



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2010-0095956
(43) 공개일자 2010년09월01일

(51) Int. Cl.

G06F 17/21 (2006.01) G06F 19/00 (2006.01)
G06F 17/00 (2006.01)

(21) 출원번호 10-2009-0015024

(22) 출원일자 2009년02월23일

심사청구일자 2009년02월23일

(71) 출원인

충남대학교산학협력단

대전광역시 유성구 궁동 220번지 충남대학교

(72) 발명자

강지훈

대전광역시 유성구 궁동 충남대학교 컴퓨터전공
웹정보시스템연구실

박종현

대전광역시 유성구 궁동 충남대학교 컴퓨터전공
웹정보시스템연구실

(74) 대리인

권오식, 박창희, 김종관

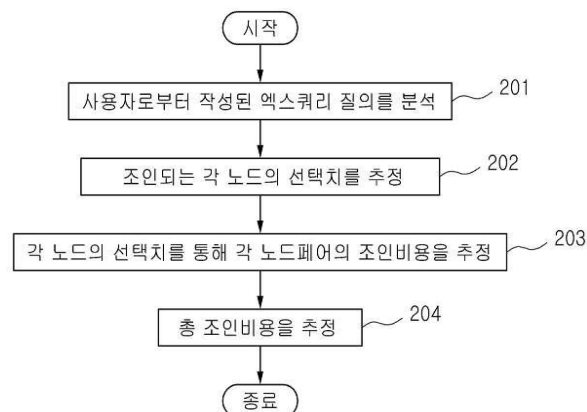
전체 청구항 수 : 총 10 항

(54) 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법

(57) 요약

본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법은, 사용자로부터 작성된 엑스쿼리(XQuery) 질의를 분석하는 단계; 조인(join)되는 각 노드(node)의 선택치(selectivity)를 추정하는 단계; 상기 추정된 상기 각 노드의 선택치를 통해 각 노드 페어(pair)의 조인 비용을 추정하는 단계; 및 상기 추정된 상기 조인 비용을 통해 총 조인 비용을 추정하는 단계를 포함한다.

대표도 - 도2



특허청구의 범위

청구항 1

사용자로부터 작성된 엑스쿼리(XQuery) 질의를 분석하는 단계;
 조인(join)되는 각 노드(node)의 선택치(selectivity)를 추정하는 단계;
 상기 추정된 상기 각 노드의 선택치를 통해 각 노드 페어(pair)의 조인 비용을 추정하는 단계; 및
 상기 추정된 상기 조인 비용을 통해 총 조인 비용을 추정하는 단계
 를 포함하는 것을 특징으로 하는 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법.

청구항 2

제1항에 있어서,
 상기 엑스쿼리(XQuery) 질의는 분산된 XML 뷰들을 대상으로 작성되고, 상기 XML 뷰는 상기 각 노드의 발생(occurrence) 정보 및 원소 수(cardinality) 정보를 포함하는 것을 특징으로 하는 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법.

청구항 3

제1항에 있어서,
 상기 사용자로부터 작성된 엑스쿼리(XQuery) 질의를 분석하는 단계는,
 상기 엑스쿼리 질의를 통해 조인의 구조 및 노드들의 연결 관계를 분석하는 단계; 및
 상기 조인되는 노드들 사이의 지정된 경로를 찾는 단계
 를 포함하는 것을 특징으로 하는 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법.

청구항 4

제1항에 있어서,
 상기 조인되는 각 노드의 선택치를 추정하는 단계는,
 동일한 문서에서 이전 조인 노드와 다음 조인 노드 간의 지정된 경로에 존재하는 노드들의 노드 선택치를 추정하는 단계
 를 포함하는 것을 특징으로 하는 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법.

청구항 5

제4항에 있어서,
 상기 지정된 경로에 존재하는 노드들의 노드 선택치를 추정하는 단계는,
 부모 노드(parent node)의 선택치를 통해 자식 노드(child node)의 선택치를 추정하는 경우, 상기 자식 노드의 선택치를 상기 부모 노드의 선택치와 동일하게 하는 단계
 를 포함하는 것을 특징으로 하는 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법.

청구항 6

제4항에 있어서,
 상기 지정된 경로에 존재하는 노드들의 노드 선택치를 추정하는 단계는,
 자식 노드의 선택치를 통해 부모 노드의 선택치를 추정하는 경우, 상기 자식 노드 및 상기 부모 노드에 각각 대응하는 발생 정보를 통해 발생 관계를 판단하는 단계; 및
 상기 판단 결과 상기 자식 노드와 상기 부모 노드의 발생 관계가 상기 하나의 부모 노드에 대하여 반드시 상기

하나의 자식 노드만을 가지는 경우(1:1) 또는 상기 하나의 부모 노드에 대하여 상기 자식 노드를 갖지 않거나 상기 하나의 자식 노드만을 가지는 경우(1:?), 상기 부모 노드의 선택치를 상기 자식 노드의 선택치와 동일하게 하는 단계

를 포함하는 것을 특징으로 하는 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법.

청구항 7

제4항에 있어서,

상기 지정된 경로에 존재하는 노드들의 노드 선택치를 추정하는 단계는,

자식 노드의 선택치를 통해 부모 노드의 선택치를 추정하는 경우, 상기 자식 노드 및 상기 부모 노드에 각각 대응하는 발생 정보를 통해 발생 관계를 판단하는 단계;

상기 판단 결과 상기 자식 노드와 상기 부모 노드의 발생 관계가 상기 하나의 부모 노드에 대하여 반드시 상기 하나 이상의 자식 노드를 가지는 경우(1:+) 또는 상기 하나의 부모 노드에 대하여 상기 자식 노드를 갖지 않거나 상기 하나 이상의 자식 노드를 가지는 경우(1:*), 상기 자식 노드가 선택될 수 있는 모든 경우의 수에 대한 상기 부모 노드가 선택될 수 있는 확률을 계산하는 단계; 및

상기 계산된 상기 확률을 통해 상기 부모 노드의 선택치를 추정하는 단계

를 포함하는 것을 특징으로 하는 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법.

