인공지능 기술을 활용한 데이터 관리 기술 동향

Trends in Data Management Technology Using Artificial Intelligence

김창수 (C.S. Kim, cskim7@etri.re.kr) 스마트데이터연구실 책임연구원
박춘서 (C.S. Park, parkcs@etri.re.kr) 스마트데이터연구실 책임연구원
이태휘 (T.W. Lee, taewhi@etri.re.kr) 스마트데이터연구실 책임연구원
김지용 (J.Y. Kim, kjy@etri.re.kr) 스마트데이터연구실 책임연구원/실장

ABSTRACT

Recently, artificial intelligence has been in the spotlight across various fields. Artificial intelligence uses massive amounts of data to train machine learning models and performs various tasks using the trained models. For model training, large, high-quality data sets are essential, and database systems have provided such data. Driven by advances in artificial intelligence, attempts are being made to improve various components of database systems using artificial intelligence. Replacing traditional complex algorithm-based database components with their artificial-intelligence-based counterparts can lead to substantial savings of resources and computation time, thereby improving the system performance and efficiency. We analyze trends in the application of artificial intelligence to database systems.

KEYWORDS 데이터베이스, 머신러닝, 비지도 학습, 제로샷 학습, 지도 학습

Ⅰ. 서론

데이터가 폭증하는 빅데이터 시대에 컴퓨팅 시스템의 발전으로 대규모 데이터에 대한 효율적인 처리가 가능해지면서 인공지능 기술의 한 부류인 머신러닝 기술의 발전이 가속화되고 있다. CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), LSTM(Long Short Term Memory) 및 GPT(Generative Pre-trained Transformer) 모델의 등장으로 머신러닝 기술은 급속히 발전하고 있으며, 다양한 문제를 해결하는 데 활용되고 있다.

빅데이터 시대에 대량의 데이터를 효율적으로 저장하고 관리할 수 있는 데이터 관리 플랫폼이 발전하였고, 인공지능 시대의 도래에 발맞춰 머신러닝

* DOI: https://doi.org/10.22648/ETRI.2023.J.380603
* 본 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[No. 2021-0-00180, 다양한 산업 분야 활성화 증대를 위한 분산 저장된 대규모 데이터 고속 분석 기술 개발].
* 본 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[No. 2021-0-00231, 빅데이터 대상의 빠른 질의 처리가 가능한 탐사 데이터 분석 지원 근사질의 DBMS 기술 개발].

모델의 학습을 위해 양질의 데이터를 제공하는 플랫폼으로서의 역할을 수행하고 있다.

대표적으로 오랫동안 중요한 데이터의 저장 및 관리를 담당해오고 있는 데이터베이스는 인공지능 모델 학습을 위한 학습 데이터를 제공하는 데이터 관리 플랫폼 역할을 수행하고 있다. 한편, 발전된 인공지능 기술을 이용해 데이터베이스 시스템을 향상시키는 연구도 활발히 진행되고 있다[1-5].

데이터베이스 시스템의 인공지능 기술 활용은 데이터베이스 시스템을 구성하는 다양한 구성요소들(예: 인덱스, 카디널리티 예측, 실행비용 예측, 실체화된 뷰 관리, 질의 최적화 등)을 적절한 머신러닝 모델로 대체하는 형태를 갖는다. 머신러닝 모델을 기반으로 하는 데이터베이스 시스템 구성요소는 기존의 휴리스틱 기반 알고리즘을 바탕으로 하는 복잡한 컴포넌트를 대체하여 시간과 비용이 많이 소요되는 작업을 감소시키고 시스템 성능과 효율을 높이는 데 기여한다. 데이터베이스 관리자가 주로 경험과 통계, 휴리스틱을 통해 수동으로 진행하는 데이터베이스 패러미터 튜닝이나 인덱스 생성 및 관리, 실체화된 뷰 관리 등의 작업은 인공지능 기술을 통해 자동화함으로써 관리 비용을 현저히 절감하고 성능을 향상시킬 수 있다. 질의 처리 계획이나 스케줄러의 자동화, 인덱스의 머신러닝 모델화 등은 시스템 성능을 향상시키는 데 효과적이다[1].

본고에서는 데이터베이스 시스템 성능 향상을 위한 인공지능 기술의 활용에 대한 동향을 살펴보고 앞으로의 발전 방향에 대하여 고찰해보고자 한다.

