構造化文書対応全文検索システム Bibliotheca2 TextSearch の開発(2)* —構造化文書処理方式—

4N－4

岡本卓哉† 多田勝己† 菅谷奈津子† 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

構造化文書対応全文検索システム Bibliotheca2 TextSearch における、構造化文書（SGML 文書）の処理方式について報告する。本稿では、構造化文書処理における構造化文書登録処理、表示処理の内容について報告し、検索処理は「構造化文書対応全文検索システム Bibliotheca2 TextSearch の開発(3) —構造指定全文検索方式—」で報告する。

Bibliotheca2 TextSearch では、登録時に文書の構造を解析し、構造内のテキストを構造情報と共に登録する。これにより、構造化文書に対する検索結果として、ヒットした箇所をハイライト表示するための文書を生成することが可能となる。

2．構造化文書処理方式

Bibliotheca2 TextSearch における構造化文書処理の概要を図１に示す。処理内容は、登録処理、検索処理、表示処理の３つから構成される。

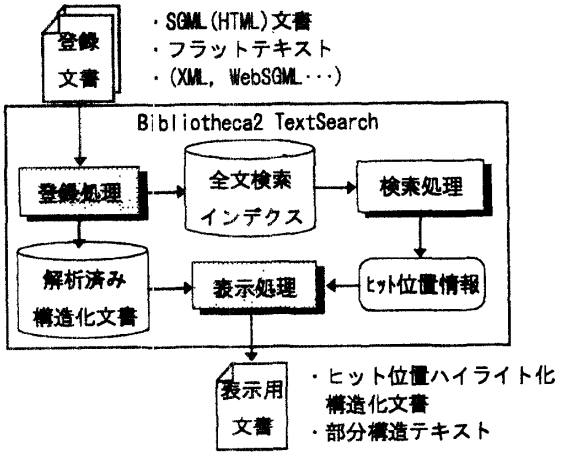

図１ 構造化文書処理の概要

登録処理では、構造化文書の論理構造を解析し、解析済み構造化文書を生成する。さらに構造化文書に対応した全文検索に利用する情報として、全文検索インデクスを生成する。

検索処理では、全文検索インデクス中に、構造情報と共に登録したテキスト中の文字列出現位置情報(n-gram インデクス)を検索することで、高速な構造化文書の全文検索を実現する。検索結果として、ヒットした文書の ID とテキスト中のヒット位置情報を出力する。

表示処理では、登録処理で得られた解析済み構造化文書と、検索処理で得られたヒット位置情報を基に、ハイライト情報を埋め込んだ表示用の SGML 文書を生成する。さらに、指定された部分構造テキストを抽出することも可能である。

3．構造化文書登録方式

図２に示したように、登録処理では登録された SGML 文書を SGML コンパイラによって構造解析し、解析済み構造化文書に変換する。

フラットテキストは文書全体を１構造とする SGML 変換を行い、HTML 文書は検索に用いられる<TITLE>などの構造だけを残した SGML 文書に変換し、SGML コンパイラに入力する。Bibliotheca2 TextSearch では、このように全ての文書を SGML の枠組みで扱うようにしている。

全文検索インデクス登録処理では、まず SGML インデクサで、解析済み構造化文書を基に各 SGML 文書の共通化した構造情報として SGML 構造インデクスを生成する。次に n-gram 抽出によって、解析済み構造化文書から文字列の出現位置情報を生成し、n-gram インデクスに登録する。

---

* Full Text Search System for Large Structured Document Database, Bibliotheca2 TextSearch(2).
† Takuya OKAMOTO, Katsumi TADA, Natsuko SUGAYA, Kanji KATO
‡ Yasushi KAWASHIMO
† Information Systems R&D Division, Hitachi, Ltd.
‡ Software Development Center, Hitachi, Ltd.

