構造化文書ベースシステムにおけるインデックス手法の検討

6P－2

長谷川 知洋† 北川 博之††
†筑波大学 理工学研究科 ††筑波大学 電子・情報工学系

1 はじめに

近年、各種文書の電子化や電子出版が普及しつつある。その国際標準規格であるSGMLはテキスト形式の文書中に構造情報も同時に格納して表現することができる。各種テキスト情報のSGML化が益々進み、SGML化されたデータの量が膨大になると、その中から必要なデータだけを高速に検索したいという要求が高まってくる。

SGML文書のような構造化文書では、各文書ごとに詳細な論理構造は異なるが、共通部分も存在するという特徴がある。また、論理構造に着目することで、構造の一部を条件として指定した検索を行なうことが可能である。

そこで本稿では、構造化文書における部分構造に基づく検索を提案する。その際、共通部分の構造の情報をDTDから抽出して利用することで効率的な検索を行なうことを目的としている。また、検索を高速化するためにシグネチャファイルを利用する。

2 SGML

SGMLとは**Standard Generalized Markup Language**の略で、電子化された各種の文書や文献を扱うための文書記述言語であり、文書データの多角的利用と異機種間の文書交換を目的とした文書の表現形式である。

SGML文書はテキスト文書でありながら、文書の論理構造を明示するためにタグを文書中に埋め込むことができるので、文書内容に関する情報と文書の論理構造に関する情報を同時に表現することができる(構造化文書)。

2.1 SGML文書の特徴

SGML文書は、SGML宣言、文書型定義(DTD)、文書インスタンス(タグ付けされたテキスト)という3つの部分から成り、この順序で文書を構成している。

文書の構造はメモや手紙のように比較的単純なものから、研究論文や技術マニュアルのようにかなり複雑なものまで様々であるが、SGMLを用いて表現可能な文書の構造はそれぞれのDTDで自由に定義できる。その際に同一のDTDに基づいて作成された文書インスタンスであっても、詳細な文書構造は異なることが多い。

An Indexing Scheme for Structured Document Base Systems
Tomohiro HASEGAWA†, Hiroyuki KITAGAWA††
†Master's Degree Program in Science and Engineering, Univ. of Tsukuba
††Institute of Information Sciences and Electronics, Univ. of Tsukuba

2.2 SGML文書を対象とした問合せ

SGML文書を対象とした検索として、[1]を参考にすると以下のような4種類の問合せが考えられる。

1. 文書内容に関する検索
2. 文書構造に関する検索
3. 文書内容とは独立にエレメントに付加される属性と属性値に関する検索
4. 上記を組み合わせた検索

本稿で対象とするのは2の文書構造に関する検索である。

問合せ例： エレメントe3のサブエレメントとしてエレメントe10を含むような文書インスタンスを探せ

3 部分構造に基づく検索

問合せの対象となる文書インスタンスは、その論理構造に着目すると木構造として表現することができる。

様々な種類のDTDが存在し、各DTDに対して複数の構造を持った文書インスタンスが存在しているような場合を考えると、膨大な数の木の中から欲しいものだけを効率的に検索するのは非常に困難である。本稿では、その効率的な検索手法として[2]に基づき、木の部分構造を条件とした検索を提案する。その基本方針は以下の通りである。

1. 検索対象となる全ての木に対して、木を構成する要素(**特徴構造要素**と呼ぶ)を抽出し、インデックス化しておく。
2. 問合せ条件として特徴構造要素を与え、それを含んでいるような木をインデックスを用いて探す。

また、あるDTDに基づく文書インスタンスの集合に注目した場合、木を構成する特徴構造要素は以下の2種類に分類される。

- 全ての文書インスタンスに必ず出現するもの
- 各文書インスタンスに固有のもの

各文書インスタンスごとに木を構成する要素の情報を保持していたのでは前者の情報が冗長になってしまう。そこで、この共通の情報は該当するDTDから抽出して別に保持し、各文書インスタンスは後者の情報だけを保持するようにする。

