Journal of Korea Multimedia Society Vol. 14, No. 2, February 2011(pp. 269-281)

XML 데이터베이스 다차원 타입상속 색인구조의 조율 알고리즘

이 종 학†

요 약

XML 데이터베이스에서 타입상속 개념의 질의처리를 지원하기 위한 다차원 타입상속 색인구조(Multidimensional Type Inheritance Index: MD-TIX)에 대하여, 본 논문에서는 질의 패턴에 따라 색인성능을 향상시키기 위한 색인구조의 조율 알고리즘을 제안한다. MD-TIX는 중첩 엘리먼트와 타입상속 계층이 포함된 복합 형태의 XML 질의처리를 지원하기 위하여 다차원 색인구조를 이용한다. MD-TIX의 조율 알고리즘에서는 먼저 사용자 질의 형태에 대한 질의 정보로서 색인구조를 구성할 색인 페이지 영역들의 모양을 결정하고, 이러한 모양의 페이지 영역들을 갖도록 하는 구간반분 전략을 적용함으로써 최적의 MD-TIX 색인구조를 구성한다. 성능평가의 결과에 의하면, 주어진 질의 패턴에 따라 제안한 조율 알고리즘을 적용함으로서 최적의 MD-TIX를 구성할 수 있었으며, 경로 길이가 2인 경우에 주어지는 중첩 술어에 대한 삼차원 질의 영역의 경우, 모양이 편향된 정도에 따라 질의처리의 성능이 매우 크게 향상됨을 알 수 있었다.

A Tuning Algorithm for the Multidimensional Type Inheritance Index of XML Databases

Jong-Hak Lee†

ABSTRACT

For the MD-TIX(multidimensional type inheritance index) that supports query processing for the type inheritance concept in XML databases, this paper presents an index tuning algorithm that enhances the performance of the XML query processing according to the query pattern. The MD-TIX uses a multidimensional index structure to support complex XML queries involving both nested elements and type inheritance hierarchies. In this index tuning algorithm, we first determine a shape of index page regions by using the query information about the user's query pattern, and then construct an optimal MD-TIX by applying a region splitting strategy that makes the shape of the page regions into the predetermined one. The performance evaluation results indicate that the proposed tuning algorithm builds an optimal MD-TIX by a given query pattern, and in the case of the three-dimensional query regions for the nested predicates of path length 2, the performance is much enhanced according to the skewed degree of the query region's shape.

Key words: XML Database(XML 데이터베이스), XML Schema(XML 스키마), XML Index(XML 색인)

1. 서 론

XML(eXtensible Markup Language)[1] 데이터베이스는 XML 문서를 저장하고 검색하기 위한 데이터베이스이다[2]. 이러한 XML 데이터베이스를 정의하기 위한 스키마 정의어로서 DTD(Data Type De-

※ 교신저자(Corresponding Author) : 이종학, 주소 : 경북 경산시 하양읍 금락1리 330번지(712-7022), 전화 : 053)850-2746, FAX : 053)850-2750, E-mail : jhlee11@cu.ac.kr
접수일 : 2010년 8월 9일, 수정일 : 2010년 11월 4일
완료일 : 2010년 12월 21일
† 정회원, 대구가톨릭대학교 컴퓨터정보통신공학부 교수
※ 본 연구는 2010년 대구가톨릭대학교 연구년에 의한 것임.

finition)와 XML 스키마(XML Schema)[3]가 있다. DTD는 엘리먼트(element)의 구조를 재사용할 수 없는 등 데이터 타입이 제한적으로 사용되는 단점을 가지고 있다. 따라서 타입상속(type inheritance)을 지원하는 XML 스키마가 W3C(World Wide Web Consortium)에 의해서 제안되었다. XML 스키마의 타입상속에 의해 정의된 XML 데이터베이스는 각 타입이 여러 개의 서브타입을 가질 수 있으므로 하나의 타입상속 계층을 형성한다. 따라서 XML 질의어는 이러한 데이터 모델상의 특징을 감안하여 질의의 대상을 하나의 타입 또는 특정 타입을 루트로 하는 타입상속 계층으로 지정할 수 있다.

XML로 작성된 문서를 효율적으로 저장하고 검색하기 위하여 XML 문서에 대한 질의 언어와 질의처리 등에 대한 분야가 현재 활발히 연구되고 있다. 특히 그 중 질의처리의 처리비용을 줄이기 위한 데이터베이스의 접근방법과 질의처리 최적화 기법에 관한 연구가 중요한 연구과제로 되고 있다. 데이터베이스의 색인구조는 탐색 조건에 따라 레코드들을 빠르고 효율적으로 검색하기 위하여 사용하는 데이터베이스의 접근 구조(access structure)이다. 최근에 제안된 XML 데이터베이스의 중첩 엘리먼트(nested element)[4]에 대한 색인기법은 XML 질의처리의 최적화에 크게 기여하는 것으로 보고되고 있으나[2], 이들 색인은 XML 데이터베이스가 가지는 타입상속 개념에 대한 고려를 하지 못하고 있다. 즉, 지금까지 사용되고 있는 중첩 엘리먼트에 관한 색인기법으로는 DataGuide[5], T-Index[6], APEX[7], SphinX[8] 등이 있으며, 이들은 구조 요약(structural summary)[5]이나 경로 색인(path index)[6-8]을 이용하여 주어진 경로 표현식에 대하여 XML 데이터베이스의 관련 있는 부분만은 검색할 수 있도록 하여 XML 데이터의 검색 속도를 향상시키는 색인기법으로 타입상속에 의한 질의를 고려하지 못하고 있다.

본 논문에서는 XML 데이터베이스의 중첩 엘리먼트에 대한 색인기법에서 타입상속에 의한 XML 질의를 효율적으로 처리하기 위하여 다차원 동적 파일구조를 다차원 색인구조로 이용하는 다차원 타입상속 색인기법(MD-TIX: Multidimensional Type Inheritance Indexing)[9]을 먼저 소개한다. B-tree와 같은 일차원 색인구조에서는 클러스터링 특성이 하나의 속성에 의해서 독점되는 반면에, 다차원 동적 파일구조는 다차원 클러스터링을 지원하는 파일구조로서 클러스터링의 특성이 파일을 구성하는 여러 속성들에 의해서 공유된다[10]. 다차원 동적 파일구조에 대한 연구는 지금까지 많이 진행되어 왔으며, 대표적인 예로는 hB-트리[11], 계층 그리드 파일(multilevel grid file)[12] 등이 있다.

MD-TIX에서는 중첩 엘리먼트의 킷값 도메인과 함께 경로 표현식에 나타나는 각 복합 엘리먼트마다 한 축의 타입식별자 도메인을 할당하여 구성된 다차원 도메인 공간에 주어진 색인 엔트리들의 클러스터링 문제를 다룬다. 이와 같은 색인기법에서는 기존의 일차원 색인구조를 이용한 색인기법들에서 문제가 되는 질의의 대상 범위가 타입상속 계층상의 임의의 타입들로 제한되거나, 경로 표현식에 나타나는 복합 엘리먼트의 도메인이 타입상속 계층상의 임의의 타입들로 제한이 되는 XML 질의들의 처리를 한 번의 색인 탐색으로 가능하게 할 수 있다.

그러나, 중첩 엘리먼트에 대한 색인구조로 다차원 색인구조를 단순히 이용하는 것은 중첩 엘리먼트의 킷값 도메인의 크기와 타입식별자 도메인의 크기가 매우 다르고, 주어지는 질의의 형태가 서로 다름으로 인하여 색인 검색의 성능이 매우 저하될 수 있다. 따라서 본 논문에서는 MD-TIX의 성능을 개선하기 위하여, 사용자에 의해서 주어지는 질의 패턴에 따라 색인 엔트리들이 도메인 공간상에서 최적의 클러스터링을 가능하게 하는 MD-TIX 색인구조의 조율 알고리즘을 제안한다. 제안한 조율 알고리즘에서는 질의 정보를 이용하여 중첩 엘리먼트의 킷값 도메인과 여러 개의 타입 상속 계층의 타입식별자 도메인으로 구성된 다차원 도메인 공간상에 주어진 색인 엔트리들의 클러스터링 문제를 다룬다. 이는 경로 표현식으로 표현된 중첩 술어를 가지는 XML 질의는 킷값 도메인과 여러 개의 타입식별자 도메인으로 구성된 다차원 도메인 공간상에 주어지는 다차원 질의 영역으로 매핑할 수 있기 때문이다.

