제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

XML Schema에 의한 XML 데이타베이스의 타입 상속 색인구조

°임윤주, 이종학
대구가톨릭대학교 컴퓨터정보통신공학부
e-mail:zasmine8071@lycos.co.kr

A Type Hierarchy Index for XML Databases with XML Schema

°Yun-Ju Lim, Jong-Hak Lee
School of Computer and Informaion Communications Engineering, Catholic Univ. of Daegu

요 약

최근 XML데이터베이스는 웹의 발전과 더불어 광범위한 인터넷의 자원 공유에 크게 기여하고 있으며 이러한 자원 공유를 위해서는 XML데이타베이스에 대한 구조적 정의로 타입 상속 구조를 가지는 XML Schema를 사용한다. 그러므로 XML Schema를 따르는 XML데이타베이스에 대한 효율적인 색인기법에 대한 연구가 필요하다. 따라서 본 논문에서는 기존의 다차원 색인구조와 사전에 분석한 사용자 질의 패턴에 대한 정보를 이용하여 주어진 질의들에 의해서 액세스되는 색인 페이지의 평균 개수가 최소가 되게 하는 최적의 이차원 타입 색인 구조를 구성할 수 있는 2D-THI를 제안한다. 제안한 2D-THI의 성능을 비교 평가하기 위해서 기존의 객체지향 데이터베이스에서 클래스 상속에 대한 색인구조로 널리 사용되고 있는 CH-index와 CG-tree를 XML데이타베이스에 적용하여 이들과 2D-THI를 비용모델을 통해서 비교 분석한다. 그 결과로 본 논문에서 제안한 2D-THI로서 다양한 질의 패턴에 대해서 최적의 색인구조를 구성할 수 있음을 보인다.

1. 서론

최근 XML(eXtensible Markup Language)데이터베이스[1]는 웹의 발전과 더불어 광범위한 인터넷의 자원 공유에 크게 기여하고 있다. 그러나 다른 사람과 자원을 공유하기 위해서는 데이타의 구성 방법에 대한 정확한 규정이 있어야만 신뢰성을 가질 수 있다. 따라서, XML데이타베이스 안에 어떤 요소형과 특성, 값들을 사용할 수 있는지에 대한 정의를 위해서 XML에서는 대표적인 데이타베이스의 구조 정의 중 DTD(Data Type Definition)를 널리 사용하고 있다[2].

하지만, DTD는 정보 요소(element), 속성(attribute), 그리고 데이터 형 정의에 대한 상속의 매커니즘을 제공하지 않는다는 문제점을 가지고 있다. 이를 해결하기 위해서 W3C(World Wide Web Consorsium)에서는 XML Schema[3, 4]를 제안하였다. XML Schema는 DTD보다 다양한 데이터 타입을 정의할 수 있고, 속성, 정보 요소, 사용자 정의 데이터 타입에 대해서 상속하는 구조를 가지고 있다. 즉, 타입 상속을 통하여 엘리먼트들의 구조를 재사용할 수 있다. 따라서 이러한 장점을 가진 XML Schema를 따르는 XML데이타베이스에 대한 효율적인 색인기법에 대한 연구가 필요하다.

한편, 기존의 객체 지향에서 널리 사용되고 있는 클래스 상속 계층에 대한 색인구조로 B$^+$-tree[8]를 확장하여 사용하는 CH-index[6]와 CG-tree[7]가 있다. 이러한 색인구조들을 XML 데이타베이스에도 적용할 수 있다. 그러나, 이들 색인구조들은 일차원의 색인구조인 B$^+$-tree의 특성 때문에 XML Schema를 따르는 XML데이타베이스에 주어지는 여러 형태의 질의들 중 특정 형태의 질의에만 좋은 성능을 보이며 다른 형태의 질의에 대해서는 비효율적인 성능을 보인다.

따라서, 본 논문에서는 이러한 문제점을 해결하기 위하여 XML데이타베이스의 타입 상속에 대한 인덱스 구조로 이차원 색인구조를 이용하는 2D-THI(2-Dimensional Hierachy Index)를 제안한다. 2D-THI는 엘리먼트 도메인과 타입 도메인으로 이루어진 이차원 도메인 공간상의 색인 엔트리들의 클러스터링 정도를 조정할 수 있는 이차원 타입 상속 색인구조이다. 2D-THI에서는 사전에 분석한 사용자 질의 패턴에 대한 정보를 이용하여 색인 엔트리들의 클러스터링 정도를 주어진 질의 패턴에 적합하도록 조정함으로써 항상 최적의 색인구조를 구성할 수 있다.