4 検索アルゴリズム

部分構造に基づく検索を高速に行なうためにシグネチャファイル[3]を利用する。図1の例を用いてシグネチャの作成手順及び検索手順を示す。

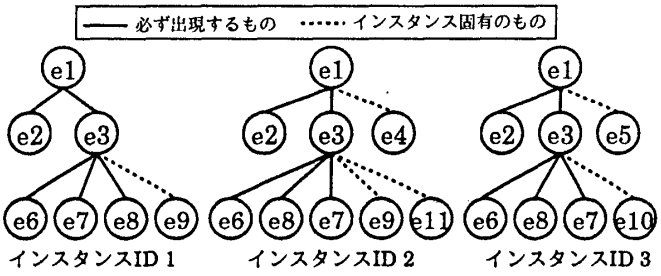

図 1: 文書インスタンスの木構造

4.1 インデックス作成手順

1. DTD の情報から全文書インスタンスに必ず出現する共通部分を抽出し、特徴構造要素となる 2 ノード間の親子関係を組とし、それに対するシグネチャを作成する。
2. 作成したシグネチャのビットごとの論理和をとり、集合シグネチャを作成する。この集合シグネチャを**最小構造シグネチャ**と呼ぶ (図 2)。

特徴構造要素	シグネチャ
e1—e2	0 0 1 0 1 0 0 0
e1—e3	0 0 0 1 1 0 0 0
e3—e6	1 0 0 1 0 0 0 0
e3—e7	0 0 1 0 0 0 1 0
e3—e8	0 0 0 1 0 0 1 0
最小構造シグネチャ	1 0 1 1 1 0 1 0

図 2: 最小構造シグネチャの作成

3. 各文書インスタンスに対し、インスタンスに固有な部分を抽出し、同様の手順で集合シグネチャを作成する。この集合シグネチャと文書インスタンスの ID を組にしてシグネチャファイルに格納する。例えば、ID 1 のインスタンスに固有な部分は e3-e9 であり、ID 2 のインスタンスでは e1-e4 と e3-e9 と e3-e11、ID 3 のインスタンスでは e1-e5 と e3-e10 である。
4. ある DTD に基づく文書インスタンス群に出現する全ての構造を抽出し、同様の手順で集合シグネチャを作成する。この集合シグネチャを**最大構造シグネチャ**と呼ぶ。この例では、e1-e2、e1-e3、e3-e6、e3-e7、e3-e8、e1-e4、e1-e5、e3-e9、e3-e10、e3-e11 に対してそれぞれシグネチャを作成し、ビットごとの論理和をとることで最大構造シグネチャを作成する。

4.2 検索手順

1. 問合せ条件で与えられた特徴構造要素に対して、シグネチャを作成する。これを**問合せシグネチャ**と呼ぶ。この例での問合せ条件は e3-e10 なる組を持つことである。
2. 問合せシグネチャと最大構造シグネチャを比較し、問合せシグネチャで 1 が立っているビット位置に最大構造シグネチャでも全て 1 が立っているならば 3 へ、そうでなければ該当する文書インスタンスが存在しないということになる。
3. 問合せシグネチャと最小構造シグネチャを比較し、最小構造シグネチャで 1 が立っていないビット位置に問合せシグネチャで 1 が立っているならば 4 へ、そうでなければ 5 へ。
4. 最小構造シグネチャで 1 が立っていなくて、問合せシグネチャで 1 が立っているビット位置の情報をもとに、シグネチャファイル中の全シグネチャの中から、そのビット位置に 1 が立っているようなシグネチャを問合せ条件を満たす候補として選び出し、フォルスドロップレゾリューションによって実際に条件を満たす文書インスタンス (インスタンス ID 3) だけを得る。
5. 該当する DTD を調べ、与えられた問合せ条件が全ての文書インスタンスに必ず出現する共通部分であるかどうかを調べる。共通部分ならば文書インスタンス全てが問合せ条件を満たすことになり、そうでない場合は全文書インスタンスを調べる必要がある。

5 おわりに

本稿では、構造化文書におけるシグネチャファイルを用いた部分構造に基づく検索を提案した。その際、必ず出現する構造の情報は DTD から抽出して利用した。本稿では、部分構造として 2 ノード間の組を特徴構造要素としたが、より多くのノード間の組を特徴構造要素とすることなども考えられる。

今後の課題としては、本手法に対する定量的な評価を行なうこと、また構造に関する検索だけでなく文書内容や属性に関する検索を扱うことができるようにすることなどが挙げられる。

参考文献

[1] Ron Sacks-Davis, Timothy Arnold-Moore and Justin Zobel, "Database Systems for Structured Documents," Proc. ADTI'94, October 1994

[2] Hans Argenton and Peter Becker, "Efficient Retrieval of Labeled Binary Trees," IEICE TRANS.INF SYST.,Vol.E78-D,No.11 November 1995

[3] Yoshiharu Ishikawa, Hiroyuki Kitagawa and Nobuo Ohbo, "Evaluation of Signature Files as Set Access Facilities in OODBs," Proc. ACM SIGMOD 1993