MD-TIX 색인구조의 조율 알고리즘에서는 먼저, 사전에 분석한 사용자 질의 패턴에 대한 정보를 이용하여 중첩 엘리먼트의 킷값 도메인과 여러 개의 타입식별자 도메인 사이의 색인 엔트리들에 대한 도메인별 클러스터링 정도의 크기를 구한다. 그리고 이러한 도메인별 클러스터링 정도를 다차원 도메인 공간상에서 유지하도록 하는 도메인 공간의 구간반분 전략

을 적용하여 다차원 색인구조를 구성한다. 이러한 조율 알고리즘의 핵심 아이디어는 다차원 도메인 공간상에서 색인 엔트리들의 클러스터링 정도를 주어진 질의 패턴에 적합하도록 조율함으로써 주어진 질의들에 의해서 액세스되는 색인 페이지의 평균 개수를 최소화하는 것이다.

본 논문의 구성은 다음과 같다. 먼저, 제 2절에서는 관련 연구로서 XML 데이터베이스의 색인 구축에 필요한 기본 개념과 함께 기존의 색인기법들을 살펴보고, 특히 XML 데이터베이스의 타입상속 개념에 의한 질의처리를 지원하는 다차원 타입상속 색인구조(MD-TIX)를 소개한다. 그리고 제 3절에서 이 MD-TIX에 대하여 질의 패턴에 따라 색인성능을 향상시키기 위한 색인구조의 조율 알고리즘을 제안한다. 제 4절에서는 제안한 조율 알고리즘의 성능평가의 결과를 제시한다. 마지막으로 제 5절에서는 결론을 내린다.

2. 관련 연구

본 절에서는 XML 데이터베이스의 색인 구축에 필요한 XML 스키마[3], XQuery[13] 등과 같은 기본 개념과 함께 기존의 색인기법들을 살펴보고, XML 데이터베이스의 중첩 엘리먼트에 대한 색인기법으로 다차원 색인구조를 이용하는 다차원 타입상속 색인구조를 소개한다.

XML 스키마는 XML 문서의 구조를 정의하기 위하여 제안된 XML 문서 정의어이다[3]. 그림 1은 루트 엘리먼트로 사람을 가지는 사람 타입과 이에 포함된 고향 엘리먼트의 타입인 지역 타입을 XML 스키마 그래프로 표현한 예이다. XML 스키마 그래프에서 타입은 네모로, 엘리먼트는 동그라미로 나타내며, 타입들 사이의 상속관계는 화살표가 있는 점선으로 나타낸다. 그리고 엘리먼트와 타입 사이의 중첩 관계를 화살표가 있는 실선으로 나타내며, 해당 엘리먼트와 그의 타입은 일반 실선으로 나타낸다. 그림 1에서 사람 타입은 서브 타입인 사원 타입과 학생 타입, 그리고 이들의 서브 타입들을 포함하는 XML 타입상속 계층구조와 복합 엘리먼트인 고향 엘리먼트의 도메인 타입인 지역 타입을 포함하는 XML 타입 집단화 계층구조의 루트이다.

타입상속 계층에서 임의의 타입 T와 그의 모든 서브 타입들을 원소로 하는 집합을 T^*로 표기한다. 예를 들어 그림 1에서 사람* 타입은 집합 {사람 타입, 사원 타입, 학생 타입, 엔지니어 타입, 대학생 타입, 대학원생 타입, 프로그래머 타입}이고, 학생* 타입은 {학생 타입, 대학생 타입, 대학원생 타입}이다.

XQuery는 XML 데이터베이스에서의 질의어로 FLWR(FOR, LET, WHERE, RETURN) 절로 구성이 된다[13]. FOR 절은 SQL 질의의 FROM 절과 의미상으로 유사하며, LET 절은 표현을 간략하게 하기

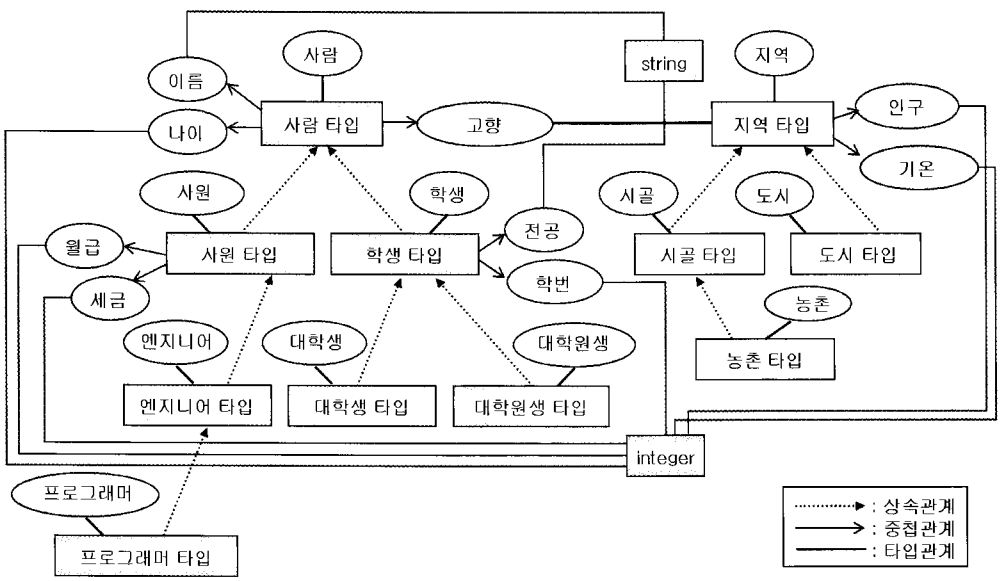

그림 1: XML 스키마 그래프의 예

위해서 복잡한 식을 변수 이름에 배치할 수 있도록 한 것이다. WHERE 절은 SQL에서의 WHERE 절과 유사하며 단순 엘리먼트에 대한 조건인 단순 술어(simple predicate)와 함께 중첩 엘리먼트에 대한 조건인 중첩 술어(nested predicate)를 사용할 수 있다. 그림 2는 그림 1에서 "기온이 평균 30도 이상인 지역이 고향인 사람들의 이름을 검색하라"는 질의를 XQuery로 작성한 예문이다.

```
FOR $h IN 사람*
WHERE $h/고향[기온 >= 30]
RETURN <name> $h/이름 </name>
```

그림 2. XQuery 예문

XPath는 중첩 엘리먼트의 경로를 표현하기 위한 경로 표현식(path expression)이다[4]. 본 논문에서는 경로 표현식에서 경로상의 엘리먼트의 타입을 타입상속 계층상의 일부 타입들로 한정하여 표현할 수 있도록 XPath를 확장하여 이를 *확장된 XPath*라 한다. 확장된 XPath는 각 엘리먼트 다음에 타입의 이름이 올 수 있도록 확장한 것으로 다음과 같은 형태를 가진다. 단, E_i 뒤의 중괄호{ }는 선택적임을 나타내는 표시이다.

$$EP=T_1/E_1\{(T_2)\}/E_2\{(T_3)\}/\ldots/E_n\{(T_{n+1})\} \quad (1)$$

경로 *EP*에서 타입 T_1을 *타겟타입*, T_{i+1}을 엘리먼트 E_i의 *도메인타입*이라 정의한다. 타겟타입과 도메인타입은 경로에서 타입상속 계층구조에 속하는 특정 타입으로 한정(limit)될 수 있으며, 이를 타입 대치(type substitution)라 한다. 이러한 타입 대치는 질의의 범위를 특정 타입으로 한정할 수 있도록 하여 타입상속의 개념을 XML 질의에 표현하도록 한 것이다. 다음 중첩 술어들은 그림 2의 질의로부터 확장된 XPath로 표현된 타입 대치에 대한 예를 보여주고 있다.