2. 관련 연구

XML은 HTML(Hyper Text Markup Language)을 대체할 목적으로 W3C에서 표준화 작업을 진행하고 있는 차세대 인터넷 전자 데이타베이스 표준이며, HTML과 SGML(Standard Generalized Markup Language)의 장점을 모두 가지도록 규정한 구조화된 정보를 포함하고 있는 데이타베이스들을 위한 마크업 언어이다[1]. 즉, XML은 데이타베이스의 내용과 형태를 다양하게 정의할 수 있는 강력한 기능을 가진 SGML에 기반을 두고 HTML의 단순함과 유연성을 고려하여 새롭게 만들어진 마크업 언어이다. 그리고 XML은 데이타베이스 내용의 구조화와 데이타를 보여주기 위한 스타일을 명백하게 구분함으로써 데이타베이스의 내용을 변경하지 않아도 다양한 형태를 갖는 데이타베이스를 만들어낼 수 있다.

또한, XML은 DTD를 통하여 정보 교환의 이점을 제공해 준다. DTD는 XML데이타베이스를 구성하는 정보 요소, 정보 요소의 구조와 속성등 데이타베이스의 형태를 구조화하여 정의한 것으로, 데이타베이스의 조직 규칙을 정의한 것이다. 하지만 DTD는 정보 요소, 속성, 그리고 데이터 형 정의에 대한 상속의 매커니즘을 제공하지 않는 등 여러 가지 문제점을 지니고 있으므로 W3C에서는 이를 개선하기 위해서 XML Schema를 제안하였다.

XML Schema는 속성, 정보 요소, 사용자 정의 데이타 타입에 대해서 상속하는 구조를 가지는 표준규약으로 DTD보다 다양한 데이타 타입을 정의 할 수 있다. XML Schema는 기본적으로 엘리먼트들간의 구조라 정의하는 type들로 구성되며 simple type과 complex type으로 나누어진다. simple type은 string,

byte, integer, date를 비롯한 40가지 이상의 built-in 타입들 을 제공하고 있으며, complex type은 속성, 엘리먼트를 포함한다 [5].

다음 그림은 XML Schema의 상속 구조의 예이다. VertebrataType안에 타입 상속을 나타내는 문법인 <extension base="AnimalType"> 을 통해서 cervical_num, lumbar_num 엘리먼트 이외에 AnimalType 안에 life_span, name을 엘리먼트로 가진다.

```
<complexType name="AnimalType">
    <element name="life_span"  type="int">
    <element name="name"    type="string">
</complexType>
<complexType name="VertebrataType">
  <complexContent>
      <extension base="AnimalType">
          <sequence>
              <element name="cervical_num" type="int">
              <element name="lumbar_num" type="int">
          </sequence>
      </extension>
  <complexContent>
</complexType>
```

그림 1. XML Schema의 상속 구조의 예

그림 2는 root를 Animal로 가지는 XML Schema Graph의 예이다. 엘리먼트는 타원으로 나타내고 타입은 사각형으로 나타내며, 엘리먼트간의 내포 관계를 화살표가 있는 실선으로 나타낸다. 그리고 화살표가 없는 실선은 엘리먼트와 타입간의 관계를 나타낸다. 그리고 타입간의 상속관계는 화살표가 있는 점선으로 나타내며, 화살촉이 있는 쪽이 부모 타입이 되고 반대편이 자식 타입이 된다.

