5J-01　XMLをベースとしたタグ格納による検索方法の提案とその効率化

近藤 博[†]　川越恭二[‡]
立命館大学大学院理工学研究科[†]
立命館大学理工学部[‡]

1 はじめに

現在，多くの分野で文書管理にXMLが利用され始めており，XMLで記述された文書が増大していくことが予想される．しかし，これらを効率良く格納，検索する技術はまだ十分ではない．

そこで本稿ではXMLをベースとするタグを利用したXML文書の効率的な検索方式を提案する．本方式は従来のように，検索対象データをシステム側に索引データとして保持するのではなく，タグと文書の場所のみを格納し直接対象データ内を検索するものである．この方法により．頻繁に更新が行われるXML文書の検索にも対応可能となる．さらに検索キャッシュと部分インデックスにより効率化した方法を以降で説明する．

2 XMLタグ格納方式

提案する方式は物理的に分散されたサーバに分割，格納されたXML文書ファイル集合を直接検索する方法を実現するものである．

2.1 基本方式

XML文書ファイルF_iの集合Fを検索対象とする．また，XML文書ファイルはXMLタグとそれに対応するタグ値のペア集合からなる．すなわち，任意の$F_i \in F$について，$F_i=\{(T_i(j),V_i(j))\}$とする．ここで，$T_i(j)$はタグ集合の要素であり，XML文書ファイルF_iの中に記述されているj番目のタグを示す．$V_i(j)$このタグに対応するタグ値である．さらに，XML文書ファイルは適当に設定された格納アドレスAddress(F_i)が割り当てられ，Address(F_i)が与えられればF_iを取り出すことが出来るものとする．

この文書群に対する検索を行うために指定する質問Qは以下のように記述されるものとする．$Q=\{(T_k, V_k)\}$ここで(T_k, V_k)は質問Qを構成するk番目のタグとタグ値のペアである．したがって，質問Qはk番目のタグとタグ値のペアを持つXML文書ファイルをFから選択することを意味する．質問Qに対する検索結果R_QはK_Q個のXML文書ファイル集合$\{F_k\}$である．したがって，{Address(F_k)}が求めるべき結果となる．

上記の結果を得るためにシステム内に保持するデータファイルDは(T_s,A_s)の形式をとる．ここで，T_sはタグであり，A_sはタグT_sを持つXML文書ファイルF_{Ts}の格納場所を示す．すなわち，A_s=Address(F_{Ts})である．

質問Qが与えられた時，システムはデータファイルDを用いて以下の手順でR_Qを求める．

Step-1

次の条件を満たす格納場所の集合$A=\{A_i\}$を求める．“任意の$A_i \in A$に関して，質問QのすべてのT_kについて$(T_k,A_i) \in D$が成り立つ”．

Step-2

任意の$A_i \in A$について，A_i=Address(F_{Ti})なる$F_{Ti} \in F$を取り出し，次の条件を満たせばA_iをR_Qに含める．“任意のj、kについて$T_{Ti}(j)=T_k$が成り立つならば，$V_{Ti}(j) \in V_k$が成り立つ”．ここで，$T_{Ti}(j)$，$V_{Ti}(j)$はタグT_iを持つXML文書ファイルF_{Ti}のタグおよびタグ値である．

2.2 処理の効率化

前節で述べた基本手順をより処理効率化するために，以下の2つの方法を導入する．

・部分インデックスの利用

Retrieval Method for XML documents by storing XML Tags
Hiroshi KONDOH,Kyoji KAWAGOE.
Ritsumeikan University
1-1-1,Noji-Higashi,Kusatsu,Shiga,525-8577,Japan

・検索キャッシュの利用

部分インデックスとは，予め頻繁にアクセスされるXML文書ファイルの部分集合 $F' \subset F$ について作成したインデックスファイル I={T(m), V(m),Address(Fm)}である．ここで $F_m \in F'$であり，T(m)はXML文書ファイル F_m の中の適当なタグ，V(m)はそのタグに対応したタグ値である．

検索キャッシュとは，最近アクセスしたタグおよびタグ値とそれを格納しているアドレスの組からなるキャッシュインデックスファイル C={(T[p],V[p],A[p])}である．ここで，T[p], V[p], A[p]は最近p番目にアクセスしたタグ，タグ値，アドレスである．

この2つの方法を用いることで，前節のStep-2をStep-2'に改良する．

Step-2'

(1)もし，"すべての k について T_k=T[p], V_k=V[p], A_i=A[p], (T[p],V[p],A[p])∈C となる p が存在する"という条件を満たすような $A_i \in A$ があれば，A_i を R_Q に含める．この A_i を A から削除する．

(2)もし"すべての k について T_k=T(m), V_k=V(m), A_i=Address(Fm), (T(m),V(m),Address(Fm))∈Iとなる m が存在する"という条件を満たすような $A_i \in A$ があれば，A_i を R_Q に含める．この A_i を A から削除する．

(3)もし A_i=Address(F_{Ti})∈A なる $F_{Ti} \in F$ を取り出し，次の条件を満たせば A_i を R_Q に含める．"任意のj, kについて $T_{Ti}(j)=T_k$ が成り立つならば，$V_{Ti}(j) \in V_k$ が成り立つ"．

2.3 効率化方法の有効性，問題点

部分インデックスを利用した場合，データを参照する時間は大幅に短縮できる。しかし，インデックスファイルを保持するためデータ量の増加によるシステムへの負荷が増加する．また，検索結果の新鮮度が低下するといったデメリットも生じる．

検索キャッシュを利用する場合，ゆるい検索条件を指定して検索を行うとキーワードに適合する確率が下がるため，キャッシュファイルを検索する処理が無駄になるという問題がある．しかし，特定の分野での検索など，必然的に検索キーワードやタグが限定されるような検索に対して適応すれば検索の効率化が実現できると考えられる．

3. 実装

提案した方式を元にXML文書論文に対する検索システムを試作した．

3.1 システムの構成

システム側は文書中のタグ(タイトル，著者名，発表年度)とその文書の所在場所(アドレス)を保持する．検索要求に対して指定されたタグに対応するアドレスを元に対象文書内を直接検索し，キーワードと合致したもののアドレスを出力する．

3.2 制約事項

事前にXMLのメタデータRDF[1]を文書ファイル内に埋め込んで置くことが必要である．

3.3 アプリケーションの試作

試作システム，実行結果を図1，図2に示す．

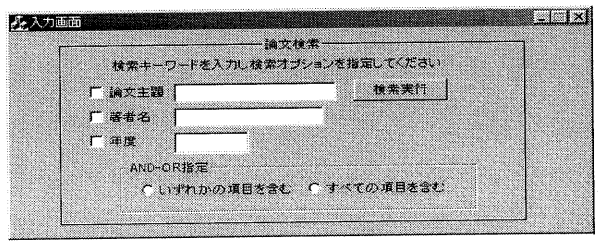

図1.論文検索システム

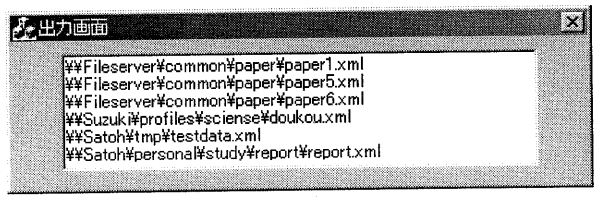

図2.実行結果

4. おわりに

今後は，処理性能の評価や検索の効率化を予定している．

参考文献

[1]"Resource Description Framework(RDF)"
http://www.w3.org/RDF/