Pn1: 사원*/고향[기온 >= 30]
Pn2: 사원/고향[기온 >= 30]
Pn3: 사원/고향(시골)[기온 >= 30]

중첩 술어 *Pn1*은 질의 대상을 *사람 타입*의 타입상속 계층에서 사원*, 즉 사원 타입, 엔지니어 타입, 프로그래머 타입에 속하는 엘리먼트들로 한정하는 조건식이며, *Pn2*는 질의 대상을 사원 타입만으로 한정하는 조건식이다. 그리고 *Pn3*는 *Pn2*에서 복합 엘리먼트 고향의 타입을 시골 타입만으로 한정하는 조건식이다.

확장된 XPath식 *EP*에서 *경로 인스턴스(path instance)*는 다음 조건을 만족하는 엘리먼트들의 리스트(E_1, E_2, ..., E_{n+1})로 정의한다. (1) 엘리먼트 E_1은 타입 T_1의 엘리먼트이다. (2) 엘리먼트 E_i ($1 < i <= n+1$)는 타입 T_i의 엘리먼트로서 엘리먼트 E_{i-1}의 구성 엘리먼트이다.

XML 질의는 전형적으로 경로 질의와 가지(twig) 질의라는 두 가지 형태의 구조적 질의로 나눌 수 있다[14]. 경로 질의는 XML 스카마에서 지금까지 소개한 단지 하나의 술어로 구성된 질의이며, 가지 질의는 두 개 이상의 술어로 구성된 가지 패턴[15]으로 나타나는 질의이다. 지금까지 제안된 XML 데이터베이스의 경로 질의에 대한 색인기법으로는, DataGuide [5], T-Index[6], APEX[7], SphinX[8] 등이 있다. DataGuide는 경로 요약 기법으로 비결정적(non-deterministic) 오토마타를 결정적(deterministic) 오토마타로 변환하는 과정과 동일한 과정으로 경로를 색인하는 기법이다. 일반적으로 비결정적 오토마타를 결정적 오토마타로 바꿀 경우, 크기가 커지게 되지만, XML 문서 내에 동일한 경로들이 많이 존재할수록 색인의 크기는 줄어든다. T-index는 루트로부터 시작되는 경로의 집합이 동일한 노드들을 모아 색인을 구축하는 기법으로서, DataGuide와 마찬가지로 XML 문서 내에 동일한 경로가 매우 많이 존재한다는 점을 이용하는 색인기법이다. 그리고 SphinX [8]는 XML 스키마 정의어로 엘리먼트의 구조를 재사용할 수 없는 등 데이터 타입이 제한적으로 사용되는 단점을 가지고 있는 DTD를 기반으로 하는 인덱스 기법이다.

그리고, 가지 질의의 처리를 위한 많은 색인기법들은 하나의 가지 패턴을 여러 개의 선형 경로(linear path)들로 분할 처리하는 방법에 기반하고 있다. 여기서는 우선 각 선형 경로 패턴에 부합하는 인스턴스들을 기존의 경로 질의에 대한 색인기법으로 검색하고, 이들을 조인함으로써 최종 결과를 얻는다. 이때 조인 과정에서 불필요한 중간 결과들이 많이 생성되면, 질의처리 성능을 크게 떨어뜨린다. 이를 해결하기 위해 TwigStack[15]을 시초로 많은 홀리스틱 가지 패턴 조인(holistic twig join) 기법들[14,16,17]이

제시되었다. 특히 TwigX-Guide[14]는 TwigStack에서 가지는 필요 없는 중간 조인결과의 생성을 줄이기 위하여 TwigStack에 있는 영역 엔코딩 기법에 DataGuide에 있는 경로 요약 기법을 추가로 확장 적용하였다. 본 논문에서는 이러한 가지 질의에 대한 처리가 아닌, 단지 하나의 경로 질의로 한정하여 경로 상의 타입상속에 의한 타입 대치에 대한 색인기법을 다룬다. 그러나 이 색인기법을 가지 질의의 처리를 위해 확대 적용할 수 있다.

앞에서 언급한 경로 질의에 대한 여러 색인기법들은 B-tree와 같은 일차원 색인구조를 이용한다. 따라서 XML 데이터 모델의 타입상속의 특징을 반영하지 못한다. 즉, 질의의 대상 범위가 타입상속 계층상의 임의의 타입들로 제한되거나, XPath식에 나타나는 어떠한 복합 엘리먼트의 도메인 타입이 타입상속 계층상의 임의의 타입들로 제한이 되는 질의들을 지원할 수 없다. 따라서 이러한 중첩 술어의 처리를 지원하기 위한 색인구조로 다차원 색인구조를 이용할 수 있다. 즉, 색인할 중첩 엘리먼트의 킷값 도메인과 함께 중첩 엘리먼트를 표현하는 경로상의 각 타입상속 계층마다 타입식별자들로 구성된 한 차원씩의 타입식별자 도메인을 할당함으로써, (*경로길이 +1*)차원의 도메인 공간을 구성하여 이를 다차원 색인구조에 적용한다. 예를 들어, 그림 1과 같은 스키마 그래프에서 경로의 길이가 2인 사람타입의 중첩 엘리먼트 인구(지역 타입의 단순 엘리먼트)에 대한 색인구조로서 X축은 중첩 엘리먼트 인구의 킷값 도메인으로 하고, Y축은 사람 타입상속 계층의 타입식별자 도메인으로 하고, 그리고 Z축은 지역 타입상속 계층의 타입식별자 도메인으로 하는 삼차원 도메인 공간을 구성할 수 있다.

참고문헌[9]에서는 XML 데이터베이스의 중첩 엘리먼트에 대한 색인구조를 다차원 색인구조의 하나인 계층 그리드 화일(multilevel grid file: MLGF)[12]을 이용하여 구성하고, 이를 *다차원 타입상속 색인기법(Multidimensional Type Inheritance Indexing: MD-TIX)*이라 하였다. MD-TIX는 디렉토리와 색인 페이지로 구성된다. 디렉토리는 다단계의 균형된 트리 구조를 가지며, 디렉토리를 구성하는 디렉토리 페이지의 구조는 MLGF[12]에서와 마찬가지이다. 색인 페이지는 색인 레코드들로 구성되며, 각 색인 레코드에는 경로상의 각 타입식별자 값(type-id value) 필드, 킷값(key value) 필드, 엘리먼트 또는 경로의 개수 필드, 및 이들에 대한 색인 엔트리들의 리스트 필드가 있다. 그리고 레코드의 크기가 페이지의 크기보다 크게 될 때 오버플로우 페이지를 할당하고 이를 포인트하기 위한 오버플로우 페이지(overflow page) 필드가 있다. 그림 3은 MD-TIX 색인구조의 색인 페이지 구조를 나타낸다.

그러나, 중첩 엘리먼트에 대한 색인구조로 다차원 색인구조를 단순히 이용하는 것은 중첩 엘리먼트의 킷값 도메인의 크기와 타입식별자 도메인들 사이의 크기가 매우 다르고, 또한 이에 따른 질의의 형태가 특정 도메인에 편향되게 주어짐으로 인하여 색인 검색의 성능이 저하될 수 있다. 따라서 본 논문에서는 MD-TIX의 색인성능을 개선하기 위하여, 사용자에 의해서 주어지는 질의 패턴에 의한 질의정보를 이용하여 MD-TIX 색인구조의 성능을 최적으로 보장할 수 있는 조율 알고리즘을 제안한다.

중첩 엘리먼트에 대한 색인기법은 중첩 술어를 만족하는 객체들의 탐색에는 매우 유용하지만, 상대적으로 경로의 길이가 길게 되면 색인구조의 유지비용을 많이 필요로 한다[9]. 따라서 경로의 길이가 4이상일 경우에는 경로를 길이가 1, 2, 또는 3이 되는 서브경로들로 분할한 다음에, 각 서브경로별로 색인구조를 할당하여야 한다라고 하였다. 본 논문에서는 경로의 길이가 2 또는 3인 경우의 삼차원 또는 사차원 MD-TIX에 한정하여 색인구조의 조율 알고리즘을 논의하고 이를 N차원으로 일반화한다.