청구항 8

제1항에 있어서,

상기 추정된 상기 각 노드의 선택치를 통해 각 노드 페어(pair)의 조인 비용을 추정하는 단계는,

조인의 구조를 분석한 결과, 상기 조인의 구조가 서로 다른 문서 사이의 조인 연산인 경우, 상기 추정된 상기 각 노드의 선택치를 통해 상기 조인 연산을 위해 선택되는 노드의 개수를 추정하는 단계;

상기 추정된 상기 노드의 개수를 통해 상기 서로 다른 문서 사이의 조인 선택치를 추정하는 단계; 및

상기 추정된 상기 조인 선택치를 통해 상기 조인 비용을 추정하는 단계

를 포함하는 것을 특징으로 하는 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법.

청구항 9

제1항에 있어서,

상기 추정된 상기 조인 비용을 통해 총 조인 비용을 추정하는 단계는,

상기 추정된 하나 이상의 조인 비용을 합하여 총 조인 비용을 추정하는 것을 특징으로 하는 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법.

청구항 10

제1항 내지 제9항 중 어느 한 항의 방법을 실행시키기 위한 프로그램을 기록한 컴퓨터 판독 가능한 기록 매체.

명 세 서

발명의 상세한 설명

기술 분야

본 발명은 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법에 관한 것으로, 더욱 상세하게는 이전 조인 노드와 연결되어 있는 다음 조인 노드의 선택치를 추정하는 알고리즘(Estimating Selectivity for Next Join: ESNJ)을 통해 조인 노드의 선택치를 추정하고, 상기 추정된 상기 조인 노드의 선택치를 통해 엑스쿼리 질의 조인 연산 비용을 추정하는 방법에 관한 것이다.

[0001]

배경 기술

[0002] XML은 인터넷 상에서 데이터 교환을 위한 표준이다. 그러므로, 이종의 레거시 데이터를 보유하고 있는 응용들은 XML 데이터와의 통합을 위하여 다양한 접근 방법들을 찾고 있으며, 그 가운데 하나가 XML 뷰를 이용하여 이종 데이터들을 XML 데이터처럼 바라볼 수 있도록 하는 방법이다. 즉, 분산 환경에 XML 데이터들을 통합하기 위한 한가지 방법은 XML 뷰를 사용하는 것이다. XML 뷰는 DTD, XML Schema, XQuery 등을 이용하여 기술할 수 있다. 사용자는 분산된 XML 뷰들을 대상으로 질의할 수 있으며, 이 때 검색을 위한 질의어로 XQuery를 사용할 수 있다. 분산된 다수의 XML 뷰들을 대상으로 작성된 XQuery 질의는 여러 지역 시스템들에 저장된 데이터들을 통합하고 검색하므로 다수의 조인 연산을 포함할 수 있다. SQL 질의의 경우, 조인 연산은 질의 처리를 위하여 비용이 많이 드는 연산자들 가운데 하나로 구분된다. 그러므로, 조인 연산을 효율적으로 처리하는 것은 결국 질의의 처리 성능을 좌우한다. XQuery 질의 역시 이와 크게 다르지 않으므로 효율적인 질의 처리를 위하여 질의에 포함된 조인 연산을 처리하기 위한 연구는 반드시 필요하다. 특별히 분산된 다수의 XML 뷰들을 대상으로 작성된 XQuery 질의는 분산된 다수의 지역 시스템들 사이에 조인 연산을 포함하므로 이 연산을 처리하기 위한 방법은 질의 처리 시간을 결정하는 중요한 요인이다. SQL 질의의 경우, 분산 환경에서 질의를 효율적으로 처리하기 위하여 조인 선택치를 추정하고 이를 기반으로 조인 처리 순서를 결정하기 위한 연구들은 이미 많이 존재한다. XQuery 질의의 경우 역시 XML 전용 데이터베이스를 연구하는 곳에서 질의에 포함된 조인을 효율적으로 처리하기 위한 몇몇 연구가 존재한다. 그러나 아직까지 분산 환경에 XQuery 질의의 조인을 처리하기 위한 연구는 많지 않다. 그러므로, XQuery 질의의 조인을 효율적으로 처리하기 위해서 SQL 질의의 조인 처리 방법을 참조하는 것은 바람직한 방법처럼 보인다.

[0003] SQL 질의의 조인 처리 방법 가운데 대표적인 한가지 방법은 분산 환경에서 SQL 질의의 조인 선택치를 추정하고 이를 기반으로 비용을 예측하기 위한 비용 모델을 제안하고 이를 기반으로 최적의 조인 순서를 결정하는 것이다. 이러한 조인 순서 결정을 위한 방법을 XQuery 질의의 조인 처리 방법을 위해 적용하는 것은 바람직해 보인다. 그러므로 종래의 SQL 질의의 조인 비용 추정 방법 중 일부를 XQuery 질의의 조인 비용 추정을 위해서 참조가 가능하다. 예를 들면, 테이블의 주 키를 이용하여 SQL 질의의 조인 선택치를 구하는 방법이 이에 해당된다. XML 문서의 ID속성이 테이블의 키와 유사한 역할을 하므로 SQL 질의의 조인 선택치를 구하는 법을 XQuery 질의의 조인 선택치를 구하는 방법에 적용할 수 있다. 표 1은 SQL 질의의 조인 선택치 추정 방법 및 상기 SQL 질의의 조인 선택치 추정 방법을 적용한 종래의 XQuery 질의의 조인 선택치 추정 방법을 나타낸다.