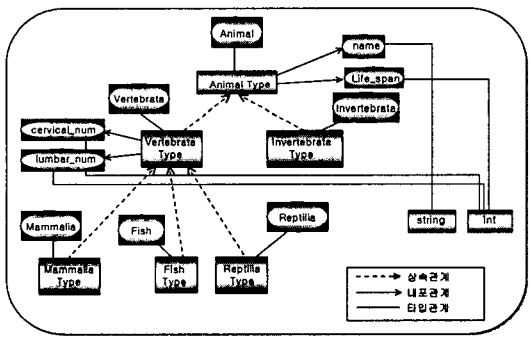

그림 2. XML Schema Graph의 예

한편, 기존의 객체 지향 데이터베이스에서 널리 사용되고 있는 클래스 상속 계층에 대한 색인기법은 클래스 구분없이 키 값 순서로 클래스 계층구조 내의 모든 객체들을 클러스터링하는 방식인 CH-index[6]와 클래스 단위로 객체들을 먼저 구분한 후 키값 순서대로 클러스터링하는 CG-tree[7]가 있다. 이러한 색인기법들은 하나의 속성에 의해 클러스터링이 이루어지는 일차원 색인구조인 B+-tree를 사용함으로써 다양한 질의 형태에 대한 질의들을 모두 잘 지원하지 못하는 문제점을 가지고 있다.

CH-index는 하나의 클래스를 대상으로 키값에 대한 범위 질의(range queries)를 처리하는 경우에, 같은 클래스의 객체들이 키값을 단위로 구분되어 여러 곳에 분리 저장되므로 많은 읽기 연산의 오버헤드가 있어 비효율적이다. 반면에, CG-tree는 클래스 계층을 대상으로 한 부합 질의(match queries)를 처리하는 경우에, 동일한 키 값을 가지는 객체들이라도 클래스 단위로 구분되어 여러 곳에 분리 저장되므로 오버헤드가 발생하게 되어 비효율적이다. 따라서, 이러한 기존의 객체지향 데이타베이스의 클래스 계층 색인기법들은 색인구조를 구성하는 색인 노드들의 형태가 고정됨으로써 특정 형태의 질의에 대해서만 효율적인 색인성능을 가지며, 그 외의 질의 형태들에 대해서는 비효율적으로 된다.

3. XML 데이터베이스의 타입상속에 대한 색인구조

먼저, 본 절에서는 XML 데이터베이스에서의 타입 계층 색인 기법을 논하기 위하여 XML Schema를 따르는 XML 데이타베이스의 상속 계층 구조에 대한 질의를 다음과 같은 세가지 형태로 분류한다. 여기서, 타입 T^*는 타입 T와 그의 모든 서브 타입들을 원소로 하는 집합으로 정의한다. 예를 들어 $Animal^*$는 집합 {Vertebrata, Invertebrata, Mammalia, Fish, Reptilia}이며, $Vertebrata^*$는 집합 {Mammalia, Fish, Reptilia}이다.

1. STR(Single Type Range queries): 특정한 하나의 타입 T를 대상으로 하는 범위 질의
2. THM(Type Hierachy Match queries): 특정 타입 T의 타입 집합 T^*를 대상으로 하는 부합 질의
3. THR(Type Hierachy Range queries): 특정 타입 T의 타입 집합 T^*를 대상으로 하는 범위 질의

Robinson[9]에 의하면, 사용자가 요구하는 모든 질의는 도메인 공간내의 영역들로 표현할 수 있으며 이 영역은 도메인을 구성하는 축에 대한 구간들의 곱으로 표현된다. 그리고 이러한 영역을 질의 영역(query region)이라고 정의하였다. 그 정의에 의해서 본 논문에서는 타입 상속 계층을 대상으로 하나의 키 엘리먼트나 속성에 대한 질의 조건으로 주어지는 질의는 키값 도메인과 타입 식별자 도메인으로 구성된 이차원 도메인 공간상의 이차원 질의 영역으로 매핑한다. 그림 3은 세가지 질의 유형을 이차원 도메인 공간에 매핑했을 경우의 예를 나타낸다.