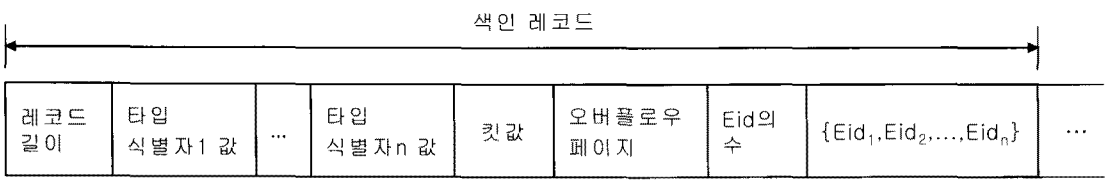

그림 3. MD-TIX 색인구조의 색인 페이지 구조

3. MD-TIX의 조율 알고리즘

본 절에서는 XML 질의에서 사용되는 중첩 술어들이 다차원 도메인 공간상에 매핑되는 질의 영역들의 형태에 대한 정보를 기반으로, 질의 영역들에 의해 교차하는 색인 페이지 영역들의 개수가 최소로 되는 *페이지 영역의 최적 구간비*를 결정하고, 가능한 이와 같은 구간비를 갖는 페이지 영역들이 되도록 하는 구간반분 전략을 사용함으로써 최적의 MD-TIX를 구성하는 조율 알고리즘을 제안한다.

3.1 MD-TIX 페이지 영역의 최적 구간비

다차원 색인구조에서는 다차원 도메인 공간에 주어진 색인 페이지 영역의 구간비에 따라 질의 영역에 의해서 교차되는 페이지 영역의 평균 개수가 달라지는 특징이 있다. 참고문헌[10]에서는 이러한 특징을 이용하여 데이터의 균일 분포와 비균일 분포 각각에 대하여 주어진 질의 영역들에 대해 페이지 영역의 평균 액세스 횟수를 최소로 하는 페이지 영역의 최적 구간비를 계산하는 방법을 제안하였다. 본 절에서는 이를 소개하고, MD-TIX 색인구조를 구성하는 페이지 영역들의 최적 구간비를 이와 같은 방법으로 계산한다.

다음은 이차원 색인구조의 도메인 공간상에서 데이터가 균일하게 분포할 때 페이지 영역의 최적 구간비를 계산하는 방법이다. 데이터가 균일하게 분포하면 도메인을 구성하는 페이지 영역들의 크기가 일정하게 되며, 주어진 질의 영역들에 의해 교차되는 페이지 영역들의 개수를 최소로 하는 페이지 영역의 최적 구간비는 모든 질의 영역들에 대해 각 축별로 구간 크기를 더한 값의 비로서 계산할 수 있다. 즉, 크기가 $p(x) \times p(y)$로 일정한 페이지 영역들로 나누어져 있는 이차원 도메인 공간상에서, 임의의 위치에 주어지는 n개의 질의 영역 $q_i(x) \times q_i(y)(i = 1, ..., n)$에 대해 각 질의 영역과 만나게 되는 페이지 영역의 총 개수를 최소로 하는 페이지 영역의 최적 구간비 $(O(x) : O(y))$는 $\sum_{i=1}^{n} q_i(x) : \sum_{i=1}^{n} q_i(y)$이다.

다차원 색인구조의 도메인 공간상에서 데이터가 비균일하게 분포한다는 것은 도메인 공간의 위치에 따라 색인 엔트리의 밀집도가 다름으로 인하여 페이지 영역의 크기가 위치에 따라 달라짐을 의미한다. 즉, 밀집도가 높은 곳에서는 밀집도가 낮은 곳에 비하여 많은 페이지가 할당되므로 각 페이지 영역의 크기는 작아지게 된다. 따라서 비균일 분포의 경우에는 질의 영역에 의해 교차되는 페이지 영역의 개수는 질의 영역의 크기뿐만 아니라 질의 영역이 주어진 위치의 데이터 밀집도에도 비례하게 되므로, 균일 분포에서와 같이 페이지 영역의 최적 구간비를 모든 질의 영역의 각 축별로 구간 크기를 단순히 더한 값의 비로서 구할 수 없다.

이와 같은 경우에는 각 질의 영역의 크기에 대해 위치에 따른 데이터 밀집도를 가중치(weight)로 곱한 질의 영역의 형태를 *정규화된 질의 영역*(*normalized query region*)이라 하고, 이러한 질의 영역의 정규화를 통하여 페이지 영역의 최적 구간비를 계산할 수 있다. 즉, 서로 다른 크기의 페이지 영역들로 나누어져 있는 이차원 도메인 공간상에서, 임의의 위치에 주어지는 n개의 질의 영역 $q_i(x) \times q_i(y)(i = 1, ..., n)$에 대해 각 질의 영역의 데이터 밀집도를 $d_i (= nr_i / q_i(x) \times q_i(y)$, 단, nr_i는 질의 영역 내의 레코드 수이다)라 할 때, 각 질의 영역과 만나게 되는 페이지 영역의 총 개수를 최소로 하는 페이지 영역의 최적 구간비$(O(x) : O(y))$는 $\sum_{i=1}^{n} q_i(x)\sqrt{d_i} : \sum_{i=1}^{n} q_i(y)\sqrt{d_i}$로 하면 된다.

따라서, 본 논문에서는 이차원에서 다차원으로 확장하여 최적의 MD-TIX 색인구조를 구성한다. 즉, 경로의 길이가 2인 경우에 적용할 삼차원 MD-TIX 색인구조인 경우에는 X, Y, 그리고 Z축으로 구성된 n개의 삼차원 질의 영역 $q_i(x) \times q_i(y) \times q_i(z)(i = 1, ..., n)$에 대해 색인 페이지 영역의 최적 구간비$(O(x):O(y):O(z))$를 $\sum_{i=1}^{n} q_i(x)d_i^{1/3} : \sum_{i=1}^{n} q_i(y)d_i^{1/3} : \sum_{i=1}^{n} q_i(z)d_i^{1/3}$로 계산한다. 그리고 경로의 길이가 3인 경우에 적용할 사차원 MD-TIX 색인구조인 경우에는 W, X, Y, 그리고 Z축으로 구성된 n개의 사차원 질의 영역 $q_i(w) \times q_i(x) \times q_i(y) \times q_i(z)(i = 1, ..., n)$에 대해 색인 페이지 영역의 최적 구간비$(O(w):O(x):O(y):O(z))$를 $\sum_{i=1}^{n} q_i(w)d_i^{1/4} : \sum_{i=1}^{n} q_i(x)d_i^{1/4} : \sum_{i=1}^{n} q_i(y)d_i^{1/4} : \sum_{i=1}^{n} q_i(z)d_i^{1/4}$로 계산한다. 다음 제 3.2절에서는 이렇게 계산된 색인 페이지 영역의 최적 구간비를 갖는 MD-TIX를 구성하기 위한 색인 구조의 구간반분 전략을 제시한다.

3.2 MD-TIX의 구간반분 전략

MD-TIX의 삽입, 삭제, 및 검색과 관련된 조작 연산의 구체적인 알고리즘은 참고문헌[12]에 기술된 MLGF의 조작 알고리즘과 거의 동일하나, 단지 삽입 연산의 구간반분 전략에서 차이가 있다. 따라서 본 절에서는 페이지 영역의 구간비가 제 3.1절에서 기술한 방법에 의하여 계산되는 페이지 영역의 최적 구간비에 근접하도록 하는 구간반분 전략을 제시한다.