표 1

조건	에스큐엘(SQL)	엑스큐리(XQuery)
하나의 피연산자가 아이디(key)인 경우	A.value=B.value → 1/ B.key	\$A/text()=\$B/@id → 1/ \$B/@id
두개의 피연산자 모두가 아이디(key)인 경우	A.key=B.key → 1/Max(A.key , B.key)	\$A/@id=\$B/@id → 1/Max(\$A/@id , \$B/@id)
두개의 피연산자 모두가 아이디(key)가 아닌 경우	A.value=B.value → 1/Max(A.value , B.value)	\$A/text()=\$B/text() → 1/Max(\$A , \$B)

[0005] 그러나 SQL은 그 질의의 대상이 테이블 구조의 데이터인 반면 XQuery는 반 구조적인 XML 문서이다. 예를 들어, 테이블의 각 속성들 사이의 관계는 모두 일대일 관계인 반면 XML 문서의 각 노드들 사이의 관계는 일 대 일의 관계, 다 대 다의 관계, 다 대 일의 관계, 또는 일 대 다의 관계로 구현될 수 있다.

[0006] 도 1은 관계형 데이터베이스의 테이블과 XML 데이터의 구조적인 차이를 나타낸다. 이처럼, SQL과 XQuery의 구조적인 특성 때문에 이들이 제안하고 있는 비용 모델을 XQuery 질의의 조인 선택치를 구하는 방법에 그대로 적용하는 것은 어렵다.

[0007] 그러므로, XQuery 질의의 조인 순서를 결정하기 위하여 XQuery 질의를 위한 조인 선택치 추정 방법 및 상기 XQuery 질의를 위한 조인 선택치 추정 방법을 통해 XQuery 질의에 포함된 조인 연산의 비용을 추정하는 방법이 요구되고 있다.

발명의 내용

해결 하고자하는 과제

[0008] 본 발명은 상기와 같은 종래 기술을 개선하기 위해 안출된 것으로서, XML 데이터의 구조적인 특성을 고려하여 앞선 조인 노드의 선택치로부터 다음 조인 노드의 선택치를 추정하고 상기 추정된 상기 조인 노드의 선택치를 통해 엑스쿼리 질의에 포함된 조인 연산의 비용을 추정함으로써, 상기 엑스쿼리 질의에 포함된 상기 조인 연산을 효율적으로 처리할 수 있도록 하는 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법을 제공하는 것을 목적으로 한다.

과제 해결수단

[0009] 상기의 목적을 이루고 종래기술의 문제점을 해결하기 위하여, 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법은, 사용자로부터 작성된 엑스쿼리(XQuery) 질의를 분석하는 단계; 조인(join)되는 각 노드(node)의 선택치(selectivity)를 추정하는 단계; 상기 추정된 상기 각 노드의 선택치를 통해 각 노드 페어(pair)의 조인 비용을 추정하는 단계; 및 상기 추정된 상기 조인 비용을 통해 총 조인 비용을 추정하는 단계를 포함한다.

[0010] 또한, 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법에서, 상기 엑스쿼리(XQuery) 질의는 분산된 XML 뷰들을 대상으로 작성되고, 상기 XML 뷰는 상기 각 노드의 발생(occurrence) 정보 및 원소 수(cardinality) 정보를 포함하는 것을 특징으로 한다.

[0011] 또한, 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법에서, 상기 사용자로부터 작성된 엑스쿼리(XQuery) 질의를 분석하는 단계는, 상기 엑스쿼리 질의를 통해 조인의 구조 및 노드들의 연결 관계를 분석하는 단계; 및 상기 조인되는 노드들 사이의 지정된 경로를 찾는 단계를 포함하는 것을 특징으로 한다.

[0012] 또한, 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법에서, 상기 조인되는 각 노드의 선택치를 추정하는 단계는, 동일한 문서에서 이전 조인 노드와 다음 조인 노드 간의 지정된 경로에 존재하는 노드들의 노드 선택치를 추정하는 단계를 포함하는 것을 특징으로 한다.

[0013] 또한, 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법에서, 상기 지정된 경로 사이에 존재하는 노드들의 노드 선택치를 추정하는 단계는, 부모 노드(parent node)의 선택치를 통해 자식 노드(child node)의 선택치를 추정하는 경우, 상기 자식 노드의 선택치를 상기 부모 노드의 선택치와 동일하게 하는 단계를 포함하는 것을 특징으로 한다.

[0014] 또한, 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법에서, 상기 지정된 경로 사이에 존재하는 노드들의 노드 선택치를 추정하는 단계는, 자식 노드의 선택치를 통해 부모 노드의 선택치를 추정하는 경우, 상기 자식 노드 및 상기 부모 노드에 각각 대응하는 발생 정보를 통해 발생 관계를 판단하는 단계; 및 상기 판단 결과 상기 자식 노드와 상기 부모 노드의 발생 관계가 상기 하나의 부모 노드에 대하여 반드시 상기 하나의 자식 노드만을 가지는 경우(1:1) 또는 상기 하나의 부모 노드에 대하여 상기 자식 노드를 갖지 않거나 상기 하나의 자식 노드만을 가지는 경우(1:?), 상기 부모 노드의 선택치를 상기 자식 노드의 선택치와 동일하게 하는 단계를 포함하는 것을 특징으로 한다.

[0015] 또한, 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법에서, 상기 지정된 경로 사이에 존재하는 노드들의 노드 선택치를 추정하는 단계는, 자식 노드의 선택치를 통해 부모 노드의 선택치를 추정하는 경우, 상기 자식 노드 및 상기 부모 노드에 각각 대응하는 발생 정보를 통해 발생 관계를 판단하는 단계; 상기 판단 결과 상기 자식 노드와 상기 부모 노드의 발생 관계가 상기 하나의 부모 노드에 대하여 반드시 상기 하나 이상의 자식 노드를 가지는 경우(1:+) 또는 상기 하나의 부모 노드에 대하여 상기 자식 노드를 갖지 않거나 상기 하나 이상의 자식 노드를 가지는 경우(1:*), 상기 자식 노드가 선택될 수 있는 모든 경우의 수에 대한 상기 부모 노드가 선택될 수 있는 확률을 계산하는 단계; 및 상기 계산된 상기 확률을 통해 상기 부모 노드의 선택치를 추정하는 단계를 포함하는 것을 특징으로 한다.