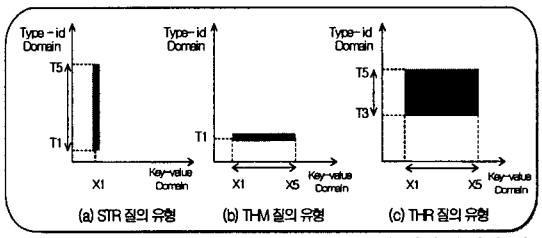

그림 3. XML 데이터베이스에 대한 세가지 질의 유형의 예

본 논문에서는 XML Schema를 따르는 XML데이터베이스내의 엘리먼트들이나 속성들의 상속을 지원하는 색인기법으로 다차원 색인구조의 하나인 MLGF[11]를 이용하여 이차원으로 구성하고 이를 2D-THI(2-Dimensional Type Hierachy Index)라 한다. 2D-THI 색인구조의 한 도메인은 색인될 엘리먼트의 키값으로 이루어지며 다른 한 도메인은 타입 상속 계층 구조에 포함된 타입들의 식별자로 이루어진다. 임의의 타입 T와 그의 서브 타입들로 구성된 타입 집합을 T^*로 표시할 때 이러한 타입 식별자들이 타입 상속 계층에 대해 preorder traversal 순서로 나열되게 타입 식별자 도메인을 구성함으로써, T^*에 포함된 타입 식별자들이 타입 식별자 도메인상에서 연속된 구간으로 표현되게 할 수 있다. 그러므로 T^*를 질의 대상으로 하는 질의처리의 색인 탐색을 이차원 도메인 공간상에서 하나의 영역 탐색으로 가능하게 한다. 그림 4는 $Vertebrata^*$에 대하여 속성 life_span의 범위(x1 < life_span < x2) 질의에 대한 색인 영역을 이차원 도메인 공간상에 표현한 것이다.

한편, 이차원 도메인 공간상에 데이터가 균일하게 분포할 경우, 도메인을 구성하는 페이지 영역들의 크기가 일정하게 되며, 주어진 질의 영역들에 의해 교차되는 페이지 영역들의 개수를 최소로 하는 페이지 영역의 최적 구간 비는 모든 질의 영역들에 대해 각 축별로 구간 크기를 더한 값의 비로서 계산할 수 있다. 따라서, 2D-THI에서는 이차원 도메인 공간상에서 임의의 위치에 주어지는 n개의 질의 영역 $q_i(x) \times q_i$ (y)(i=1, ..., n)에 대해 페이지 영역의 최적 구간비 p(x) : p(y)를 사전에 주어지는 모든 질의 영역들에 대해 각 축별로 구간 크기를 더한 값의 비인 $\sum_{i=1}^{n} q_i(x)$:

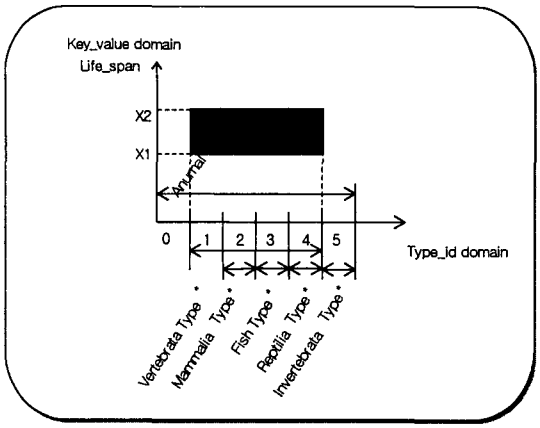

그림 4. Animal에 대한 life_span의 범위 질의(range queries) 영역

$\sum_{i=1}^{n} q_i(y)$가 되도록 한다.

2D-THI의 구조는 디렉토리와 색인페이지들로 구성된다. 디렉토리는 다단계의 균형된 트리 구조를 가지며, 디렉토리의 최하위 단계에 있는 디렉토리 레코드는 색인 페이지를 가리키는 포인터를 가진다. 그 색인 페이지가 할당된 페이지 영역을 해당 페이지 영역의 키값 도메인과 타입 식별자 도메인에 대한 구간을 리전 벡터(region vector)를 이용하여 표현한다. 색인 페이지는 디렉토리 레코드에 의해서 표현된 영역내에 속하는 색인 레코드만을 저장한다. 그리고, 다단계의 디렉토리 구조는 재귀적으로 구성된다.

2D-THI의 색인 페이지의 각 색인 레코드에는 타입 식별자 값(Type-id value) 필드, 키 값(key value) 필드, 이 두 값을 갖는 객체의 개수(No. Oids) 필드와 이들 객체에 대한 색인 엔트리(Oids)들의 리스트 필드가 있다. 그리고 레코드의 크기가 페이지의 크기보다 크게 될 경우에 할당하게 되는 오버플로어 페이지(overflow page)가 있다.

그리고 2D-THI의 연산 알고리즘은 MLGF의 삽입 연산에서 영역 분할 전략에만 차이가 있으며 다른 연산은 MLGF에서와 동일하다. 본 논문에서는 페이지 영역의 구간비가 앞에서 설명한 바와 같이 최적구간비에 근접하도록 하는 영역 분할전략을 사용한다. 이러한 영역분할 전략은 참고문헌[10]을 참고하면 된다.

