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—

4N－5

多田勝己† 岡本卓哉† 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

本稿では、構造化文書対応全文検索システム Bibliotheca2 TextSearch における、SGML 文書の論理構造を指定した検索(構造指定全文検索)の処理方式について述べる。

Bibliotheca2 TextSearch では、登録時に SGML 文書の木構造を重ね合わせることにより、登録済み文書に対して論理構造を一意に識別するための SGML 構造インデクスを生成する。また、登録文書から抽出した各 n-gram に対し、文書番号、構造番号、文字位置を格納した n-gram インデクスを生成し登録する。

そして、検索時には SGML 構造インデクスを参照することにより検索対象とする論理構造の構造番号を取得する。次に、ここで得た構造番号を基に n-gram インデクスを参照することにより、指定された論理構造中に検索タームが含まれる文書を高速に検索することが可能になる。

2．構造指定全文検索方式

2．1 検索方式の概要

Bibliotheca2 TextSearch における構造指定全文検索方式の概要を図１に示す。

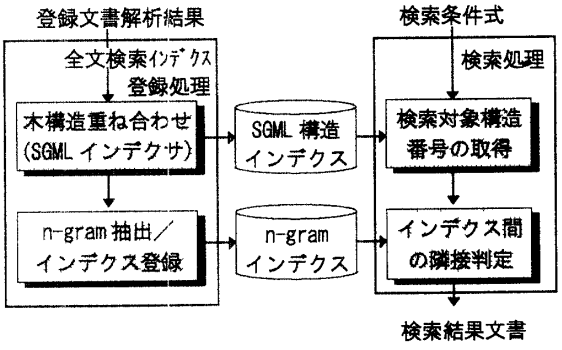


図１ 構造指定全文検索方式の概要

以下、本図における各処理の内容について、登録処理と検索処理に分けて説明する。

2．2 登録処理

（１）木構造の重ね合わせ（SGML インデクサ）

木構造の重ね合わせの概要を図２に示す。

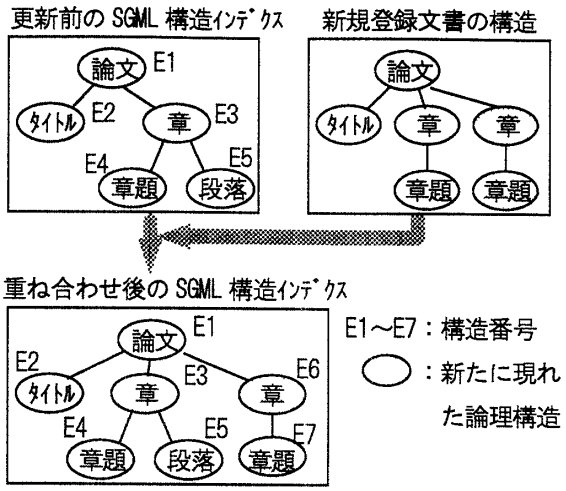


図２ 論理構造の重ね合わせ処理

木構造の重ね合わせ処理では、登録対象文書の持つ論理構造を登録順に従って順次重ね合わせていく。つまり、文書中における出現位置および種別が同じである要素を、同一のメタ要素によって代表させることにより、登録済みの各文書について論理構造を一意に識別するための SGML 構造インデクスを生成する。

なお、SGML 文書における各論理構造の属性値を対象とした検索が実現できるように、属性値に対しても図２に示す論理構造の重ね合わせ処理と同様の処理を行い、論理構造とは別種類の構造番号を割り振るようにしている。

---

* Full Text Search System for Large Structured Document Database, Bibliotheca2 TextSearch(3).
† Katsumi TADA, Takuya OKAMOTO, Natsuko SUGAYA, Kanji KATO
‡ Yasushi KAWASHIMO
† Information Systems R&D Division, Hitachi, Ltd.
‡ Software Development Center, Hitachi, Ltd.

（2）n-gram抽出およびインデクス登録

インデクス作成の対象となるn-gram情報の管理部(トライ)を参照することにより、登録対象文書からn-gramを抽出する。この時、検索に利用されることが少ない文字種の組み合わせで構成される2-gram以上のn-gramを排除する。そして、インデクス作成対象n-gramに対し、該当文書の文書番号、構造番号、文字位置を抽出するとともにデータ圧縮を行い、これをn-gramインデクスとして登録する。

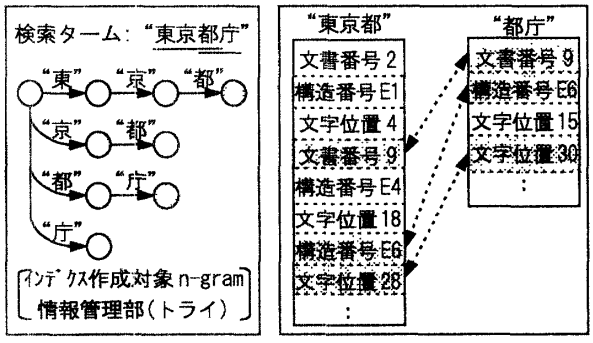
2．3 検索処理

（1）検索対象構造番号の取得

SGML構造インデクスを参照することにより、検索対象に指定された論理構造に該当する構造番号を取得する。また、この時、検索対象構造の上位の構造番号を抽出することにより、例えば図2に示した例では、「章題に“SGML”を含み、同一章内の段落に“全文検索”を含む文書」といった検索が実現できるようになる。

（2）インデクス間の隣接判定

インデクス隣接判定処理の概要を図3に示す。



（a）n-gram抽出処理　（b）隣接判定処理

図3 隣接判定処理

インデクス間の隣接判定処理では、はじめに図3(a)に示すように、インデクス作成対象となるn-gram情報(トライ)を参照し、検索タームを構成する最長のn-gramを抽出することにより、隣接判定処理に使用するn-gramを抽出する。

次に、検索タームから抽出されたn-gramに対するインデクスを取得する。そして、検索対象に指定した論理構造に該当する文字位置情報に着目し、図3(b)に示すように同一文書番号で同一構造番号であり、かつ文字位置が所定の文字数(本図の例では2文字)で隣接している文書を抽出することにより、指定された検索条件を満たす文書を検索することが可能となる。

また、本方式によると検索条件に指定された論理構造中の検索タームの出現頻度や出現位置を正確に取得することができる。これにより、検索ターム間の位置関係を指定した検索や、氏名やタイトルなどの論理構造内における前方・後方一致検索も高速に実現することができる。

3．まとめ

構造化文書対応全文検索システムBibliotheca2 TextSearchにおいて、検索対象構造を一意に識別するためのSGML構造インデクスを作成するSGMLインデクサ方式、および構造指定全文検索対応のn-gramインデクス方式を開発した。これにより、以下の検索を高速に実現することが可能になった。

（1）検索対象とする論理構造を指定した構造指定全文検索

（2）指定した論理構造の下位構造中のいずれかに検索タームが含まれる文書の検索

（3）SGML文書における論理構造の属性値を指定した検索

（4）検索対象論理構造における検索タームの出現頻度によるスコアリングとランキング

（5）検索ターム間で上位構造に同一の論理構造が含まれる文書の検索

（6）氏名やタイトルなどの指定構造内での前方一致・後方一致検索

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