構造化文書対応全文検索システム Bibliotheca2 TextSearch の開発(1)* —システムの概要—

4N－3

川下靖司† 坂田淳† 日野隆教† 多田勝己 橋本和広†
(株)日立製作所 ソフトウェア開発本部† 情報・通信開発本部

1．はじめに

近年、インターネット、イントラネットの利用増大に伴い、生成される電子化文書も爆発的な勢いで増加している。こうした状況下で、大規模な文書データベースの中から所望の文書を効率的に検索できる全文検索システムに対する要求が急速に高まっている。

筆者らは、このような要求に応えるシステムとして、階層型プリサーチ方式に基づく全文検索システム Bibliotheca/TS[1]を開発した。また、ここで用いている文字成分表に改良を加え、特許 100 万件のデータベース(テキスト容量：約15GB)を１秒で検索できる技術を開発した。

今回、SGML/HTML にも対応でき、ランキング検索が可能な検索システムとして、全文検索システム Bibliotheca2 TextSearch を開発したので、その概要について報告する。

2．システムの特長

今回開発したシステムの特長と開発技術の概要について述べる。

（1）高速ノイズレス検索の実現

Bibliotheca2 TextSearch では、インクリメンタル n-gram インデクス方式[2]と呼ぶ検索方式を採用している。本方式の概要を図１に示す。

通常の n-gram インデクス方式では、登録の対象とする文書から固定長の部分文字列(n-gram)を抽出し、これに対する文書番号と文字位置をインデキシングする。これに対し、本方式では、システムに期待する検索レスポンス時間を基に基準となるインデクス容量を設定する。そして、この容量を超えた n-gram に対して n-gram 長を拡張したインデクスを既作成のインデクスを参照して生成する。これにより、総インデクス容量を抑えながら出現頻度の高い n-gram を含む検索タームが指定された場合でも高速な全文検索が可能になる。

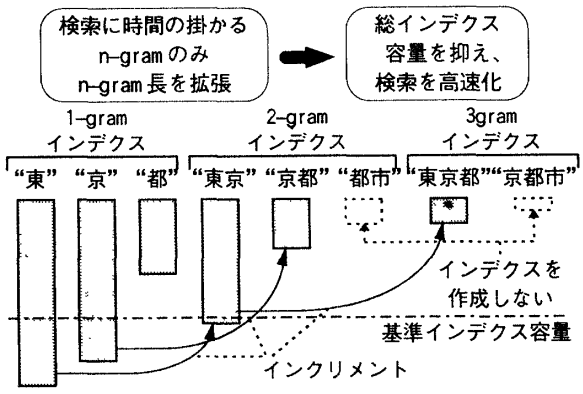

図１ インクリメンタル n-gram インデクス方式

（2）構造化文書対応

構造化文書の標準形式である SGML 形式の文書を登録、検索するための技術を開発した。これにより、これまで別々の情報として管理してきた本文と書誌を、標準形式の枠組で一体管理することが可能になるとともに、他の文書管理システムとのデータ交換性を飛躍的に向上することが可能になった。また、指定した論理構造だけを対象とした文書の検索や表示など、文書の論理構造を利用した木目の細かい検索や表示機能を提供することが可能になった。

（3）文書閲覧の容易化

大規模文書データベースの検索時に検索結果として得られる大量のヒット文書中から、文書閲覧を容易化するために、検索結果文書に対し、検索式に対する適合度に応じて得点付けを行なうスコアリング検索機能と、スコア順に検索結果一覧を作成、表示するランキング検索機能を

---

* Full Text Search System for Large Structured Document Database, Bibliotheca2 TextSearch(1).

† Software Development Center, Hitachi, Ltd.
† Information Systems R&D Division, Hitachi, Ltd.

