OR 연결 질의에 대한 능력 기반 질의 재작성 과정

윤정기, 이지행, 문강식, 이전영
지능정보시스템연구실, 포항공과대학교

Capabilities-based Query Rewriter for Disjunctive Query with Single Source

Jeong-Ki Yun, Ji-Haeing Lee, Kang-Sik Moon, Jeon-Young Lee
Intelligent Information Systems Lab., POSTECH

요 약

능력기반 질의 처리기는 데이터 소스의 질의 처리 능력과 데이터 소스에 대한 목적 질의(target query)를 입력으로 받아 목적 질의와 동일한 결과를 내는 질의 수행 계획을 출력하는 시스템이다. 이전 능력기반 질의 처리기에서는 AND 결합 질의(Conjunctive query)만을 처리하였기 때문에 데이터 소스의 질의 처리 능력을 충분히 이용하지 못했다. 본 논문에서는 AND 및 OR 연결 (Disjunctive query) 형태의 목적 질의에 대한 질의 재작성 방법을 제시한다 재 작성된 질의는 데이터 소스에서 처리 가능한 지원 질의(supported query)들의 유니온으로 표현된다. 제안된 시스템에서는 각 지원 질의의 질의 수행 계획에 대한 기여도와 수행 비용을 고려함으로써 질의 재작성에 필요한 탐색 공간을 줄이고, 최적화 된 질의 수행 계획을 생성할 수 있다.

1. 서론

능력기반 질의 처리기는 목적 질의와 데이터 소스의 질의 처리 능력을 입력받아 목적 질의와 동일한 결과를 생성하는 질의 수행 계획을 출력하는 시스템이다. 그러나 지금까지 발표된 질의 처리 시스템[1, 2, 3, 4]는 데이터 소스의 질의 처리 능력을 표현하고 질의 수행 계획을 세우는데 AND를 포함하는 질의만을 처리한다는 제한사항이 있었다. 실제 사용되고 있는 많은 데이터 소스는 AND 결합 질의 뿐 아니라 OR를 포함하는 질의도 처리하고 있다. 따라서 OR를 포함하고 있는 목적 질의는 데이터 소스에서 곧바로 처리될 수 있음에도 불구하고 AND 결합 질의 형태로 변형된 후 처리되고 그 결과를 상위 전역 시스템 단계에서 한번 더 가공해야 하는 비효율적인 면이 있었다.

본 논문에서는 이러한 문제를 해결하기 위하여 OR 연결 질의 표현법과 입력된 목적 질의에 대해 데이터 소스의 질의 처리 능력을 최대한 활용한 효과적인 질의 수행 계획 작성과정을 기술한다. 제안된 질의 처리기는 목적 질의와 래퍼의 처리 능력으로부터 생성된 지원 질의의 집합을 입력 사항으로 전달받아 최적의 질의 수행 계획을 출력한다. 논문의 구성은 다음과 같다. 2장에서는 기존에 연구된 질의 처리기를 정리하고 제안된 시스템과 차이점을 알아본다. 3장에서는 제안된 시스템의 질의 표현 방법과 질의 처리 능력을 이용한 질의 재처리 과정 알고리즘을 제시한다 4장에서는 제안된 시스템의 의의와 결론 및 향후 연구 방향에 대해 소개한다.

2. 관련연구

질의 처리 능력을 고려한 질의 처리 시스템으로는 TSIMMIS[1], Information Manifold[2], Garlic[3], DISCO[4], GenCompact[5] 등이 있다. 이 중 OR를 포함하는 질의를 처리할 수 있는 시스템으로는 DISCO와 GenCompact이 있다. 그러나 DISCO의 경우 래퍼의 질의 처리 능력을 연산자만으로 표현하기 때문에 다른 시스템에 비해 표현력이 떨어진다[9]. 따라서 이 장에서는 GenCompact 시스템의 OR 연결 질의 재작성에 대해 알아본다.