MD-TIX에서는 색인 엔트리가 삽입되고 삭제되는 상황에 따라 분할과 병합을 반복함으로써 동적 변화에 적응한다[12]. 새로운 색인 엔트리가 삽입되는 경우, 다단계의 디렉토리를 루트로부터 최하위 디렉토리까지 탐색하여 그 색인 엔트리가 속하는 페이지 영역을 찾게 되고, 그 영역에 할당된 색인 페이지에 색인 엔트리를 삽입하게 된다. 삽입 결과 색인 페이지의 용량이 초과되면(overflow), 해당 영역은 같은 크기를 갖는 두 개의 영역으로 분할(half splitting)되고 각 영역에 해당하는 새로운 두 개의 색인 페이지가 할당되며, 색인 레코드들은 두 색인 페이지에 분산된다.

지금까지의 MD-TIX에서는 하나의 페이지 영역을 두 개의 영역으로 분할할 때 각 축을 번갈아 가며 분할시키는 순환분할 전략을 사용하고 있다. 그러나 MD-TIX는 페이지 영역을 분할할 때 분할 축을 임의로 선택할 수 있으며, 분할 축을 선택하는 방법에 따라 페이지 영역의 모양을 결정할 수 있다. 따라서 본 논문에서는 분할 축으로서 분할된 후의 페이지 영역의 구간비가 최적 구간비에 가깝게 되는 축을 선택함으로써, 엘리먼트의 지속적인 삽입으로 인한 연속된 분할시에 도메인 공간내의 모든 페이지 영역의 구간비를 최적 구간비에 근접하도록 조율할 수 있다.

아래 정리 1은 특정 모양의 질의 영역이 이차원 도메인 공간상의 임의의 위치에 주어질 때, 특정 크기의 한 페이지 영역과 만나게 되는 위치 영역의 크기는, 그 페이지 영역의 모양이 주어진 질의 영역의 모양과 같을 때 최소가 됨을 나타낸다.

정리 1 구간비가 $q_x : q_y$ 인 $q_x \times q_y$ 형태의 질의 영역 QR이 이차원 도메인 공간상의 임의의 위치에 주어질 때, 크기가 S인 $p(x) \times p(y)$ 형태의 한 페이지 영역 PR과 만나게 되는 위치 영역의 크기가 최소로 되는 경우는 PR의 구간비($p(x) : p(y)$)가 주어진 QR의 구간비($q_x : q_y$)와 같을 때이다.

증명: 아래 그림 4는 $q_x \times q_y$ 형태의 질의 영역 QR이 이차원 도메인 공간상의 임의의 위치에 주어질 때, 크기가 $S(=p(x)\times p(y))$인 특정 페이지 영역 PR과 만나게 되는 QR의 위치를 질의 영역 QR의 우하점이 위치하는 영역(음영 부분) QLR로 나타낸 것이다.

그림 4에서 QLR의 크기 $S_QLR(p(x),p(y))$는 다음 식과 같다.

$$S\_QLR(p(x),p(y)) = (p(x)+q_x)(p(y)+q_y) \quad (2)$$

$p(x)\times p(y) = S$ 이므로, 수식(2)의 $p(y)$를 $\frac{S}{p(x)}$로 치환하면,

$$S\_QLR(p(x),\frac{S}{p(x)}) = (p(x)+q_x)(\frac{S}{p(x)}+q_y)$$
$$= S + \frac{q_x S}{p(x)} + p(x)q_y + q_x q_y \quad (3)$$

이다. 따라서 수식(3)의 값을 최소로 하는 $p(x)$를 구하면, $p(x) = \sqrt{(q_x/q_y)S}$ 이다. 또한, 이러한 $p(x)$에 대한 $p(y)$는 $p(x)\times p(y) = S$에 의하여 $p(y) = \sqrt{(q_y/q_x)S}$ 이다. 그러므로, $S_QLR(p(x), p(y))$를 최소로 하는 페이지 영역 PR의 구간비 $p(x) : p(y) = q_x : q_y$ 이다. □

정리 1을 삼차원으로 확장하여 이용하면, 삼차원 MD-TIX의 경우 페이지 영역을 분할할 때 분할된 페이지 영역의 구간비가 최적 구간비에 가깝게 되는 분할 축을 선택할 수 있다. 즉, X, Y, Z축으로 이루진 삼차원 MD-MAI의 경우 페이지 영역의 분할 시 분할된 페이지 영역의 구간비가 제 3.1절에서와 같이 계산된 최적 구간비에 가장 가깝게 되는 분할 축을 선택하는 방법은 다음과 같다. 먼저, 계산된 색인 페이지 영역의 최적 구간비(O(x) : O(y) : O(z))를 갖는

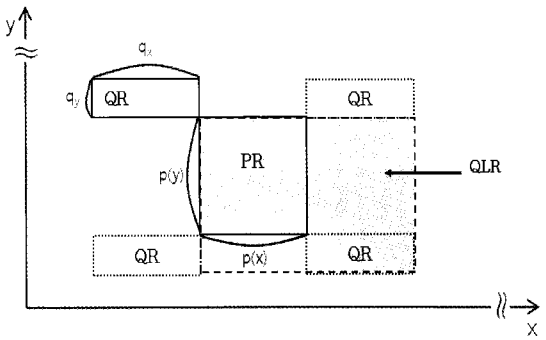

그림 4. 임의의 한 페이지 영역과 만나게 되는 질의 영역의 위치 영역

$O(x) \times O(y) \times O(z)$형태의 질의 영역이 삼차원 도메인 공간상에 임의의 위치에 주어진다고 가정하고, 분할이 요구되는$p(x)\times p(y)\times p(z)$형태의 페이지 영역이 각 축에 대해 분할된 후의 한 페이지 영역과 만나게 되는 질의 영역의 위치 영역 QLR의 크기를 계산한다. 그리고, 이 QLR 크기의 값이 가장 작게 되는 축을 분할 축으로 선택한다. 예를 들어, 분할 축으로 X축을 선택했을 때 QLR의 크기는 $(p(x)/2+O(x))(p(y)+O(y))(p(z)+O(z))$이다.

3.3 MD-TIX의 조율 알고리즘

다음은 제 3.1절에서의 다차원 MD-TIX의 최적 조건과 제 3.2절에서의 구간반분 전략을 이용하여 경로의 길이가 $N-1$인 중첩 엘리먼트에 대한 N차원의 MD-TIX로 일반화한 색인구조에 대한 조율 알고리즘을 나타낸다.

알고리즘 1 경로의 길이가 N-1인 중첩 엘리먼트에 대한 N차원 MD-TIX의 조율 알고리즘

• 조율 정보

한 개의 색인된 중첩 엘리먼트의 킷값 도메인과 N-1개의 타입식별자 도메인으로 구성된 N차원 도메인 공간상에 주어진 n개의 질의 영역에 대하여,

(1) 각 질의 영역의 형태: $q_i(1)\times q_i(2)\cdots q_i(j)\cdots \times q_i(N)(i=1,\cdots,n)$

(2) 각 질의 영역에 포함되는 엘리먼트 개수: $NE_i\,(i=1,\cdots,n)$

• 조율 알고리즘

단계 1: 각 질의 영역의 정규화$(i=1,\cdots,n)$

(1) 각 질의 영역의 엘리먼트 밀집도 d_i를 계산한다.

$$d_i = \frac{NE_i}{q_i(1)\times q_i(2)\ldots q_i(j)\ldots \times q_i(N)}$$

(2) 밀집도 d_i로서 각 질의 영역의 정규화된 질의 영역의 형태를 구한다.

$$q_i'(1) = q_i(1)\times d_i^{1/N}$$
$$q_i'(2) = q_i(2)\times d_i^{1/N}$$
$$\vdots$$
$$q_i'(j) = q_i(j)\times d_{ii}^{1/N}$$
$$\vdots$$
$$q_i'(N) = q_i(N)\times d_{ii}^{1/N}$$

단계 2: 색인 페이지 영역의 최적 구간비(O(1):O(2)$\cdots$:O(j)$\cdots$:O(N))의 결정

$$O(1):O(2)\cdots:O(j)\cdots:O(N) = \sum_{i=1}^{n} q_i'(1) : \sum_{i=1}^{n} q_i'(2) \cdots : \sum_{i=1}^{n} q_i'(j) \cdots : \sum_{i=1}^{n} q_i'(N)$$