[0016] 또한, 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법에서, 상기 추정된 상기 각 노드의 선택치를 통해 각 노드 페어(pair)의 조인 비용을 추정하는 단계는, 조인의 구조를 분석한 결과, 상기 조인의 구조가 서로 다른 문서 사이의 조인 연산인 경우, 상기 추정된 상기 각 노드의 선택치를 통해 상기 조인 연산을 위해 선택되는 노드의 개수를 추정하는 단계; 상기 추정된 상기 노드의 개수를 통해 상기 서로 다른 문서 사이의 조인 선택치를 추정하는 단계; 및 상기 추정된 상기 조인 선택치를 통해 상기 조인 비용

을 추정하는 단계를 포함하는 것을 특징으로 한다.

[0017] 또한, 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법에서, 상기 추정된 상기 조인 비용을 통해 총 조인 비용을 추정하는 단계는, 상기 추정된 하나 이상의 조인 비용을 합하여 총 조인 비용을 추정하는 것을 특징으로 한다.

효 과

[0018] 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법에 따르면, XML 데이터의 구조적인 특성을 고려하여 앞선 조인 노드의 선택치로부터 다음 조인 노드의 선택치를 추정하고 상기 추정된 상기 조인 노드의 선택치를 통해 엑스쿼리 질의에 포함된 조인 연산의 비용을 추정함으로써, 상기 엑스쿼리 질의에 포함된 상기 조인 연산을 효율적으로 처리할 수 있도록 하는 효과를 얻을 수 있다.

발명의 실시를 위한 구체적인 내용

[0019] 이하에서는 첨부된 도면을 참조하여 본 발명의 실시예를 상세히 설명한다.

[0020] 도 2는 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법의 흐름을 도시한 도면이다.

[0021] 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기는 사용자로부터 작성된 엑스쿼리 질의를 분석한다(단계(201)).

[0022] 상기 엑스쿼리(XQuery) 질의는 분산된 XML 뷰들을 대상으로 작성될 수 있고, 상기 XML 뷰는 상기 각 노드의 발생(occurrence) 정보 및 원소 수(cardinality) 정보를 포함할 수 있다.

[0023] 예를 들어, 수학적 식 1을 통해 기재된 알고리즘과 같이, 사용자는 분산된 XML 뷰들을 대상으로 엑스쿼리 질의를 작성할 수 있다.

수학적 식 1

[0024] For \$A in doc("A.xml")/root_A/a

[0025] For \$A2 in \$A/a2

[0026] For \$B in doc("B.xml")/root_B/b

[0027] For \$B2 in \$B/b1/b2

[0028] For \$B3 in \$B/b3

[0029] For \$C in doc("C.xml")/root_C/c

[0030] For \$C1 in \$C/c1

[0031] Where \$A2 = \$B2 and \$B3 = \$C1

[0032] Return

[0033] <result>{\$A/a1, \$C/c2}</result>

[0034] 도 3은 본 발명의 일실시예에 따른 수학적 식 1의 엑스쿼리 질의에 대한 조인의 구조 및 노드들의 연결 관계를 도시한 도면이다.

[0035] 상기 XML 뷰들의 각 노드에 표현된 라벨(Label)은 각 노드의 발생(occurrence) 정보 및 원소 수(cardinality) 정보를 나타낼 수 있다.

[0036] 예를 들어, 지역 시스템 B의 b1노드에 대응하는 라벨 (?,4)는 b1노드가 지역 시스템 B에 4개 존재하며, b1의 부모노드인 b노드에 대하여 b1노드가 한 개 발생하거나 발생하지 않음을 나타낼 수 있다.

[0037] 또한, 지역 시스템 B의 b2노드에 대응하는 라벨 (*,12)는 b2노드가 지역 시스템 B에 12개 존재하며, b2의 부모노드인 b1노드에 대하여 b2노드가 한 개 이상 발생하거나 발생하지 않음을 나타낼 수 있다.

[0038] 또한, 지역 시스템 B의 b3노드에 대응하는 라벨 (1,25)는 b3노드가 지역 시스템 B에 25개 존재하며, b3의 부모

노드인 b노드에 대하여 b3노드가 한 개 발생함을 나타낼 수 있다.

- [0039] 또한, 지역 시스템 B의 b노드에 대응하는 라벨 (+,25)는 b노드가 지역 시스템 B에 25개 존재하며, b의 부모 노드에 대하여 b노드가 한 개 이상 발생함을 나타낼 수 있다.
- [0040] 상기 원소 수 정보는 XQuery Function and Operations의 Count()함수뿐만 아니라, 원소 수를 계산하기 위한 다양한 종류의 기술을 통해 취득될 수 있다.
- [0041] 단계 (201)에서, 상기 엑스쿼리 질의 조인 비용 추정기는 상기 엑스쿼리 질의를 통해 조인의 구조 및 노드들의 연결 관계를 분석한다.
- [0042] 예를 들어, 다시 도 3을 참조하면 상기 엑스쿼리 질의 조인 비용 추정기는 지역 시스템 A의 변수 a2와 지역 시스템 B의 변수 b2가 조인으로 연결되어 있고, 지역 시스템 B의 b3와 지역 시스템 C의 c1이 조인으로 연결되어 있음을 분석할 수 있다.
- [0043] 또한, 상기 엑스쿼리 질의 조인 비용 추정기는 상기 조인되는 노드들 사이의 지정된 경로를 찾는다.
- [0044] 예를 들어, 다시 도 3을 참조하면 상기 엑스쿼리 질의 조인 비용 추정기는 b2노드로부터 b3노드까지의 지정된 경로를 찾을 수 있다.
- [0045] 상기 엑스쿼리 질의 조인 비용 추정기는 조인(join)되는 각 노드(node)의 선택치(selectivity)를 추정한다 (단계(202)).
- [0046] 예를 들어, 다시 도 3을 참조하면 상기 엑스쿼리 질의 조인 비용 추정기는 표 1을 통해 지역 시스템 A의 a2노드와 지역 시스템 B의 b2노드 사이의 조인 선택치를 추정할 수 있다. 상기 지역시스템 A의 a2노드와 지역 시스템 B의 b2노드 사이의 조인 선택치는 수학적식 2를 통해 구현될 수 있다.

