シグネチャファイルを用いた構造化文書の索引・検索方式の設計と評価

2Q－4

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等が普及しつつある。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化が今後益々進み、SGML化されたデータの量が膨大になると、それらをデータベース化して管理したり、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。

我々は、構造化文書を対象とした検索における検索方式、またその検索を支援するための索引方式についての研究を行なっている。本稿では、構造化文書の論理構造に関する検索をより効率的に行なうことを可能にする索引方式を提案し、その評価を行なう。

2 構造化文書

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のタグを文書中に埋め込み、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という特徴がある(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や文書構造に関する検索やこれらを組み合わせた検索などが考えられる。我々は、これらの検索を支援可能にする索引方式の研究を行なっているが、本稿ではこのうち特に、**文書構造に関する検索**を支援する索引部についての説明を行なう。

SGMLでは、文書の論理構造は**文書型定義(DTD)**で自由に定義できるが、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という特徴がある。以下にDTD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)>
```

2.2 部分構造検索

様々な種類のDTDが多数存在し、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、膨大な数の文書インスタンス群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。希望の文書インスタンスを高速に検索する一手段として、文書の論理構造に着目し、部分構造検索を行なうことが考えられる。本研究では、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[2]を利用する。左記DTDに基づく文書インスタンス群に対する問合せの例を以下に示す。

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

3 シグネチャファイルを用いた構造化文書の検索方式

膨大な数の文書インスタンスの中から希望の文書インスタンスを高速に検索するために、検索手順を二段階に分割して考える。まず初めに、問合せ条件を満たし得る文書インスタンスを持つDTDの検索を行なう。次に、条件を満たしたDTDに基づく文書インスタンス群に対しての検索を行なう。

3.1 DTDの絞り込み

多数存在する文書インスタンス群の中には、明らかに検索条件を満たさないものが存在し得る。これらは全文書インスタンスを調べなくても該当するDTDを解析することで判別可能である。そこでまず、DTDレベルでの検索を行なう。DTDレベルでの検索を可能にするため、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シグネチャは、DTD中に出現する全エレメントに対してシグネチャを作成し、それらのビットごとの論理和をとることで作成する。

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検索を支援するため、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。しかしDTDの情報を用いると論理構造中には、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

An Indexing and Retrieval Scheme for Structured Documents based on Signature Files

Tomohiro HASEGAWA†, Hiroyuki KITAGAWA††

†Master's Degree Program in Science and Engineering, Univ. of Tsukuba

††Institute of Information Sciences and Electronics, Univ. of Tsukuba

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のどれか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： DTDの解析からエレメント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」という性質が存在する。この性質のことをグループ間の**出現従属性**と呼び、グループBはグループAに出現従属しているという。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。

3.4 検索手順

前述の問合せに対する処理手順を以下に示す。

1. 問合せ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から問合せシグネチャを作成する。
2. 問合せシグネチャとDTDシグネチャを比較し、問合せ条件を満たす文書インスタンスがどのDTDに基づいているものなのかを調べる。
3. グループ化規則の適用によりエレメントprolog, from, subjectをグループ2とグループ3に分類する。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群をグループ化したもの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が出現していれば、必ずグループ2も出現していることになる。よって、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条件を満たす候補となる。

4 評価

提案した索引方式の有効性を調べるため、通常のシグネチャファイルを用いた場合と本索引方式を用いた場合について、[3]を参考に**検索コストとフォルスドロップ確率**のコスト式を確率論的に見積もり、計算することで比較を行なった[4][5]。さらにまた、実データに基づくパラメータ値を設定したシミュレーションを行ない、コスト式の検証を行なった。本索引方式では、検索効率向上のための手法を三種類(DTDの絞り込み、エレメントのグループ化、グループ間の出現従属性)提案したが、それぞれの効果が明確になるようにそれらを単独で採用した場合のシミュレーションを行なった。そのうち本稿では、**DTDの絞り込み**を採用した場合の検索コストと[4]で見積もった検索コストの計算結果を図1に示す。

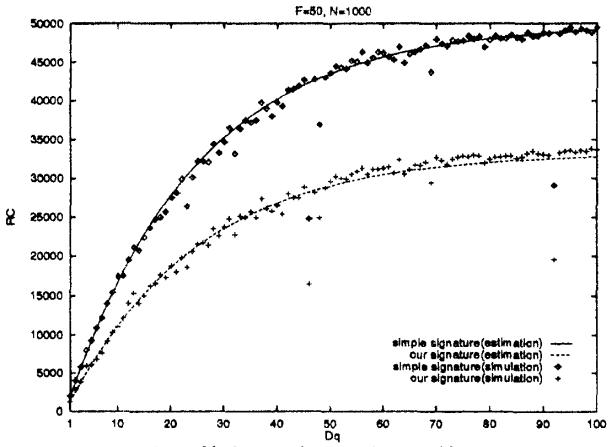

図1: 検索コスト(RC)の比較

図1より[4]で見積もりを行なったコスト式がほぼ正しいことが検証できた。このことから、実データを用いた実験でもDTDの情報を用いることが検索コスト削減の有効な手段であることが予想される。

5 おわりに

シグネチャファイル技法を用いた構造化文書検索の索引・検索方式について論じた。特に、文書構造に関する検索を支援する索引部において検索効率を向上させるために、構造情報の取得にDTDを利用すること、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等を提案し、評価を行なった。

参考文献

[1] R. Sacks-Davis, T. Arnold-Moore and J. Zobel, "Database Systems for Structured Documents," Proc. ADTI'94

[2] C. Faloutsos and S. Christodoulakis, "Signature Files: An Access Method for Documents and Its Analytical Performance Evaluation," ACM Trans. Office Inf. Syst.,Vol.2,No.4,1984

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

[4] 長谷川, 北川, "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," 情報処理学会第52回全国大会 1996

[5] 長谷川, 北川, "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," 情報処理学会第53回全国大会 1996