開発した。また、スコアリング検索機能については、検索ターム毎に重要度を設定する検索ターム重みと、論理構造毎に重要度を設定する論理構造重みなどを設定できるようにした。これにより、AND/OR 条件で接続された複数の検索ターム間で、特定のタームに対して高い得点を付与できるようになった。例えばタイトルの論理構造中に指定された検索タームを含む文書に対して高い得点を付与するなどといった柔軟なスコア算出指定が可能になった。

3．システムの構成

（1）システムの構成

Bibliotheca2 TextSearch は、以下に示す3つのシステムにより構成される。

（a）Bibliotheca2 TextSearch サーバ

文書の登録、検索および表示機能を提供する。本サーバでは検索セッション管理機能を提供し、ユーザの検索履歴や過去の検索結果集合を管理する。このため、過去の検索結果の参照や、過去の検索結果を対象とした絞り込み検索を高速に実現することができる。

（b）Bibliotheca2 TextSearch ライブラリ

サーバへの API (Bibliotheca2 TextSearch API)を提供するために使用する。検索セッションの接続、切断や検索コマンドの送出など，標準で 17 個の API 関数を提供している。

（c）Bibliotheca2 TextSearch ゲートウェイ

インターネット、イントラネット環境下で利用するための専用ゲートウェイであり、HTTPD から CGI(Common Gateway Interface)により起動され、検索結果を HTML 形式に変換し Web ブラウザに返送する機能を持つ。標準インタフェースに対し、簡単に画面をカスタマイズすることができる。

（2）システムの利用形態

システムの利用形態としては、図2に示すように Bibliotheca2 TextSearch ゲートウェイを介して Bibliotheca2 TestSearch サーバに格納された文書を直接検索、参照する利用と、図3に示すようにユーザ開発の文書管理システムから Bibliotheca2 TextSearch API を通して検索結果を取得するバックエンドサーバとしての利用を想定している。また、サーバ内部の登録制御部および検索・表示制御部は関数として切り出し易いよう設計しており、DBMS の検索用関数として利用することも可能である。

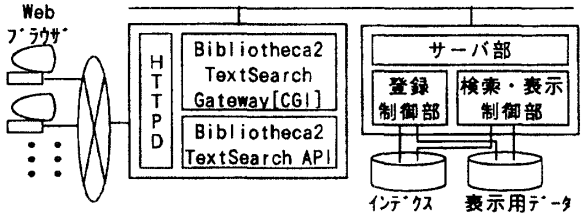

図2 ゲートウェイを介した利用

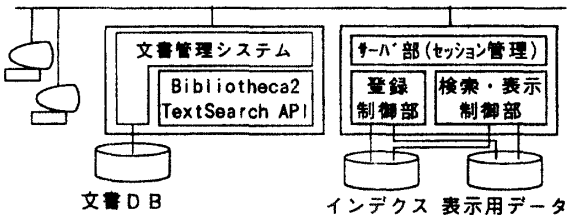

図3 文書管理システムを介した利用

4．まとめ

今回、インターネット、イントラネット対応の大規模全文検索システムとして、以下に示す特長を持つ構造化文書対応全文検索システム Bibliotheca2 TextSearch を開発した。

（1）ノイズレス高速全文検索

インクリメンタル n-gram インデクス方式と呼ぶ全文検索方式により、大規模な文書データベースに対しノイズのない高速検索が可能となった。

（2）構造化文書対応

構造化文書の標準形式(SGML)に基づき本文と書誌の一体管理が容易に行なえるようになった。

（3）スコアリング/ランキング機能のサポート

大規模な文書データベースに対しても効率的な文書の検索、閲覧が行なえるようになった。

5．参考文献

[1]畠山他：「ソフトウェアによるテキストサーチマシンの実現」，情報処理学会情報学基礎研究報告，Vol.92,No.32,25-2,pp.19-25(1992.5)

[2]菅谷他：「n-gram 型大規模全文検索方式の開発 —インクリメンタル型 n-gram インデクス方式—」，情報処理学会第 53 回全国大会 5T-2