4. 성능평가

본 절에서는 2D-THI로서 XML데이타베이스에 주어지는 다양한 질의 형태에 대해서 최적의 색인구조를 구성할 수 있음을 보이기 위해서 성능 비교를 위한 비용 모델을 개발하고, 이를 통하여 CH-index와 CG-tree와 성능을 비교 평가 한다. 성능 분석에 사용된 비용은 질의처리를 위해 액세스해야 할 색인 페이지의 개수로 한다.

먼저, 탐색 비용을 위한 모델링의 공정한 성능 비교를 위해서 다음과 같은 두 가지 가정들을 사용한다.

1. 모든 키 값들은 같은 길이를 지니며 각 키값에 대해서 그 값을 가지는 객체가 있는 타입들의 개수는 일정하다.
2. 키값들은 타입 상속 계층내에서 균일하게 분포한다.

비용 모델에 의한 각 색인구조의 성능 비교를 위한 실험 모델로 31개의 타입 T로 구성된 균형된 이진트리 형태로 사용하고 각 타입당 엘리먼트의 개수 E_i를 3000개를 사용한다. 그리고 동일한 키값을 갖는 객체들의 평균 개수 K로 1, 2, 5, 10을 사용하며, 각 타입이 가지는 서로 다른 키값의 개수 D_i는 3000, 1500, 900, 300으로 한정된다. 그리고 사용자 질의 패턴으로는 STR, THM, THR형태의 질의로만 주어지는 각각의 경우와 여러 질의 형태들이 혼합되어 주어지는 경우를 사용한다. 그리고, 색인 탐색 비용을 모델링하기 위해서 표1과 같은 매개변수를 사용한다.

표 1. 비용모델을 위한 매개변수

매개변수	의 미
H	단말 색인 페이지를 제외한 색인구조의 높이
HB	CH-index의 단말 색인 페이지의 블로킹 인수
GB	CG-tree의 단말 색인 페이지의 블로킹 인수
TB	2D-THI의 단말 색인 페이지의 블로킹 인수
RKN	범위 질의에 나타나는 키값의 개수
TQT	타입 계층에 대한 질의에서 질의의 대상이 되는 타입의 개수
DKN	하나의 단말 색인 페이지에 포함된 서로 다른 키값의 개수
TN	하나의 단말 색인 페이지에 포함된 타입 식별자의 개수

4.1 비용 모델

STR형태의 질의 패턴이 주어지는 경우에는 2D-THI에서는 색인 페이지의 영역분할 전략에 의해 타입 식별자 도메인을 이루는 X축에 대해서만 계속해서 분할하게 되므로 2D-THI는 CH-index에서와 같은 페이지 영역들로 구성되는 색인구조가 된다. 모든 색인구조에서 각 색인 레코드에는 하나의 키값이 할당되므로, 단말 색인 페이지에 할당되는 서로 다른 키값의 개수 DKN은 각 색인 구조의 블로킹 인수와 같게 된다. 그리고 단말 색인 페이지에 오버플로우가 발생하면 DK의 값은 1이 된다. 그리고 CH-index, CG-index, 2D-THI 색인구조의 색인 레코드의 길이를 각각 HRL GRL, DRL이라 하며 색인 페이지의 크기를 PS라 한다. STR형태의 질의에서 색인 탐색의 비용 C는 다음 식과 같이 나타낼 수 있다.

- CH-index인 경우

$$DKN = \begin{cases} HB & \text{if } HB \geq 1 \\ 1 & otherwise \end{cases}$$

$$C = \begin{cases} h + \lceil RKN/DKN \rceil & \text{if } DKN > 1 \\ h + RKN \times \lceil HRL/PS \rceil & \text{if } DKN = 1 \end{cases}$$

- CG-tree인 경우

$$DKN = \begin{cases} GB & \text{if } GB \geq 1 \\ 1 & otherwise \end{cases}$$

$$C = \begin{cases} h + \lceil RKN/DKN \rceil & \text{if } DKN > 1 \\ h + RKN \times \lceil GRL/PS \rceil & \text{if } DKN = 1 \end{cases}$$