GenCompact에서 질의 재작성을 위한 입력 사항으로 목적 질의는 $Q = \pi_A (\sigma_C (R))$로 표시되고 이와 동일한 의미로 SP(C, A, R)를 사용하기도 한다. 여기서 질의 조건 C는 조건 트리(Condition Tree CT) 형태로 표현된다 CT의 마지막 노드는 부울린 조건이고 중간 노드들은 부울린 연결자(connector)로 AND와 OR를 사용하고 있어서 OR 연결 질의의 표현을 가능하게 하고 있다. 또 다른 입력 사항인 래퍼에서의 질의 처리 능력은 Context Free Grammar 형태로 표현된다. 이에 대한 표현은 [5]를 참고한다. GenCompact은 효율적인 질의 수행 계획을 세우기 위하여 제거 규칙(Pruning Rule)과 비용 함수(Cost Function)를 사용하고 있다 질의 재처리 과정을 통하여 생성되는 질의 수행 계획은 질의 트리(Query Tree QT)로 나타난

다 QT의 마지막 노드는 래퍼에서 처리할 수 있는 SP형태의 질의이고 중간 노드들은 상위 전역 시스템 단계에서 처리되는 실렉션, 프로젝션, 인터섹션, 유니온으로 구성된다

제안된 시스템과 GenCompact은 다음과 같은 차이점이 있다 GenCompact은 목적 질의를 질의 처리 능력을 바탕으로 생성 가능한 모든 CT로 표현하고, 각 CT의 노드마다 질의 재작성 비용을 구한다. 재작성 비용은 데이터 소스에 질의를 던지기 위한 오버헤드와 결과 데이터 전송 비용의 합이 된다. 마지막으로 재작성 비용을 이용한 제거 규칙을 통해 질의 수행 계획을 세운다. 반면, 제안된 시스템에서는 SQL 실렉션 구문 형태의 목적 질의와 지원 질의 집합을 입력받아 각 지원 질의의 기여도와 기여도를 만족시키기 위해 상위 전역 정보 시스템에서 추가로 처리해야할 나머지 조건을 비용으로 계산하고 이를 바탕으로 질의 수행 계획을 세운다. 따라서 모든 질의 처리 능력에 해당하는 CT를 생성하여 노드마다 비용을 계산해야 하는 GenCompact보다 질의 재처리 과정이 간단해진다.

3. OR 연결 질의의 재처리

2장에서는 처리 능력을 고려한 질의 처리 시스템들과 그 중에서 OR를 포함한 질의를 처리할 수 있는 질의 처리기에 대하여 살펴보았다. 이들 능력기반 질의 처리기에서 중요한 사항 중 하나는 질의 처리 능력에 대한 표현 방법이다 본 논문에서는 지면 관계상 질의 처리 능력에 대한 표현 방법은 생략하고 OR 연결 질의를 처리할 수 있는 질의 처리기의 질의 표현과 재처리 과정을 설명하겠다.

제안할 질의 처리 능력을 고려한 질의 처리기의 구조는 그림 1과 같다.

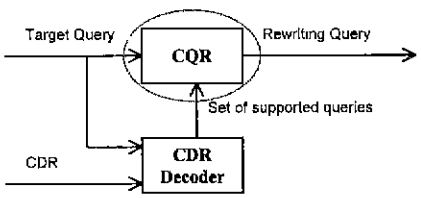

그림 1. 질의처리능력 기반 질의 처리기

질의 처리기(Capabilities-Based Query Rewriter : CQR)는 목적 질의와 래퍼의 질의 처리 능력으로부터 생성된 지원 질의 집합 SQ를 입력으로 받아 질의 수행 계획(Rewriting Query)을 생성한다.

3.1 질의의 표현

입력된 목적 질의 Q = π_A (σ_C (R))는 다음과 같은 형태로 표현되고 질의 조건 C는 아래와 같다.

[정의] target query Q = (A, C, R) where

C a selection condition

$\{c_1, c_2, \ldots, c_n\}$ c_1 AND c_2 AND ... AND c_n where c_i, $1 \le i \le n$

$\{c_{11}, c_{12}, \ldots, c_{1m}\}$ c_{11} OR c_{12} OR . OR c_{1m}

예를 들어 Q = ({A}, [(A=10), {(B=10), (C=10)}], {R})은 Select A From R Where (A=10) AND ((B=10) OR (C=10))을 의미한다

CQR의 또 다른 입력 사항인 지원 질의 집합 SQ는 각 래퍼의 처리 능력(Capabilities Description Record . CDR)을 CDR Decoder에서 입력받아 목적 질의와 릴레이션 매핑, 조건 매핑, 어트리뷰트 매핑을 통해 생성한 지원 질의들의 집합이다. 지원 질의도 목적 질의와 마찬가지로 sq = (A, C, R)로 표현된다.