단계 3: 최적의 MD-TIX 구축

(1) 엘리먼트의 삽입으로 색인 페이지에 오버플로우가 발생하면, 색인 페이지 분할

⇒ 대응하는 페이지 영역 $(p(1)\times p(2)\ldots p(j)\ldots \times p(N))$의 구간반분 전략:

다음 식들의 값들 중 최소가 되는 차례의 축을 분할 축으로 선택

- 첫 번째: $(p(1)/2+O(1))(p(2)+O(2))\cdots(p(j)+O(j))\cdots(p(N)+O(N))$
- 두 번째: $(p(1)+O(1))(p(2)/2+O(2))\cdots(p(j)+O(j))\cdots(p(N)+O(N))$

⋮

- j 번째: $(p(1)+O(1))(p(2)+O(2))\cdots(p(j)/2+O(j))\cdots(p(N)+O(N))$

⋮

- N 번째: $(p(1)+O(1))(p(2)+O(2))\cdots(p(j)+O(j))\cdots(p(N)/2+O(N))$

(2) 연속된 엘리먼트의 삽입에 따라 (1)번 항목의 반복 적용 □

알고리즘 1에서 N차원 MD-TIX의 조율과정은 다음과 같은 세 가지 단계로 구성된다. 첫째, 질의 패턴으로 주어진 각 질의 영역 즉, 색인된 중첩 엘리먼트의 킷값 도메인과 N-1개의 타입식별자 도메인으로 구성된 N차원의 각 질의 영역에 대해서 정규화 과정을 거친다. 정규화 과정은 제 3.1절에서 언급한 바와 같이 각 질의 영역의 크기뿐만 아니라, 각 질의 영역이 주어진 도메인 공간상의 색인엔트리 밀집도를 색인 페이지 영역의 모양을 결정하는데 반영하기 위함이다. 알고리즘에서는 정규화 과정을 제 3.1절에서 언급한 삼차원, 사차원의 경우를 N차원으로 일반화하였다.

둘째, 정규화된 모든 질의 영역에 대해서 제 3.1절에서 언급한 바와 같이 각 축별 구간의 크기를 합산

한 값의 비율로서 색인 페이지 영역의 최적 구간비 (*O(1):O(2)…:O(j)…:O(N)*)를 얻는다.

셋째, 최적 구간비에 최대한 가까운 모양의 페이지 영역으로 구성된 MD-TIX를 구축한다. 여기서는 제 3.2절의 삼차원 MD-TIX의 구간반분 전략을 N차원으로 일반화하여 적용하였다. 즉, 계속되는 엘리먼트의 삽입으로 MD-TIX의 색인 페이지에 오버플로우가 발생하면, 이 색인 페이지에 대응하는 페이지 영역은 구간반분 전략을 사용하여 같은 크기의 두 영역으로 분할되고, 원 색인 페이지의 색인 레코드들은 분할된 페이지 영역에 대응하는 두 개의 색인 페이지로 나뉘어 저장된다. 이때 페이지 영역의 구간반분 전략으로 제 3.2절의 구간반분 전략을 적용하는 것이다. 즉, 둘째 단계에서 결정된 최적 구간비 *O(1):O(2)…:O(j)…:O(N)*과 같은 모양을 가지는 가상의 질의 영역 *O(1)×O(2)…×O(j)…×O(N)*이 임의의 위치에 주어진다고 가정하고, 분할이 요구되는 페이지 영역(p(1)×p(2) …×p(j) …×p(N))이 각 축에 대해 구간반분에 의한 분할 후의 한 페이지 영역과 만나게 되는 질의 영역의 위치 영역의 크기(예를 들어, j번째 축의 구간을 이등분 했을 때 그 크기는 (p(1)+O(1)) (p(2)+O(2)) … (p(j)/2+O(j)) … (p(N)+O(N)이다)를 각각 계산한 다음 그 값이 가장 작게 되는 축을 분할 축으로 선택한다.

4. 성능평가

본 절에서는 MD-TIX의 조율 알고리즘에 대한 타당성과 유용성을 다양한 실험을 통하여 평가한다. 실험의 목적은 MD-TIX를 구성하는 색인 페이지 영역의 모양에 대한 XML 질의의 질의 영역별 색인 검색의 성능을 알아보고, 주어진 질의 정보로서 최적의 MD-TIX를 구성할 수 있음을 실제 실험을 통하여 검증하는 것이다. 본 성능평가에서 사용된 비용은 질의처리를 위하여 액세스해야 할 색인 페이지의 개수로 한다. 제 4.1절에서는 성능평가를 위하여 사용된 실험 환경에 대하여 기술하고, 제 4.2절에서는 실험 결과를 제시하고 이를 비교분석한다.

4.1 실험 환경

본 실험에서 구현한 MD-TIX 색인구조의 조율 알고리즘은 C++ 버전 4.3으로 단일 스레드 실행 코드로 구현 되었다. 모든 실험은 Intel Core2 Duo 2Ghz와 5,400 RPM의 2.5′ HDD, Fedora 리눅스 10이 설치된 시스템에서 수행되었다. 그리고 실험의 간결성을 위하여 이차원과 삼차원에 한정하는 이차원 MD-TIX와 삼차원 MD-TIX인 두 종류의 색인구조를 구축한다. 하나는 중첩 엘리먼트의 킷값 도메인인 X축과 타겟 타입식별자 도메인인 Y축으로 구성된 이차원 MD-TIX이고, 다른 하나는 이차원 MD-TIX에 도메인 타입식별자 도메인인 Z축을 하나 더 할당하여 구성한 삼차원 MD-TIX이다. 그리고 색인구조의 구축을 위하여 사용한 각 색인구조의 구성 요소들의 값으로 일반적인 구현에서 널리 사용되고 있는 다음과 같은 값들을 사용한다. 엘리먼트 식별자 Eid의 크기는 6바이트, 타입식별자의 크기는 4바이트, 색인된 키의 크기는 10바이트, 포인터는 4바이트, 각종 길이와 개수 필드의 크기는 2바이트, 디렉토리 페이지의 리젼 벡터의 크기는 12바이트, 그리고 색인 페이지의 크기는 4K바이트로 한다.

그리고, 실험을 위한 데이터베이스를 다음과 같이 구성한다. 먼저, 타겟 타입식별자 도메인과 도메인 타입식별자 도메인의 구성을 위하여 사용한 타입상속 계층구조로서 63개의 타입들(T_1, T_2, …, T_{63})로 구성된 균형된 이진 트리 형태를 사용한다. 이런 경우 타입식별자 도메인의 구간 크기는 타입 집합 T_i^*에 속하는 원소의 개수로 1, 3, 7, 15, 31, 63인 6가지가 가능하다. 그리고 각 타입에 대해 2000개의 킷값을 [0, 1000]의 구간내에서 표준 편차σ가 1000 × 2/5인 $N(\mu,\sigma^2)$의 정규 분포를 취하게 하여 평균값 μ를 임의로 조정함으로써, 타입에 따라 색인 엔트리들이 다차원 도메인 공간상에서 집중되는 위치가 다르게 한다.

질의 패턴의 구성을 위하여 사용한 질의 영역들의 형태는 이차원 질의 영역인 경우에는 크기가 도메인 공간의 1/1000로서 일정하고, 질의 영역의 구간비가 64:1, 16:1, 4:1, 1:1, 1:4, 및 1:16인 Q_64:1, Q_16:1, Q_4:1, Q_1:1, Q_1:4, 및 Q_1:16 형태의 질의 영역 등이다. 그리고 삼차원 질의 영역인 경우에는 크기가 도메인 공간의 1/10000로 일정하고, 질의 영역의 구간비가 1:1:1, 1:2:4, 1:4:16, 1:8:64, 및 1:16:256인 Q_1:1:1, Q_1:2:4, Q_1:4:16, Q_1:8:64, 및 Q_1:16:256 형태의 질의 영역 등이다.

