構造化文書対応全文検索システム Bibliotheca2 TextSearch の開発(4)* —検索機能および検索性能—

4N－6

菅谷奈津子† 多田勝己† 岡本卓哉† 加藤寛次† 川下靖司‡
(株)日立製作所 情報・通信開発本部† ソフトウェア開発本部‡

1．はじめに

本稿では、今回開発した構造化文書対応全文検索システム Bibliotheca2 TextSearch における検索機能について報告する。

また、新聞記事データを対象として実測した検索用ファイルの容量および検索性能について報告する。

2．検索機能

今回開発した構造化文書対応全文検索システム Bibliotheca2 TextSearch では、表１に示す検索機能が実現されている。

表１ 検索機能の一覧

	検索機能	概　要
1	単純検索	検索タームを含む文書の検索
2	論理条件検索	検索タームによる論理演算
3	構造指定全文検索	論理構造を指定した検索
4	スコアリング検索	検索結果に対する得点付け
5	ランキング検索	スコア順での結果表示
6	近傍条件検索	ターム間の隣接関係を指定
7	同義語・異表記検索	同義語と異表記を含む検索
8	前方・後方一致検索	項目内での先頭／末尾を指定
9	絞り込み検索、検索結果間の論理演算	過去の検索結果を対象とした検索、検索結果間の論理演算

以下、各機能の概要について述べる。

（１）単純検索

登録文書全文中に指定した検索タームが含まれる文書を、漏れなく、ノイズを生じることなく検索することができる。

（２）論理条件検索

AND 検索, OR 検索および NOT 検索が指定できる。

（３）構造指定全文検索

「タイトル中に“SGML”が含まれる文書の検索」というように、検索の対象とする論理構造を指定した検索を行なうことができる。また、指定した論理構造の下位構造に、指定した検索タームが含まれる文書を検索することができる。さらに、章タイトルに“SGML”が含まれ、かつ同じ章の章内容に“全文検索”が含まれる文書というように、検索タームが存在する論理構造間の関係を指定した検索を行うことができる。

（４）スコアリング検索

検索結果文書に対し、検索条件への適合度に応じて得点付け(スコアリング)を行なうことができる。また、複数の検索ターム間で重要度に応じて重みを設定することができる。さらに、タイトルの構造中に検索タームが含まれる文書に特に得点を高くするといった論理構造毎の重みを設定することもできる。

（５）ランキング検索

得点(スコア)順に検索結果文書の一覧を作成、表示することができる。

（６）近傍条件検索

“行政”と“改革”が 10 文字以内に隣接する文書を検索するなど、検索ターム間の隣接関係を指定することにより、各検索タームの意味的な関連が深い文書を検索することができる。

（７）同義語・異表記検索

同義語辞書により、指定した検索タームと同じ意味を持つ語を含む文書を検索することができる。また、カタカナやアルファベットに対しては、ルールベースの異表記展開処理により、辞書を用いることなく表記の揺らぎを吸収した検索を行うことができる。

---

* Full Text Search System for Large Structured Document Database, Bibliotheca2 TextSearch(4).
† Natsuko SUGAYA, Katsumi TADA, Takuya OKAMOTO, Kanji KATO
‡ Yasushi KAWASHIMO
† Information Systems R&D Division, Hitachi, Ltd.
‡ Software Development Center, Hitachi, Ltd.

（8）前方・後方一致検索

氏名やタイトルなどの論理構造に対し、検索タームが構造の先頭や末尾に現われる文書を検索することができる。

（9）絞り込み検索および検索結果間の論理演算

Bibliotheca2 TextSearch の検索セッション管理機能により、過去の検索結果集合を対象とした絞り込み検索を指定することができる。また、検索結果集合間の論理演算も指定できる。

3．検索性能

Bibliotheca2 TextSearch を用いて 50 万件の新聞記事データを対象として全文データベースを作成し、検索用ファイルの容量および検索処理時間を測定した。

（1）測定条件

◎使用機器：

ワークステーション：日立 3050RX/340

◎使用磁気ディスク：

2GB の HDD を HFS として使用

◎測定対象データベース：

日本経済新聞記事 50 万件(1981/10～88/5)

SGML 原文書容量：770MB

◎測定環境：

同一 WS 上でクライアントとサーバプログラムを動作させるスタンドアロン環境で測定

（2）測定項目と測定方法

サーバ内にタイマを組み込むことにより、検索処理に要する時間を測定した。なお、測定に際しては、ファイルシステムのキャッシングの影響を排除するために、検索の度にディスクをアンマウントし再マウントした後に、ダミーの条件式で検索を行い、その直後に検索を実行して処理時間を測定した。

（3）性能測定結果

（a）データ容量

検索用インデクス	：2.26GB
解析済み構造化文書	：845MB
n-gram 情報管理部(トライ)	：39MB
SGML 構造インデクス	：120KB

（b）検索処理時間

表 2 に示す検索条件に対し、検索処理時間を測定した結果を図 1 に示す。

なお、本図に示す測定結果はインクリメント処理前のインデクスを対象としたものである。したがって、インクリメント処理を加えることにより、さらに高速化できるものと考えている。

表 2 測定に用いた検索条件

	検索条件	ヒット件数	ヒット率
1	全文:“行政改革”	3,424 件	0.69 %
2	全文:“千代の富士”	57 件	0.01 %
3	見出し:“行政改革”	56 件	0.01 %
4	全文:“行政改革”[スコア]	3,424 件	0.69 %
5	全文:“行政改革”[ランク]	3,424 件	0.69 %
6	全文:AND(“行政”,“改革”)	5,179 件	1.04 %
7	全文:OR(“行政”,“改革”)	34,559 件	6.94 %
8	全文: “行政”と“改革”が 10 文字以内	3,557 件	0.71 %

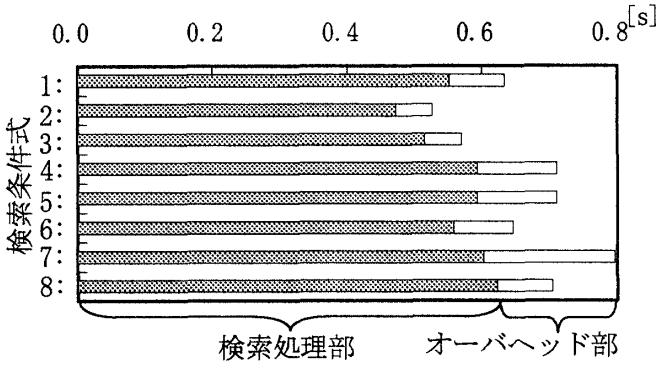

図 1 検索処理時間

4．まとめ

新聞記事 50 万件のデータベースを対象として検索性能を測定した結果、0.6 秒程度で検索が完了する見通しを得た。今後は、さらに大規模な文書データベースを対象とした検索性能を測定するとともに、検索性能の要因を分析し検索性能を高める方式の開発を進める。

最後に、性能測定に用いた新聞記事データを提供頂いた日経新聞社殿に感謝致します。

5．参考文献

[1]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

[2]川口他：「n-gram 型大規模全文検索方式の開発 —文字種適応型 n-gram インデクス方式—」，情報処理学会第 53 回全国大会 5T-3