構造化文書の部分構造検索のための索引方式の設計と評価

3S－3

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は文書中にテキスト情報と論理構造情報を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文書では、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。

従来のテキスト検索で用いられてきた索引では構造化文書に特有な文書の論理構造に関する検索を十分に支援することができない。このような検索を支援するためには、構造化文書が持つ構造情報をいかに扱うかが大変重要であると考えられる。そこで本稿では、この構造情報を手がかりとして、高速かつ効率的な検索を支援する索引機構の検討及び評価を行なう。

2 構造化文書検索

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 検索対象となるSGML文書

SGML文書を対象とした検索として、[1]を参考にすると文書内容に関する検索、文書構造に関する検索、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、上記を組み合わせた検索等が考えられるが、本稿で対象としているのは**文書構造に関する検索**である。

このような検索を支援するためには、構造化文書が持つ構造情報をうまく扱うことが大変重要である。そこで我々は文書の構造定義をしているDTD(Document Type Definition)に着目し、DTDの情報を用いて文書構造に関する検索を行なう。

以下にDTDの例とそのDTDに基づく文書インスタンス群に対する問合せの例を示す。

```
<!ELEMENT memo      - -  (prolog?, body)>
<!ELEMENT prolog    - O  (date, (from & to),
                          subject?)>
<!ELEMENT (date | from | to)  - O  (#PCDATA)>
<!ELEMENT body      O O  (p)+>
<!ELEMENT (subject | p)  - O  (#PCDATA | q)*>
<!ELEMENT q         - -  (#PCDATA)>
```

DTDの例

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

2.2 部分構造検索

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ことで木構造として表現することができる。

ラベル付き2分木に対して部分構造検索を行なった[2]を参考に、本稿では木構造とみなした文書インスタンス群に対する部分構造検索のための索引方式を提案する。

その基本方針は以下の通りである。

- 検索対象となる全ての木に対して、木を構成するノード(エレメント)を抽出し、索引を作成する。
- 問合せ条件として与えられた部分構造からも同様にノードを抽出し、それら全てを含んでいるような木を索引を用いて探す。

3 シグネチャファイルを用いた2段階検索

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群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。

そこで、初めにDTDレベルでの絞り込みを行ない、次に該当するDTDに基づく文書インスタンス群に対しての絞り込みを行なうことにする。

また、部分構造検索を高速に行なうための索引としてシグネチャファイル[3]を利用する。

3.1 DTDの絞り込み

DTDレベルでの絞り込みを可能にするために、各DTDに対してシグネチャ(DTDシグネチャと呼ぶ)を作成する。DTDシグネチャの作成方法は以下の通りである。

1. 1つのDTD中に出現し得るエレメントを全て抽出し、ハッシング等によりそれらに対するシグネチャをそれぞれ作成する。

An Indexing Scheme for Substructure Retrieval of Structured Documents

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. 作成したシグネチャのビットごとの論理和をとったものをDTDシグネチャとする。

3.2 エレメントのグループ化

文書インスタンスレベルでの絞り込みを可能にするために、各文書インスタンスに対してシグネチャ(インスタンスシグネチャと呼ぶ)を作成する。DTDシグネチャの作成方法と同様に、1つの文書インスタンス中に出現し得る全てのエレメントを抽出し、それらに対してそれぞれシグネチャを作成して、ビットごとの論理和をとることによってインスタンスシグネチャを作成する。

しかしDTDの情報を用いると、文書インスタンスの論理構造中には、ある決まったエレメント群で構成される部分構造が出現し得ることがわかる。そのような部分構造を1つのグループと見なすと、グループに含まれるエレメントがどれか1つでも出現すれば、残りのエレメントも必ず出現するので、同一グループ中の各エレメントに対して異なるシグネチャを割り当てても意味がない。

そこでグループ中の代表となるエレメントを決め、代表エレメントのシグネチャをグループ中のエレメントのシグネチャとして共有することで効率化を図る。

例： エレメント prolog が出現する文書インスタンス中には、エレメント date、from、to が必ず出現することがわかるので、これらのエレメントは1つのグループ(この場合はグループ2)として扱う。

グループ1： {memo, body, p}
グループ2： {prolog, date, from, to}
グループ3： {subject}
グループ4： {q}

エレメントのグループ化の例

3.3 グループ間の出現従属性

前述のようにしてできたグループ間には、「あるグループ A が出現しなければ、別のグループ B は出現し得ない」という**出現従属性**が存在し得る。問合せの際にこの性質を用いることによって効率化を図る。

例： グループ2に含まれるエレメント prolog はグループ1に含まれるエレメント memo に依存して出現するので、グループ2はグループ1に出現従属している(図1)。

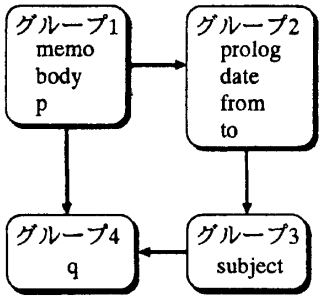

図 1: グループ間の出現従属性

4 検索手順

前述の問合せに対する処理手順を示す。

1. 問合せの条件として与えられた部分構造を構成するエレメント prolog、from、subject に対してそれぞれシグネチャを生成し、ビットごとの論理和をとることで問合せシグネチャを生成する。
2. 問合せシグネチャとDTDシグネチャを比較することよって、問合せ条件を満たす文書インスタンスが、様々な種類のDTDに基づいた全文書インスタンスのうち、どのDTDに基づいた文書インスタンス群に含まれるのかという絞り込みを行なう。
3. グループ化の規則を用いることで、エレメント prolog、from、subject がグループ2とグループ3に分類できることがわかる。
   この時、グループ1はこのDTDに従う全文書インスタンスに必ず出現するエレメントのグループなので、問合せ条件で与えられたエレメントが全てグループ1に属す場合は、全文書インスタンスが検索条件を満たすことになり、検索は終了する。
4. グループ3はグループ2に出現従属しているので、グループ3の出現があるならば、必ずグループ2も出現していることになるので、グループ3のシグネチャだけを問合せシグネチャとして用いればよいことになる。
5. 問合せシグネチャで1が立っているビット位置に全て1が立っているようなインスタンスシグネチャが問合せを満たす候補となる。

5 おわりに

本稿では、シグネチャファイルを用いた構造化文書の部分構造検索のための索引方式を提案した。その際、構造情報はDTDの情報を利用することを提案した。また、エレメントのグループ化やグループ間の出現従属性を利用することにより検索効率の向上を図った。

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