4.2 실험 결과

첫 번째 실험에서는 이차원 MD-TIX 색인구조에

대해서, 다양한 타겟 페이지 영역의 구간비를 갖는 여러 개의 MD-TIX를 생성하고, 각각에 대하여 고유의 질의 영역 형태를 갖는 각 질의를 처리할 때 발생하는 평균 페이지 액세스 수를 측정하였다. 그림 5는 첫 번째 실험의 결과를 나타낸다. 가로축은 구성된 이차원 MD-TIX의 타겟 페이지 영역의 구간비를 나타내고, 세로축은 질의처리 시 발생되는 색인 페이지의 액세스 수를 나타낸다. 그림 5에서 알 수 있듯이, 모든 형태의 질의 영역에 대하여 그 질의 영역의 구간비를 타겟 페이지 영역의 구간비로 가지는 MD-TIX에서 가장 좋은 성능을 보였다. 질의 영역의 구간비가 16:1로 주어진 경우 타겟 페이지 영역의 구간비가 16:1(최적 구간비)인 MD-TIX에서 (평균 4개의 색인 페이지를 액세스) 타겟 페이지 영역의 구간비가 1:1로 구성된 MD-TIX에 (평균 10개의 색인 페이지를 액세스) 비해서는 2.5배까지의 성능 향상을 보였으며, 타겟 페이지 영역의 구간비가 1:16으로 구성된 MD-TIX에 (평균 21개의 색인 페이지를 액세스) 비해서는 5배까지 성능이 좋았다. 이것은 이차원 MD-TIX의 경우 반드시 주어진 질의 형태에 따라 색인구조를 구성하는 색인페이지의 모양을 조율할 수 있어야함을 보여주는 것이다.

두 번째 실험에서는 삼차원 MD-TIX 색인구조에 대해서, 첫 번째 실험에서와 같은 실험을 실시하였다. 그림 6은 두 번째 실험의 결과를 나타낸다. 그림 6에서 알 수 있듯이, 삼차원 MD-TIX에 대해서도 모든 형태의 질의 영역에 대하여 그 질의 영역의 구간비를 타겟 페이지 영역의 구간비로 가지는 MD-TIX에서 가장 좋은 성능을 보인다. 질의 영역의 구간비가 1:16:256으로 주어진 경우 타겟 페이지 영역의 구간비가 1:16:256(최적 구간비)인 MD-TIX에서 (평균 5개의 색인 페이지를 액세스) 타겟 페이지 영역의 구간비가 1:1:1로 구성된 MD-TIX에 (평균 25.5개의 색인 페이지를 액세스) 비해서 5.1배까지의 성능향상을 보인다. 이것은 삼차원 MD-TIX의 경우에도 반드시 주어진 질의 형태에 따라 색인구조를 구성하는 색인페이지의 모양을 조율할 수 있어야함을 보여주는 것이다.

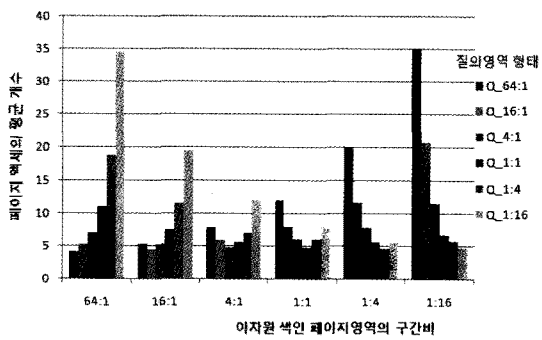

그림 5. 서로 다른 모양의 색인 페이지 영역으로 구성된 이차원 MD-TIX들에 대한 질의 형태별 성능비교

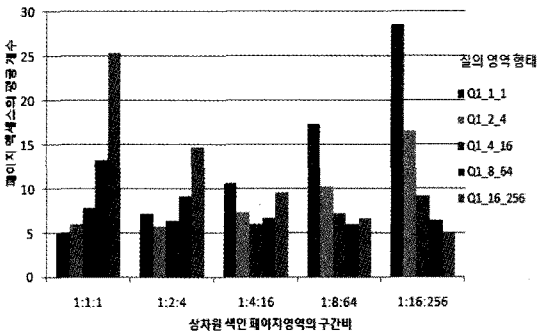

그림 6. 서로 다른 모양의 색인 페이지 영역으로 구성된 삼차원 MD-TIX들에 대한 질의 형태별 성능비교

세 번째 실험에서는 서로 다른 타겟 페이지 영역의 구간비를 갖는 여러 개의 삼차원MD-TIX 색인구조에 대하여, 여러 가지 형태의 질의 영역들이 혼합되어 주어지는 하나의 혼합 질의 패턴을 처리하기 위한 평균 색인 페이지 액세스 수를 측정하였다. 혼합 질의 패턴을 구성하기 위하여, 삼차원 질의 영역의 형태(Q_1:1:1, Q_1:2:4, Q_1:4:16, Q_1:8:64, 및 Q_1:16:256)별로 200개씩 도메인 공간상에 균일하게 분포하도록 생성하여 이들을 혼합하여 사용하였다. 이들 모든 질의들에 대하여 정규화 과정을 거쳐 생성된 정규화된 질의 영역들에 대하여 각 축의 구간 크기를 더한 값의 비는 1:6:68로 계산되었다. 그림 7은 세 번째 실험의 결과를 나타낸다. 그림 7에서 알 수 있듯이 실험에서 주어진 모든 질의들을 처리하기 위하여 측정한 색인페이지의 평균 액세스 수는 구간비가 1:6:68인 페이지 영역들로 구성된 MD-TIX에서 4.8개로서 가장 작은 액세스 수를 나타내었다. 이와 같은 실험 결과는 다양한 형태의 질의 영역들에 의해 교차되는 페이지 영역의 개수를 최소로 하는 페이지 영역의 최적 구간비는 정규화 과정을 통하여 주어진 질의 영역들에 대해 각 축별로 구간 크기를 더한 값의 비로서 계산할 수 있음을 보이기 위한 것으로 MD-TIX 색인구조의 조율알고리즘의 타당성을 입

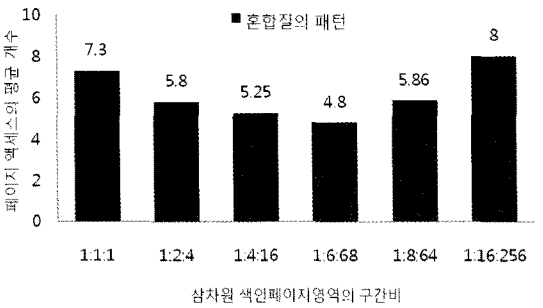

그림 7. 서로 다른 모양의 색인 페이지 영역으로 구성된 삼차원 MD-TIX들에 대한 혼합 질의 패턴의 성능비교

증하는 것이다.

끝으로, 네 번째 실험에서는 본 논문에서 제안한 다차원 MD-TIX의 조율 알고리즘을 이용하여 구성한 MD-TIX가 기존의 조율과정을 거치지 않고 구축한 MD-TIX[9]과 비교하여 얼마만큼의 성능개선 효과가 있는지를 알아본다. 먼저, 다섯 가지의 삼차원 질의 영역의 형태인 Q_1:1:1, Q_1:2:4, Q_1:4:16, Q_1:8:64, 및 Q_1:16:256에 대하여, 각 형태별로 1000개의 질의 영역들이 도메인 공간상에 균일하게 주어지는 다섯 가지의 질의 패턴을 생성한다. 그리고 각 질의 패턴에 대하여 조율 알고리즘에 의한 MD-TIX를 생성하여 그 질의 패턴을 처리할 때 발생하는 평균 색인 페이지 액세스 수를 구하고, 이 값에 대한 기존의 MD-TIX(즉, 색인 페이지 영역의 구간비 = 1:1:1)에서 같은 질의 패턴을 처리할 때 발생하는 평균 페이지 액세스 수의 비율을 측정한다. 그림 8은 이에 대한 실험 결과를 나타낸 것이다. 그림 8에서 나타난 바와 같이 질의 패턴을 구성하는 질의 영역의 구간비가 1:1:1에서 멀어질수록 제안된 조율기법을 사용하는 경우의 성능개선 효과가 뚜렷해짐을 볼 수

