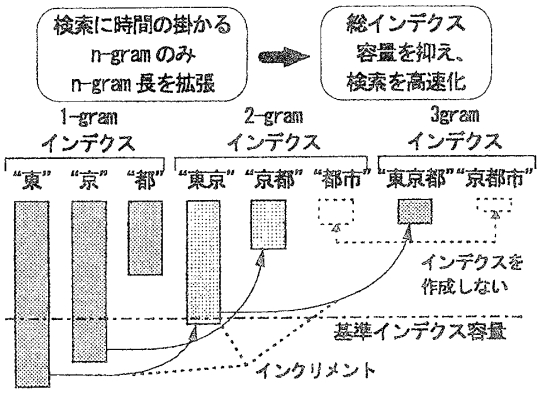

図2 インクリメンタルn-gramインデクス方式

また、HiRDBに登録する文書のn-gramインデクスは、BLOBとして格納することにより、HiRDB管理下でパラレル処理を利用した大量文書の高速ノイズレスな全文検索が可能になった。

3.2. 属性検索と全文検索の複合検索

全文検索機能をプラグインとして組込むことで、全文検索で扱う文書も、多種多様な他のデータと同様にHiRDBで一元管理できるため、SQLを利用して全文検索とその文書の関連情報による属性検索の条件を同時指定するといった複合検索が可能になった。

3.3. 運用性の容易化と信頼性の確保

全文検索機能をプラグインとして組込むことで、全文検索で扱う文書も他のデータと同様にHiRDBで一元管理できるため、バックアップ等の運用が容易になった。また、HiRDBで取得する更新履歴情報を取得することにより、障害時の回復が容易になった。

3.4. 文章の内容で簡易に検索できる概念検索

Bibliotheca2 TextSearchで実現していない全文検索の拡張機能として、概念検索機能[4]を開発した。概念検索では、検索条件に検索語ではなく文章を入力し、それと似通った文書を検索するため、複雑な条件式を作成せずに所望の文書を検索することが可能になった。また、検索条件に入力した文章から抽出した概念情報を検索に利用することにより、検索語で検索するよりも幅広く目的の文書を探し出すことが可能になった。

4. まとめ

今回、ORDB向け構造化文書全文検索プラグインとして、Text Search Plug-inを開発した。柔軟なスケーラビリティと高い性能・信頼性を備え、また、ORDBとして拡張可能なHiRDBに、全文検索機能をプラグインとして提供することで、以下の内容を実現することができた。

①HiRDBの大規模並列処理とインクリメンタルn-gram方式による大量文書の高速な構造全文検索
②全文検索と他データの属性検索との複合検索
③HiRDBの管理下で高信頼性を確保

また、新規機能として

④検索条件に文章を入力して所望の文書を幅広く検索する概念検索

を実現することができた。

参考文献

[1]菅谷他：「n-gram型大規模全文検索方式の開発ーインクリメンタル型n-gramインデクス方式ー」，情報処理学会第53回全国大会 5T-2
[2]川下他：「構造化文書対応全文検索システムBibliotheca2 TextSearchの開発(1)～(4)」，情報処理学会第55回全国大会 4N-3～6
[3]小林他：「ORDBにおけるプラグイン組込み仕様と実行制御」，情報処理学会第58回全国大会 2T-02
[4]松林他：「知識指向文書管理基盤の開発（5）n-gram方式に基づく概念検索」，情報処理学会第59回全国大会 5P-11