- 2D-THI인 경우

$$DKN = \begin{cases} TB & \text{if } TB \geq 1 \\ 1 & otherwise \end{cases}$$

$$C = \begin{cases} h + \lceil RKN/UK \rceil & \text{if } DKN > 1 \\ h + RKN \times \lceil DRL/PS \rceil & \text{if } DKN = 1 \end{cases}$$

THM형태의 질의 패턴이 주어지는 경우에 2D-THI에서는 색인 페이지의 영역 분할 전략에 의해 키값 도메인을 이루는 Y축에 대해서만 계속해서 분할하게 되므로 TH-index에서 같은 페이지 영역들로 구성되는 색인구조가 된다. STR형태의 질의에서 색인 탐색의 비용 C는 다음 식과 같이 나타낼 수 있다.

- CH-index인 경우

$$C = \begin{cases} h+1 & \text{if } HB \geq 1 \\ h + \lceil HRL/PS \rceil & otherwise \end{cases}$$

- CG-tree인 경우

$$C = \begin{cases} h + TQT & \text{if } GB \geq 1 \\ h + TQT \times \lceil GRL/PS \rceil & otherwise \end{cases}$$

- 2D-THI인 경우

$$DKN = \begin{cases} BT & \text{if } BT \geq 1 \\ 1 & otherwise \end{cases}$$

$$C = \begin{cases} h + \lceil TQT/TB \rceil & \text{if } TB \geq 1 \\ h + TQT \times \lceil DRL/PS \rceil & otherwise \end{cases}$$

THR형태의 질의 패턴이 주어지는 경우에는 앞에서의 색인 페이지의 영역 분할전략에 의해서 페이지 영역들의 구간비가 TQT:RKN에 가깝게 된다. THR형태의 질의에서 색인 탐색의 비용 C는 다음 식과 같이 나타낼 수 있다.

- CH-index인 경우

$$DKN = \begin{cases} HB & \text{if } HB \geq 1 \\ 1 & otherwise \end{cases}$$

$$C = \begin{cases} h + RKN/DKN + 1 & \text{if } DKN > 1 \\ h + RKN \times \lceil HRL/PS \rceil & \text{if } DKN = 1 \end{cases}$$

- CG-tree인 경우

$$C = \begin{cases} h + (RKN/DKN + 1) \times TQT & \text{if } DKN > 1 \\ h + RKN \times \lceil GRL/PS \rceil \times TQT & \text{if } DKN = 1 \end{cases}$$

- 2D-THI인 경우

$$DKN = \begin{cases} \sqrt{\frac{RKN}{TQT} \times TB} & \text{if } TB \geq 1,\ TB > \frac{RKN}{TQT},\ TB > \frac{TQT}{RKN} \\ TB & \text{if } TB \geq 1,\ TB \leq \frac{RKN}{TQT} \\ 1 & otherwise \end{cases}$$

$$TN = \begin{cases} \sqrt{\frac{TQT}{RKN} \times TB} & \text{if } TB \geq 1,\ TB > \frac{TQT}{RKN},\ TB > \frac{RKN}{TQT} \\ TB & \text{if } TB \geq 1,\ TB \leq \frac{RKN}{TQT} \\ 1 & otherwise \end{cases}$$

$$C = \begin{cases} h + (\frac{RKN}{DKN} + 1)(\frac{TQT}{TN} + 1) & \text{if } DKN > 1,\ TN > 1 \\ h + (\frac{RKN}{DKN} + 1) \times TQT & \text{if } DKN > 1,\ TN = 1 \\ h + RKN \times (\frac{TQT}{TN} + 1) & \text{if } DKN = 1,\ TN > 1 \\ h + RKN \times TQT \times \lceil DRL/PS \rceil & \text{if } DKN = 1,\ TN = 1 \end{cases}$$

그리고, STR, THM, THR형태의 혼합 질의 패턴이 주어지는 경우 질의영역의 크기를 각각 $RKN_i \times TQT_i (i=1,...,n)$이라 하면, 2D-THI에서는 페이지 영역의 최적구간비가 모든 질의 영역의 각 축별로 구간 크기를 단순히 더한 값의 비가 되므로 $\sum_{i=1}^{n} RKN_i : \sum_{i=1}^{n} TQT_i$ 에 근접하게 된다. STR, THM, THR형태의 혼합 질의에서 색인 탐색의 비용 C는 THR형태의 질의가 주어지는 경우와 동일하다. 단, 2D-THI의 경우에 하나의 단말 색인 페이지에 있는 유일 키값의 개수 DKN과 타입 식별자 개수 TN은 다음과 같이 계산한다.