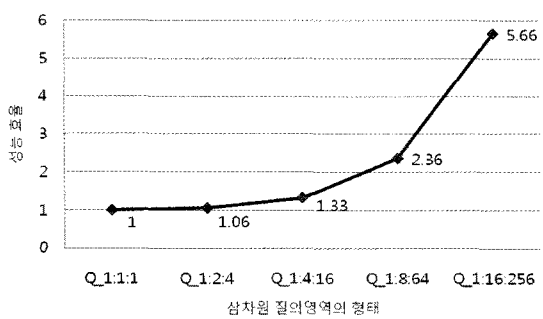

그림 8. 질의 패턴을 구성하는 질의 영역의 형태별 다차원 MD-TIX의 조율기법에 의해서 생성된 삼차원 MD-TIX의 성능 효율

있다. 즉, 질의 영역의 구간비가 1:16:256인 경우 질의 처리 성능이 다섯 배 이상까지 향상됨을 볼 수 있으며, 구간비가 더 커질수록 더욱더 성능이 더욱더 크게 향상될 수 있음을 나타낸다. 이러한 결과는 제 3.3절에서 제시한 다차원 MD-TIX 조율 알고리즘의 성능개선 효과를 잘 나타내는 것이다.

5. 결 론

본 논문에서는 XML 데이터베이스의 타입상속 계층과 중첩 엘리먼트가 포함된 복합 형태의 XML 질의의 처리를 지원하기 위하여 제안된 다차원 타입상속 색인구조인 MD-TIX에 대하여, 질의 패턴에 따라 질의처리 성능을 최적으로 보장할 수 있는 색인구조의 조율 알고리즘을 제안한다. MD-TIX는 XML 데이터베이스의 중첩 술어에 타겟 타입상속 계층뿐만 아니라 복합 엘리먼트의 도메인 타입상속 계층 모두에 타입 대치가 있는 경우에도 질의처리를 잘 지원할 수 있는 색인구조이다.

MD-TIX의 조율 알고리즘에서는 먼저, XML 질의에서 사용되는 중첩 술어들이 다차원 도메인 공간상에 매핑되는 질의 영역들의 형태에 대한 정보를 기반으로, 질의 영역들에 의해 교차하는 색인 페이지 영역들의 개수가 최소로 되는 색인 페이지 영역의 최적 구간비를 결정한다. 이는 MD-TIX의 색인 킷값 도메인과 함께 타겟 타입상속 계층으로부터 색인 키 엘리먼트가 정의된 타입상속 계층까지의 여러 개의 타입식별자 도메인 사이의 색인 엔트리들에 대한 클러스터링 정도를 나타낸다. 그리고 다차원 색인구조에서 이러한 최적 구간비를 갖는 페이지 영역들이 되도록 하는 구간반분 전략을 적용함으로써 최적의 다차원 MD-TIX를 구성한다.

또한, 본 논문에서는 MD-TIX의 조율 알고리즘의 성능평가를 위하여 다양한 실험을 수행하였으며, 실험 결과로서 주어진 질의 패턴과 데이터분포에 따라 색인구조를 조율함으로써 최적의 MD-TIX를 구성할 수 있음을 확인하였다. 중첩 엘리먼트의 경로 길이가 2인 경우에 주어지는 중첩 술어에 대한 삼차원 질의 영역의 경우, 모양이 편향된 정도에 따라 조율없이 구성된 삼차원 MD-TIX에 비해 질의처리의 성능이 매우 크게 향상됨을 알 수 있었다. 특히, 주어진 중첩 술어에 대한 질의 영역의 모양이 매우 편향되게

주어지는 경우 조율 알고리즘에 의한 색인 탐색의 효율이 다섯 배 이상까지 향상됨을 확인하였다. 이는 제안된 조율 알고리즘이 실제적으로 매우 유용함을 보여주는 것이다.

참 고 문 헌

[1] T. Bray et al., *Extensible Markup Language, (XML) 1.0. W3C Recommendation*, http: // www.w3.org /TR /REC-xml-19980210, Feb. 2004.

[2] W. Meier, "eXist: An Open Source native XML Database," *Web, Web-Services, and Database Systems, NODe 2002 Web- and Database-Related Workshops,* Revised Papers (Lecture Notes in Computer Science Vol.2593), pp. 169-183, 2003.

[3] C.D. Fallside and P. Walmsley, *XML Schema Part 0. W3C Recommendation*, http: //www. w3.org /TR /xmlschema-0, Oct. 2004.

[4] A. Berglund et al., "XML Path Language (XPath) 2.0. W3C Working Draft 30 Apr. 2002," *http: //www.w3.org /TR /xpath20*, Working Draft, 2002.

[5] R. Goldman and J. Widom, "DataGuides: Enable Query Formulation and Optimization in Semistructured DataBases," In *Proc. Int'l Conf. on Very Large Data Bases*, Athens, Greece, pp. 436-445, 1999.

[6] T. Milo and D. Suciu, "Index Structures for Path Expression," In *Proc. Int'l Conf. on Database Theory*, Jerusalem, Israel, pp. 277-295, 1999.

[7] C.W. Chung, J.K. Min, and K. Shim. "APEX: An Adaptive Path Index for XML Data," In *Proc. Intl. Conf. on Management of Data, ACM SIGMOD*, Madison, Wisconsin, pp. 121-132, 2005.

[8] K.P. Leela, and J.R. Haritsa, "Schema-conscious XML indexing," *Information Systems 32*, pp. 344-364, 2007.

[9] J.H. Lee, "MD-TIX: Multidimensional Type Inheritance Indexing for Efficient Execution of XML Queries," *Journal of Korea Multimedia Society*, Vol. 10, No. 9, pp. 1093-1105, Sept. 2007.

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

[11] D. Lomet and B. Salzberg, "The hB-tree: A Multiattribute Indexing Method with Good Guaranteed Performance," *ACM Trans. on Database Systems*, Vol. 15, No. 4, pp. 625-658, Dec. 1990.

[12] K.Y. Whang and R. Krishnamurthy, "The Multilevel Grid File- A Dynamic Hierarchical Multidimensional File Structure," In Proc. Intl. Conf. on Database Systems for Advanced Applications(DASFAA), Tokyo, pp. 449-459, 1991.

[13] S. Boag et al., *XQuery 1.0: An XML Query Language*, http://www.w3.org/ TR/xquery, Nov. 2005.

[14] S.C. Haw and C.S. Lee, "Extending Path Summary and Region Encoding for Efficient Structural Query Processing in Native XML Databases," *The Journal of Systems and Software 82*, pp. 1025-1035, 2009.

[15] N. Bruno, N. Koudas, and D. Srivastava, "Holistic twig joins: optimal XML pattern matching," *In* Proceeding of the 2002 ACM SIGMOD International Conference on Management of Data, pp. 310-321, 2002.

[16] T. Chen, J. Lu, and T.W. Ling, "On Boosting Holism in XML Twig Pattern Matching Using Structural Indexing Techniques," In Proceeding of the 2005 ACM SIGMOD International conference on Management of Data, pp. 455-466, 2005.

[17] R. Kaushik et al., "On the Integration of Structure Indexes and Inverted Lists," In Proceeding of the 2004 ACM SIGMOD inter-

national conference on Management of data, pp. 779-790, 2004.

이 종 학

1982년 경북대학교 전자공학과 (전자계산 전공) 학사
1984년 한국과학기술원 전산학과 석사
1997년 한국과학기술원 전산학과 박사
1991년 정보처리기술사
1984년~1987년 금성통신(주) 부설연구소 주임연구원
1987년~1998년 한국통신 연구개발본부 선임연구원
1998년~현재 대구가톨릭대학교 컴퓨터정보통신공학부 교수
관심분야 : 객체관계형 데이터베이스, 다차원 파일구조, 물리적 데이터베이스 설계, XML 데이터베이스, 데이터 웨어하우스 등