3.2 질의 재작성

OR 연결 질의의 재작성은 목적 질의 Q = (A, C, R)와 SQ를 입력받아 목적 질의와 동일한 결과를 나타내는 아래의 Q'를 표현하는 것이다[6,7].

[정의]

$Q' = q_1$ OR q_2 OR . OR q_n, $1 \le i \le n$, q_i = (A, C', sq_i) where $sq_i \in SQ$

즉, sq로 이루어진 q_i를 유니온 한 결과인 Q'는 목적 질의 Q와 같다

그림 2는 질의 재작성 알고리즘이다. 첫 단계로 목적 질의로 입력된 Q를 DNF(Disjunctive Normal Form)로 변형하고 Q를 구성하는 서브 목적 질의 TQ를 구한다. 둘째 입력된 지원

Input *: target query Q*
set of supported queries SQ
Output. *rewriting target query Q'*

Method
Convert Q to DNF form Q = q_1 OR q_2 OR . . OR q_n
Let TQ = set of q_i, where $1 \le i \le n$

For each sq in SQ
For each q in TQ
Calculate the contribution and remaining condition of each sq for q

For each sq in SQ
Remove sq having ϕ contribution

Classify sqs into same contribution using their contribution

Find the minimum contribution set

Construct Q' from the minimum cost sqs which are selected from their contribution class

그림 2. 질의 재작성 알고리즘

질의 각각에 대해 TQ에 대한 기여도(Contribution)와 기여도를 만족시키기 위하여 상위 전역시스템 단계에서 수행해야 할 추가적인 조건을 구한다. 기여도란 지원 질의 sq가 주어진 TQ를 처리할 수 있는 정도로 sq = (A, C, R), q_i = (A, C, R)이라 할 때 아래와 같이 정의된다.

[정의] Contribution ' maximum subset S of sq.R for q_i where
for all $x \in$ ((attributes of q_i.A) $\cap$ (attributes of S)), $x \in sq.P+$
for all $y \in$ ((attributes of (q_i.C-sq.C) $\cap$ (attributes of S)), $y \in sq.P-$

for all z ∈ ((attributes of sq.C) ∩ (attributes of (sq.R-S))), z ∈ sq.P+

*sq.P+*은 확장된 어트리뷰트의 집합으로 sq.A와 sq.C를 구성하는 어트리뷰트들로 이루어져 있다. 위의 정의는 *sq*에서 q_i의 S에 해당하는 실렉션 항목과 조건을 모두 만족시키고 S에서의 조인 항목을 제공하는 sq.R의 가장 큰 서브 집합 S를 찾는 것이다. 예를 통해 기여도를 설명하기로 한다.

[예제] *Q = ((A), [(A=10), ((B=10), (C=10))], (R)) = q1 OR q2 where*

q_1 = (A, [(A=10), (B=10)], (R))

q_2 = (A, [(A=10), (C=10)], (R))

sq_1 = ((A,B,C), [(B=10)], (R))

SQ	Contribution Class	Remaining Condition
sq_1	{R}/{q_1}	[(A=10)]
sq_2	{R}/{q_2}	[(A=10)]
sq_3	{R}/{q_1, q_2}	ϕ

sq_2 = ((A,B,C), [(C=10)], (R))

sq_3 = ((A,B,C), [(A=10), ((B=10), (C=10))], (R))

sq_1의 기여도는 q_1의 R에 대한 실렉션 항목과 조건을 만족하고 있으므로 {R}/{q_1}이고 나머지 조건은 [(A=10)]이다. 즉, sq_1은 추가 조건 (A=10)이 주어지는 경우 q_1의 릴레이션 R을 만족시킬 수 있다는 의미이다. sq_3는 TQ인 q_1과 q_2를 나머지 조건이 없어도 모두 만족시키는 것을 알 수 있다. 셋째 기여도가 없는 *sq*들은 작업 범위에서 삭제한다 이 과정을 통해 질의를 재 작성하기 위해 필요한 뷰의 범위를 줄일 수 있다. 넷째 앞에서 생성한 기여도를 분석하여 동일한 기여도를 갖는 *sq*들을 같은 그룹으로 분리한다. 다섯째 단계는 기여도 그룹의 릴레이션을 유니온 한 결과가 목적 질의 TQ의 릴레이션을 모두 만족시키는 가장 작은 기여도 집합 (Minimal Contribution Set : MCS)을 찾는다. 위의 예에서 {R}/{q_1}과 {R}/{q_2}를 q_i에 대해 유니온 한 결과와 {R}/{q_1, q_2}가 TQ의 릴레이션을 만족시키고 있으므로 최소 기여도 집합은