• 2D-THI인 경우의 DKN과 TN

$$DKN = \begin{cases} \frac{\sum_{i=1}^{n} RKN_i}{\sum_{i=1}^{n} TQT_i} \times TB & \text{if } TB \geq 1,\ TB > \frac{\sum_{i=1}^{n} RKN_i}{\sum_{i=1}^{n} TQT_i},\ TB > \frac{\sum_{i=1}^{n} TQT_i}{\sum_{i=1}^{n} RKN_i} \\ TB & \text{if } TB \geq 1,\ TB \leq \frac{\sum_{i=1}^{n} RKN_i}{\sum_{i=1}^{n} TQT_i} \\ 1 & otherwise \end{cases}$$

$$TN = \begin{cases} \frac{\sum_{i=1}^{n} TQT_i}{\sum_{i=1}^{n} RKN_i} \times TB & \text{if } TB \geq 1,\ TB > \frac{\sum_{i=1}^{n} TQT_i}{\sum_{i=1}^{n} RKN_i},\ TB > \frac{\sum_{i=1}^{n} RKN_i}{\sum_{i=1}^{n} TQT_i} \\ TB & \text{if } TB \geq 1,\ TB \leq \frac{\sum_{i=1}^{n} TQT_i}{\sum_{i=1}^{n} RKN_i} \\ 1 & otherwise \end{cases}$$

4.2 성능 비교 평가

다음 그림 6은 각 질의 패턴이 주어질 경우에 액세스 해야할 색인 페이지의 수를 도표로 나타낸 것이다. 그림 6의 (a)에서는 STR형태의 질의 패턴이 주어지는 경우로서 CH-index에서는 한 색인 레코드에 타입 계층의 모든 타입에서 같은 키값을 가지는 색인 엔트리들을 모두 저장하기 때문에 범위질의에 나타나는 키값의 개수에 따라 급격하게 증가함을 보인다. 그림 6의 (b)에서는 THM형태의 질의 패턴이 주어지는 경우로서 각 단말 색인 페이지에 하나의 타입에 속하는 객체들의 색인 엔트리들만 저장되어 있는 CG-tree는 타입의 개수에 따라서 급격하게 증가한다. 그리고 그림 6의 (c)에서는 THM 형태의 질의 패턴이 주어지는 경우의 성능 비교의 결과를 나타낸다. 이 경우에는 CH-index의 경우에서 질의의 대상이 되는 타입의 개수에 상관없이 질의에 주어지는 키값의 범위에 따라 일정한 값을 가지는 반면에, CG-tree에서는 질의의 대상이 되는 타입의 개수에 비례하고 키값의 범위가 커짐에 따라 증가함을 알 수 있다. 그리고 2D-THI에서는 질의의 대상이 되는 타입의 개수와 키값의 범위에 따라 영역 분할 전략에 의해서 최적의 색인구조를 구성함으로써 다른 두가지 색인구조에 비하여 항상 좋은 성능을 보임을 보인다. 마지막으로 그림 6의 (d)에서 각 색인구조에 혼합 질의 패턴이 주어지는 경우로서 두 가지 혼합 질의 패턴 모두에 2D-THI가 다른 두 색인구조에 비하여 좋은 성능을 보임을 나타낸다.

5. 결론

본 논문에서는 XML Schema에 의해 구조가 정의된 XML데이타베이스의 타입 상속 구조에 대한 색인 기법으로 다차원 색인구조를 이용하는 이차원의 타입 상속 색인 구조인 2D-THI를 제안하였다. 2D-THI는 엘리먼트 도메인과 타입 식별자 도메인으로 이루어진 이차원 도메인 공간으로 색인구조를 구성한다. 그리고 사전에 수집한 사용자 질의 정보를 바탕으로 도메인 사이에서 색인 엔트리들의 클러스터링 정도를 주어진 질의 패턴에 적합하도록 조정한다.