수학적식 2

- [0047] $\text{Join Selectivity}(A.a2, B.b2) = 1/\text{Max}(|A.a2|, |B.b2|) = 1/|B.b2| = 1/12$
- [0048] 상기 엑스쿼리 질의 조인 비용 추정기는 상기 추정된 상기 조인 선택치를 통해 첫 번째 조인 연산인 $|A.a2 \bowtie B.b2|$ 의 조인 비용을 추정할 수 있다. 상기 $|A.a2 \bowtie B.b2|$ 의 조인 비용은 수학적식 3을 통해 구현될 수 있다.

수학적식 3

- $$|A.a2 \bowtie B.b2| = \text{Join Selectivity}(A.a2, B.b2) \times |A.a2| \times |B.b2| = 1/12 \times 3 \times 12 = 3$$
- [0049]
- [0050] 단계 (202)에서, 상기 엑스쿼리 질의 조인 비용 추정기는 동일한 문서에서 이전 조인 노드와 다음 조인 노드 간의 지정된 경로에 존재하는 노드들의 노드 선택치를 추정한다.
- [0051] 예를 들어, 다시 도 3을 참조하면 상기 엑스쿼리 질의 조인 비용 추정기는 b2노드로부터 b3노드까지의 지정된 경로 사이에 존재하는 노드들의 노드 선택치를 추정할 수 있다.
- [0052] 상기 엑스쿼리 질의 조인 비용 추정기는 부모 노드(parent node)의 선택치를 통해 자식 노드(child node)의 선택치를 추정하는 경우, 상기 자식 노드의 선택치를 상기 부모 노드의 선택치와 동일하게 한다.
- [0053] 또한 상기 엑스쿼리 질의 조인 비용 추정기는 자식 노드의 선택치를 통해 부모 노드의 선택치를 추정하는 경우, 상기 자식 노드 및 상기 부모 노드에 각각 대응하는 발생 정보를 통해 발생 관계를 판단하고, 상기 판단 결과 상기 자식 노드와 상기 부모 노드의 발생 관계가 상기 하나의 부모 노드에 대하여 반드시 상기 하나의 자식 노드만을 가지는 경우(1:1) 또는 상기 하나의 부모 노드에 대하여 상기 자식 노드를 갖지 않거나 상기 하나의 자식 노드만을 가지는 경우(1:?), 상기 부모 노드의 선택치를 상기 자식 노드의 선택치와 동일하게 한다.
- [0054] 또한 상기 엑스쿼리 질의 조인 비용 추정기는 자식 노드의 선택치를 통해 부모 노드의 선택치를 추정하는 경우, 상기 자식 노드 및 상기 부모 노드에 각각 대응하는 발생 정보를 통해 발생 관계를 판단하고, 상기 판단 결과 상기 자식 노드와 상기 부모 노드의 발생 관계가 상기 하나의 부모 노드에 대하여 반드시 상기 하나 이상의 자

식 노드를 가지는 경우(1:+) 또는 상기 하나의 부모 노드에 대하여 상기 자식 노드를 갖지 않거나 상기 하나 이상의 자식 노드를 가지는 경우(1:*), 상기 자식 노드가 선택될 수 있는 모든 경우의 수에 대한 상기 부모 노드가 선택될 수 있는 확률(P(i))을 계산한다. 상기 확률은 수학식 4을 통해 구현될 수 있다.

수학식 4

$$P(i) = \frac{\left({}_{i-r}C_q - \sum_{j=MinP}^{i-1} E(j,i) \right) \times {}_m C_i}{{}_n C_q}$$

$$E(j,i) = \left({}_{j-r}C_q - \sum_{l=MinP}^{j-1} E(l,j) \right) \times {}_i C_j$$

[0055]

[0056] 수학식 4에서 i는 부모 노드가 선택될 수 있는 모든 경우의 수, m은 부모 노드의 총 수, MinP는 선택 가능한 부모 노드의 최소 수, MaxP는 선택 가능한 부모 노드의 최대 수, n은 자식 노드의 총 수, r은 부모에 대한 평균 자식 수(= n/m), q는 n × Selectivity(C), E(j,i)는 부모 노드가 i개 선택될 경우 부모 노드가 j개 선택될 경우의 수를 의미한다.

[0057]

상기 엑스쿼리 질의 조인 비용 추정기는 상기 계산된 상기 확률을 통해 상기 부모 노드의 선택치를 추정한다. 상기 부모 노드의 선택치는 수학식 5를 통해 구현될 수 있다.

수학식 5

$$Selectivity(P) = \sum_{i=MinP}^{MaxP} i \cdot P(i) / m$$

[0058]

[0059] 예를 들어, 다시 도 3을 참조하면 상기 엑스쿼리 질의 조인 비용 추정기는 상기 계산된 b2노드의 선택치(3/12)로부터 b1노드의 선택치를 추정할 수 있다. 이 경우는 자식 노드의 선택치를 통해 부모 노드의 선택치를 추정하는 경우 및 상기 자식 노드 및 상기 부모 노드에 각각 대응하는 발생 정보를 통해 발생 관계를 판단 결과 상기 하나의 부모 노드에 대하여 상기 자식 노드를 갖지 않거나 상기 하나 이상의 자식 노드를 가지는 경우(1:*)에 해당하므로, 상기 엑스쿼리 질의 조인 비용 추정기는 상기 자식 노드가 선택될 수 있는 모든 경우의 수에 대한 상기 부모 노드가 선택될 수 있는 확률을 계산하고, 상기 계산된 상기 확률을 통해 상기 부모 노드의 선택치를 추정할 수 있다.