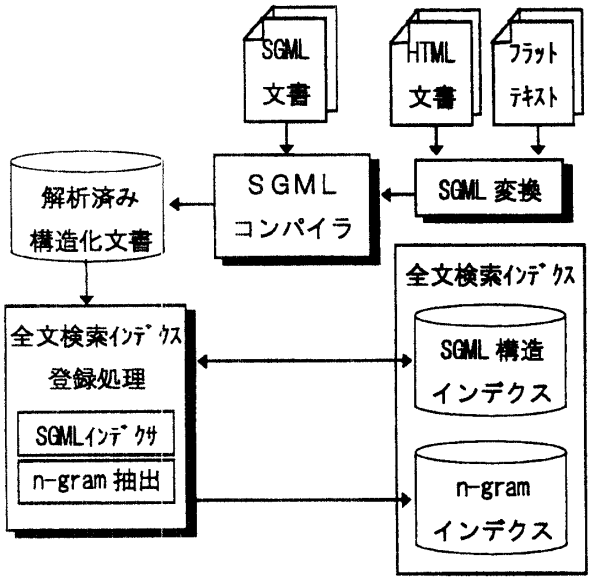

図２ 構造化文書登録処理の内容

４．構造化文書ハイライト表示処理方式

構造化文書ハイライト表示処理では、図３に示すように、解析済み構造化文書と、ヒット位置情報を基に、SGML エクストラクタによってハイライト情報を付加した表示用の SGML 文書を生成する。

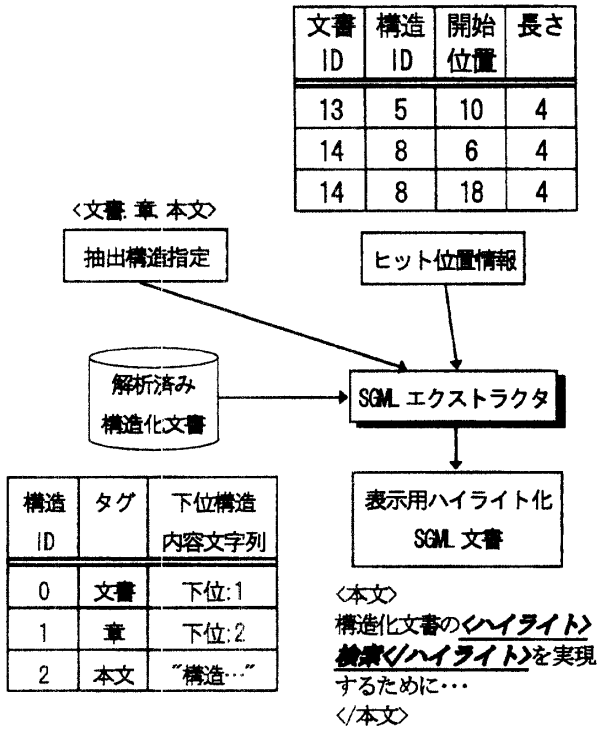

図３ 構造化文書ハイライト表示処理の内容

ヒット位置情報は、文書 ID、構造 ID、ヒットしたタームの構造内での開始位置および長さの情報を持つ。SGML エクストラクタでは、解析済み構造化文書からテキストを読み出し、ヒットしたタームの前後にハイライト用のタグを埋め込んだ SGML 文書を生成する。

また、SGML エクストラクタでは、抽出構造の指定により、部分構造の内容だけを出力することができる。本機能を利用することで、各セクションのタイトル情報を抽出し、目次を作成するといった表示も可能となる。

５．まとめ

構造化文書対応全文検索システム Bibliotheca2 TextSearch における、構造化文書の登録方式、構造全文検索方式、ハイライト表示方式を開発し、以下の機能を実現した。

（１）構造化文書を解析し、全文検索インデクスおよび解析済み構造化文書を生成する構造化文書登録機能
（２）構造を指定した構造化文書全文検索機能
（３）解析済み構造化文書を利用した、検索結果のハイライト表示用文書生成機能
（４）HTML 文書やフラットテキストなど、SGML 以外の文書の SGML 変換機能

上記機能の開発により、フラットテキストのみならず、SGML などで記述された構造化文書と HTML 文書の全文検索、ならびに検索結果のハイライト表示を実現した。今後は、XML や WebSGML など、標準化が進められている構造化文書への対応を行なう。

６．参考文献

[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