제안한 2D-THI의 성능을 비교 평가하기 위해서 기존의 객체지향 데이터베이스의 클래스 상속 계층에 대한 색인구조로 널리 사용되고 있는 CH-index와 CG-tree를 XML데이타베이스에 적용하여 이들과 2D-THI를 비용모델을 통해서 비교 분석하였다. 성능 분석의 결과로 XML데이타베이스에 하나의 타입을 대상으로 하는 범위질의인 STR형태의 질의가 주어지는 경우에는 각 색인 구조의 탐색 성능이 2D-THI와 CG-tree가 동일하게 CH-index보다 월등히 좋음을 보였고, 타입 계층을 대상으로 하는 부합 질의인 THM형태의 질의가 주어지는 경우에는 각 색인 구조의 탐색 성능이 2D-THI와 CH-index가 동일하게

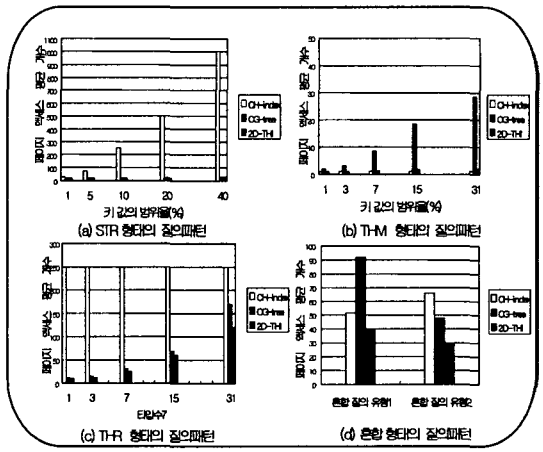

그림 6. 각 질의패턴에 대한 성능 비교

CG-tree 보다 월등히 좋음을 보였다. 그리고 타입 계층을 대상으로 하는 범위 질의인 THR이 주어지는 경우에는 2D-THI가 다른 두 색인 구조보다 좋은 성능을 보였다. 마지막으로 혼합 질의가 주어지는 경우에도 2D-THI가 제일 좋은 성능을 보였다. 따라서 2D-THI 색인구조로서 XML데이타베이스에 주어지는 다양한 질의 유형에 대해서 최적의 색인구조를 구성할 수 있음을 알 수 있다.

향후 연구로서는 데이터가 비균일하게 분포하는 경우에 실제로 색인구조들을 구축하여 실험을 통한 성능을 비교 분석하는 것이다.

참고문헌

[1] eXtensible Markup Language(XML) 1.0, http://www.w3.org/TR/1998/REC-xml-19980210.

[2] 민영수 외 5인, "XML 문서를 위한 구조정보 추출기의설계 및 구현," *한국정보과학회 '99 가을 학술발표논문집(I)*, 한국정보과학회, pp. 81-83, 1999.

[3] Migrating from XML DTD to XML Schema using UML, http://www.rational.com/produ- cts/ whitepapers/412.jsp.

[4] XML Schema Part 0:Primer, http://www.w3.org/TR/xmlschema-0/.

[5] 김정섭, 박창원, 정진완, "XML Schema를 위한 관계형 스키마 자동생성기의 개관", *한국정보과학회 2002 춘계 학술발표논문집(B)*, pp. 10-12, 2002년 4월.

[6] Kim, W. et al., "Indexing Techniques for Object-Oriented Databases," In book *Object-Oriented Concepts, Databases*, and Applications, (Kim, W and Lochovsky, F. eds.), Addison-Wesley, 1989.

[7] Kiger, C. and Moerkotte, G., "Indexing Multiple Sets," In Proc. *Int'l Conf. on Very Large Databases*, pp. 180-191, Santiago, Chile, 1994.

[8] Comer, D., "The Ubiquitous B-tree," *ACM Computing Surveys*, Vol. 11, No. 2, pp. 121-137, 1979.

[9] Robinson, J. T., "The K-D-B-Tree: A Search Structure for Large Multidimensional Dynamic Indexes," In Proc. *Int'l Conf. on Management of Data*, ACM SIGMOD, pp. 10-18, Ann Arbor, Michigan, Apr. 1981.

[10] Lee, J. H. et al., "A Region Splitting Strategy for Physical Database Design of Multidimensional File Organizations," In Proc. *Int'l Conf. on Very Large Data Bases* pp. 416-425, Athens, Greece, Aug. 1997.

[11] Whang, K. Y. and Krishnamurthy, R., Multilevel Grid Files, IBM Research Report RC 11516, IBM Thomas J. Watson Research Center, Nov. 1985.