[0060]

상기 b2노드의 선택치(3/12)로부터 12개의 자식 노드 중 3개의 자식 노드가 선택되었으므로 선택될 수 있는 부모 노드의 개수는 1개 내지 3개이다. 12개의 자식 노드 중 3개의 자식 노드가 선택되는 총 경우의 수는 ${}_{12}C_3$ 이다. 상기 12개의 자식 노드 중 3개의 자식 노드가 선택되는 경우에서 부모 노드가 1개 선택되는 경우를 찾아보면, ('b2₁', 'b2₂', 'b2₃'), ('b2₄', 'b2₅', 'b2₆'), ('b2₇', 'b2₈', 'b2₉'), ('b2₁₀', 'b2₁₁', 'b2₁₂')의 4가지 경우가 된다.

[0061]

또한, b1노드의 자식노드인 b2₁노드, b2₂노드 및 b2₃ 중에서 3개의 자식노드를 선택할 수 있는 경우의 수는 ${}_3C_3$ 이고, 경우의 수 ${}_3C_3$ 은 선택되는 부모 노드가 b1₂노드, b1₃노드 또는 b1₄노드인 경우에도 적용되므로, 상기 부모 노드들에서 하나의 부모노드를 선택하는 조합인 ${}_4C_1$ 을 ${}_3C_3$ 에 곱한다.

[0062]

따라서, 부모노드가 1개 선택되는 경우의 수는 ${}_3C_3 \times {}_4C_1$ 로 구할 수 있고, P(1)은 $(C_3 \times {}_4C_1) / {}_{12}C_3$ 이다.

[0063]

다음으로 부모노드가 2개 선택되는 경우를 고려해보자. 만약 부모 노드 b1₁노드 및 b1₂노드가 선택되려면 b2₁노드 내지 b2₆노드 가운데 3개가 선택되어야 하며, 그 경우의 수는 ${}_6C_3$ 이다. 그러나 ${}_6C_3$ 경우의 수에는 부모노드가 1개 선택되는 경우인, ('b2₁', 'b2₂', 'b2₃'), ('b2₄', 'b2₅', 'b2₆')도 포함되어 있기 때문에, 경우의 수 ${}_3C_3 \times {}_2C_1$ 를 제외시켜야 한다. 그러므로 부모 노드인 b1₁노드 및 b1₂노드가 선택되는 경우의 수는 ${}_6C_3 - {}_3C_3 \times {}_2C_1$ 이

된다. 경우의 수 ${}_6C_3 - {}_3C_3 \times {}_2C_1$ 는 b1노드 내지 b4노드의 부모 노드들 4개 중에서 2개를 선택하는 조합만큼 적용되므로 부모노드가 2개 선택되는 경우의 수는 $({}_6C_3 - {}_3C_3 \times {}_2C_1) \times {}_4C_2$ 가 되며, 결국 P(2)는 $({}_6C_3 - {}_3C_3 \times {}_2C_1) \times {}_4C_2 / {}_{12}C_3$ 이 된다.

[0064] 3개의 부모노드가 선택될 확률 역시 앞선 경우들과 유사하며 위의 예를 계산한 식은 다음과 같다.

[0065]
$$P(1) = ({}_3C_3 \times {}_4C_1) / {}_{12}C_3 = 4 / 220 = 1/55$$

[0066]
$$P(2) = (({}_6C_3 - {}_3C_3 \times {}_2C_1) \times {}_4C_2) / {}_{12}C_3 = (({}_6C_3 - E(1,2)) \times {}_4C_2) / {}_{12}C_3 = 108 / 220 = 27/55$$

[0067]
$$P(3) = (({}_9C_3 - ({}_3C_3 \times {}_3C_1 + ({}_6C_3 - {}_3C_3 \times {}_2C_1) \times {}_3C_2)) \times {}_4C_3) / {}_{12}C_3$$

[0068]
$$= (({}_9C_3 - (E(1,3) + E(2,3))) \times {}_4C_3) / {}_{12}C_3 = 108 / 220 = 27/55$$

[0069] 결국 상기 b2노드의 선택치(3/12)로부터 추정된 b1 노드가 선택되는 노드의 개수는 약 2.5개이며 상기 b1노드의 선택치는 수학적 6을 통해 구현될 수 있다.

수학적 6

[0070]
$$selectivity(P) = \sum_{i=MinP}^{MaxP} i \cdot P(i) / 4 = \left(1 \times \frac{1}{55} + 2 \times \frac{27}{55} + 3 \times \frac{27}{55} \right) / 4 \approx 2.5/4$$

[0071] 상기 엑스쿼리 질의 조인 비용 추정기는 상기 계산된 b1노드의 선택치(2.5/4)로부터 b노드의 선택치를 추정할 수 있다. 상기 b1노드의 선택치로부터 b노드의 선택치를 추정하는 경우는 자식 노드의 선택치를 통해 부모 노드의 선택치를 추정하는 경우 및 상기 자식 노드 및 상기 부모 노드에 각각 대응하는 발생 정보를 통해 발생 관계를 판단 결과 상기 하나의 부모 노드에 대하여 상기 자식 노드를 갖지 않거나 상기 하나의 자식 노드만을 가지는 경우(1:?)에 해당하므로, 상기 엑스쿼리 질의 조인 비용 추정기는 상기 부모 노드의 선택치를 상기 자식 노드의 선택치와 동일하게 할 수 있다.