MCS = {{{R}/{q_1}, {q_2}, {q_1, q_2}}}

이다 여섯 번째 단계는 MCS 중에서 가장 비용이 적은 *sq*들로 질의 수행 계획을 세우는 것이다. 여기서 가장 작은 비용이란 레퍼에게 요구해야 하는 질의 즉, *sq*의 수행회수가 적고 기여도를 생성하기 위하여 상위 전역 시스템 단계에서 추가되는 나머지 조건이 가장 작은 것을 말한다. 앞의 예제에서는 아래와 같은 2가지 질의 수행 계획이 생성 가능 하지만

Q_1' = ((A), [(A=10)], ($sq_1 \cup sq_2$))

Q_2' = ((A), [], (sq_3))

Q_1'은 Q_2'보다 더 많은 *sq*와 나머지 조건을 사용하고 있으므로 최종적으로 출력되는 질의 수행 계획은 Q_2'가 된다.

이상으로 OR 연결 질의를 처리하는 능력기반 질의 처리기의 질의 표현 방법과 재처리 알고리즘을 살펴보았다 제안된 시스템은 첫째 질의 최적화 규칙(query optimization)[8]을 사용하여 질의를 재작성 하기 때문에 처리단계가 간단하고, 둘째 기여도를 이용해 질의 작성에 필요한 탐색 공간을 줄여 재처리 과정의 복잡도(complexity)를 줄일 수 있었다 또, 최소의 sqs와 나머지 조건을 사용해 효율적인 수행 계획 생성하고 있다.

4. 결론 및 향후 연구방향

본 논문에서는 OR를 포함한 질의를 표현하는 방법과 능력기반 질의 처리기에서 래퍼의 질의 처리 능력을 이용한 질의 재처리 과정을 설명하였다. 사용자는 제안된 질의 처리기를 통하여 다양한 데이터 소스에 원하는 질의를 던질 수 있다. 또한 시스템 내에서는 목적 질의로 입력된 OR 연결 질의를 처리하기 위하여 기존 시스템들이 수행하던 질의 분리나 질의 결과 통합과 같은 불필요한 작업을 없앨 수 있다 특히 제안된 능력기반 질의 처리기에서는 질의 재처리 과정을 효과적으로 처리하기 위하여 *sq*의 목적 질의에 대한 질의 처리 능력 즉, 기여도를 측정하였고 그 과정에서 생성되는 비용을 계산함으로써 기존 시스템보다 짧은 시간 내에 최적화 된 질의 수행 계획을 세울 수 있었다.

본 논문에서 발표한 질의 표현 방법과 재처리 과정을 이용하여 향후에는 멀티 소스에서 OR를 포함한 질의 재처리 문제에 대하여 연구를 할 계획이다.

참조문헌

[1] Y. Papakonstantinou, A. Gupta, and L Haas, "Capabilities-based query rewriting in mediator systems" , *IEEE Int. Conf. On Parallel and Distributed Information Systems*, 1996

[2] Alon Y. Levy, A.Rajaraman and Joann J. Ordille, "Querying Heterogeneous Information sources Using Source Descriptions" , *the 22nd VLDB Conf.* India, 1996.

[3] V. Vassalos, Y. Papakonstantinou, "Describing and Using Query Capabilities of Heterogeneous Sources" , *the 23rd VLDB Conference*, 1997

[4] Olga Kapitskaia, Anthony Tomasic, Patrick Valduriez, "Dealing with Discrepancies in Wrapper Functionality" , *TR*. INRIA, 1997

[5] Hector Garcia-Molina, Wilburt Labio, Ramana Yerneni "Capability Sensitive Query Processing on Internet Sources" , ICDE 1999

[6] A Levy, A. Mendelzon, Y. Sagiv, and D Srivastava, "Answering queries using views". *Int. Conf. on Principles of Database Systems*, 1995.

[7] A. Levy, A. Rajaraman, and J. Ullman, "Answering queries using limited external query processors" , *Int. Conf. On Principles of Database Systems*, 1996

[8] J.D. Ullman, Principles of Database and Knowledge-base Systems Vol II, *Computer Science Press*, 1989.

[9] 김지운, "전역 정보 시스템에서의 처리 능력 표현과 질의 분해", 석사학위논문, 포항공과대학교, 1998