[0072] 즉, b노드의 선택치는 b1노드의 선택치와 동일하므로 상기 b노드의 선택치는 15.625/25이다.

[0073] 상기 엑스쿼리 질의 조인 비용 추정기는 상기 계산된 b노드의 선택치로부터 b3노드의 선택치를 추정할 수 있다. 상기 b노드의 선택치로부터 b3노드의 선택치를 추정하는 경우는 부모 노드(parent node)의 선택치를 통해 자식 노드(child node)의 선택치를 추정하는 경우에 해당하므로, 상기 엑스쿼리 질의 조인 비용 추정기는 상기 자식 노드의 선택치를 상기 부모 노드의 선택치와 동일하게 할 수 있다. 즉, b3노드의 선택치는 b노드의 선택치와 동일하므로 상기 b3노드의 선택치는 15.625/25이다.

[0074] 상기 엑스쿼리 질의 조인 비용 추정기는 상기 추정된 상기 각 노드의 선택치를 통해 각 노드 페어(pair)의 조인 비용을 추정한다 (단계(203)).

[0075] 단계 (203)에서, 상기 엑스쿼리 질의 조인 비용 추정기는 조인의 구조를 분석한 결과, 상기 조인의 구조가 서로 다른 문서 사이의 조인 연산인 경우, 상기 추정된 상기 각 노드의 선택치를 통해 상기 조인 연산을 위해 선택되는 노드의 개수를 추정한다.

[0076] 예를 들어, 다시 도 3을 참조하면 상기 엑스쿼리 질의 조인 비용 추정기는 |B'.b3 $\bowtie$ C.c1|의 조인 연산을 위해 선택되는 노드의 개수를 추정할 수 있다.

[0077] 즉, |B'.b3 $\bowtie$ C.c1|의 조인 연산을 위해 b3 노드가 선택되는 노드의 개수는 15.625개이며, c1 노드가 선택되는 노드의 개수는 10개이다.

[0078] 상기 엑스쿼리 질의 조인 비용 추정기는 상기 추정된 상기 노드의 개수를 통해 상기 서로 다른 문서 사이의 조인 선택치를 추정한다.

[0079] 예를 들어, 다시 도 3을 참조하면 상기 엑스쿼리 질의 조인 비용 추정기는 표 1을 통해 지역 시스템 B의 b3노드와 지역 시스템 C의 c1노드 사이의 조인 선택치를 추정할 수 있다. 상기 지역시스템 B의 b3노드와 지역 시스템 C의 c1노드 사이의 조인 선택치는 수학적 7을 통해 구현될 수 있다.

수학식 7

[0080] $\text{Join Selectivity}(B'.b3, C.c1) = 1/\text{Max}(|B'.b3|, |C.c1|) = 1/|B'.b3| = 1/15.625$

[0081] 상기 엑스쿼리 질의 조인 비용 추정기는 상기 추정된 상기 조인 선택치를 통해 상기 조인 비용을 추정한다. 예를 들어, 다시 도 3을 참조하면 상기 엑스쿼리 질의 조인 비용 추정기는 상기 추정된 상기 조인 선택치를 통해 두 번째 조인 연산인 $|B'.b3 \bowtie C.c1|$ 의 조인 비용을 추정할 수 있다. 상기 $|B'.b3 \bowtie C.c1|$ 의 조인 비용은 수학식 8을 통해 구현될 수 있다.

수학식 8

[0082] $|B'.b3 \bowtie C.c1| = \text{조인 선택치}(B'.b3, C.c1) \times |B'.b3| \times |C.c1| = 1/15.625 \times 15.625 \times 10 = 10$

[0083] 상기 엑스쿼리 질의 조인 비용 추정기는 상기 추정된 상기 조인 비용을 통해 총 조인 비용을 추정한다 (단계 (204)).

[0084] 단계 (204)에서, 상기 엑스쿼리 질의 조인 비용 추정기는 상기 추정된 하나 이상의 조인 비용을 합하여 총 조인 비용을 추정한다.

[0085] 예를 들어, 다시 도 3을 참조하면 상기 엑스쿼리 질의 조인 비용 추정기는 세 엑스엠엘 문서를 조인하는 경우 'A.xml과 'B.xml'을 먼저 조인하고 'B.xml과 'C.xml'을 나중에 조인하는 총 비용은 상기 추정된 첫 번째 조인 연산 $|A.a2 \bowtie B.b2|$ 의 조인 비용인 3 및 상기 추정된 두 번째 조인 연산 $|B'.b3 \bowtie C.c1|$ 의 조인 비용인 10을 합하여 총 조인 비용은 13이라는 것을 추정할 수 있다.

[0086] 이와 같이, 본 발명에 따른 조인 비용 추정 방법에 따르면, 동일한 문서에서 이전 조인 노드와 다음 조인 노드 간의 지정된 경로에 존재하는 노드들의 노드 선택치를 추정하고, 상기 추정된 노드들의 노드 선택치를 통해 조인 비용을 추정할 수 있다. 상기 엑스쿼리 질의 조인 비용 추정기는 상기 추정된 상기 조인 비용을 통해 최적의 조인 순서를 결정할 수 있다.

[0087] 본 발명에 따른 조인 연산 비용 추정 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 상기 매체는 프로그램 명령, 데이터 구조 등을 지정하는 신호를 전송하는 반송파를 포함하는 광 또는 금속선, 도파관 등의 전송 매체일 수도 있다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0088] 이상과 같이 본 발명은 비록 한정된 실시예와 도면에 의해 설명되었으나, 본 발명은 상기의 실시예에 한정되는 것은 아니며, 이는 본 발명이 속하는 분야에서 통상의 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형이 가능하다. 따라서, 본 발명 사상은 아래에 기재된 특허청구범위에 의해서만 파악되어야 하고, 이의 균등 또는 등가적 변형 모두는 본 발명 사상의 범주에 속한다고 할 것이다.

도면의 간단한 설명

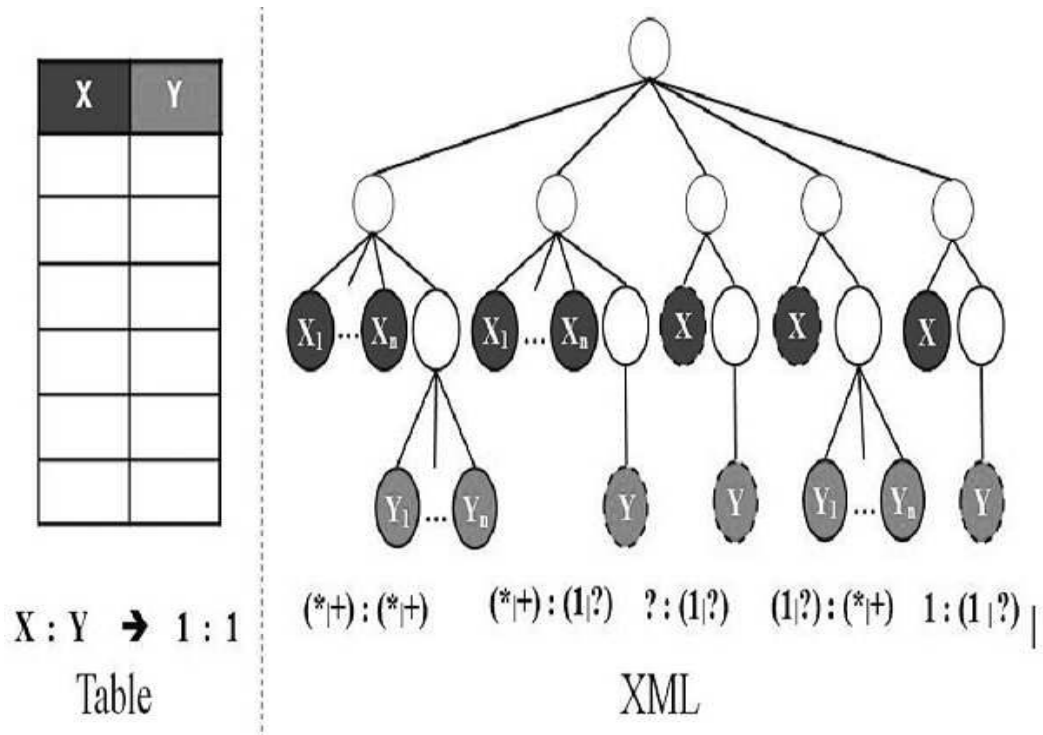
[0089] 도 1은 관계형 데이터베이스의 테이블과 XML 데이터의 구조적인 차이를 도시한 도면.

[0090] 도 2는 본 발명의 일실시예에 따른 엑스쿼리 질의 조인 비용 추정기의 조인 연산 비용 추정 방법의 흐름을 도시한 도면.

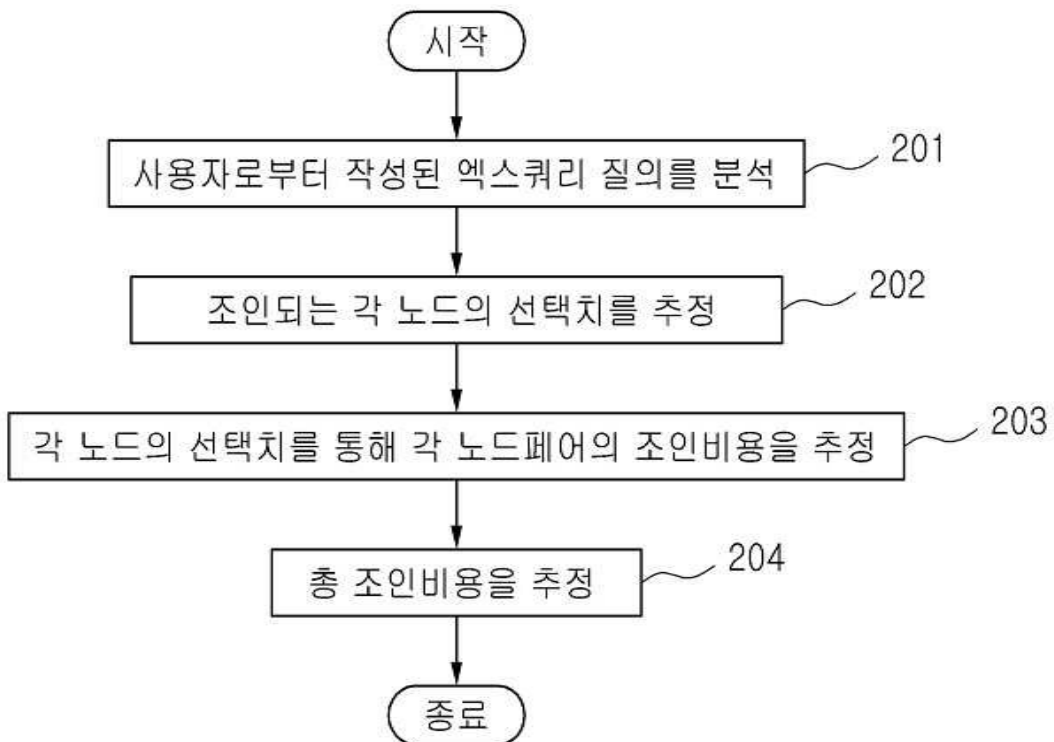
[0091] 도 3은 본 발명의 일실시예에 따른 수학식 1의 엑스쿼리 질의에 대한 조인의 구조 및 노드들의 연결 관계를 도시한 도면.

도면

도면1



도면2



도면3