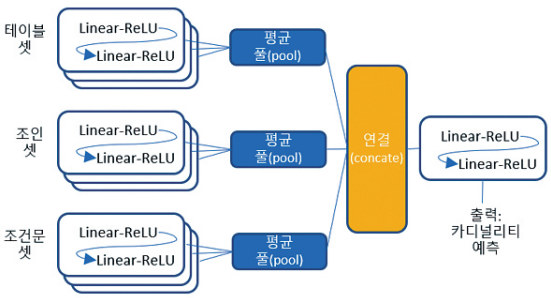
본고의 구성은 다음과 같다. 먼저 Ⅱ장에서는 워크로드 기반 지도 학습을 활용한 데이터베이스 컴포넌트 기술에 대하여 간략히 살펴본다. Ⅲ장에서는 데이터나 워크로드의 변화에 따라 새롭게 학습 데이터를 준비하고 학습해야 하는 워크로드 기반 지도 학습 방법의 문제점을 해결하기 위한 데이터 기반 비지도 학습을 활용한 데이터베이스 컴포넌트 기술에 대하여 살펴보고, Ⅳ장에서는 데이터 기반 비지도 학습의 한계를 극복하기 위한 제로샷 학습 기반 기술에 대하여 살펴본다. Ⅴ장에서는 한국전자통신연구원에서 다양한 인공지능 모델을 활용하여 실제 데이터베이스 시스템에 적용하고 있는 기술 연구 동향을 소개하고, Ⅵ장에서 결론 및 향후 연구 방향을 기술한다.

Ⅱ. 워크로드 기반 지도 학습 기술

데이터베이스 컴포넌트를 위한 워크로드 기반 지도 학습 기술은 학습 단계에서 대량의 학습 워크로드를 수집하고 대상 데이터베이스를 기반으로 실제 워크로드를 실행하여 그 결과를 함께 학습 데이터로 활용하여 머신러닝 모델을 학습한다. 모델 학습이 완료되면 실제 환경에서 워크로드를 모델에 입력하여 모델이 예측하는 결과를 활용하는 형태이다.

1. MSCN(Multi-Set Convolutional Network)

MSCN[2]은 2019년 발표된 카디널리티 예측을 위한 딥러닝 접근방법이다. 데이터베이스 시스템에서 카디널리티 예측은 질의를 수행하는 과정에서 질의 처리 중간 결과 및 최종 결과의 튜플 개수를 예측하는 것으로, 다양한 조인 순서 중 최적의 순서를 정하는 등의 질의 최적화에 널리 활용되고 있는 중요한 요소이다. 기존의 통계에 기반한 예측 알고리즘은 종종 부정확한 결과를 초래하여 질의 처리 성능을 저하하는 요인으로 지적되어왔다. 이러한 부정확한 예측 알고리즘을 개선하기 위하여 컨볼루션 뉴럴 넷(CNN: Convolutional Neural Network)을 활용하여 다양한 조인 질의에 대하여 학습을 수행하



출처 Reproduced from [2], CC BY 3.0.

그림 1 MSCN의 모델

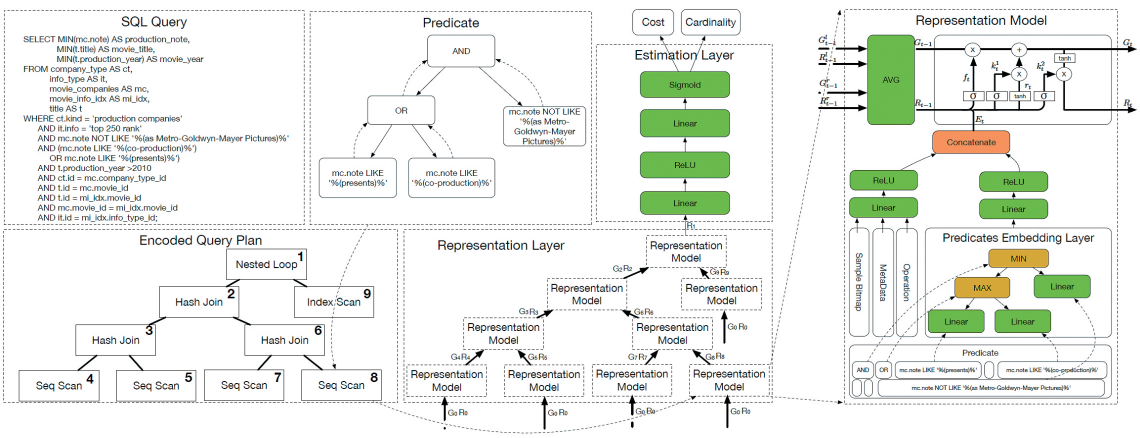
여 모델을 정립하고, 해당 모델을 기반으로 실제 질의에 대하여 카디널리티 예측을 수행한다. MSCN은 조인 질의를 지원하기 위해 그림 1에서 보는 것처럼 테이블 셋, 조인 셋, 조건문 셋 모듈을 별도로 구성하여 선형모델을 구성하고 풀링을 수행한 후 각 모듈의 결과를 합쳐 출력하는 형태를 취하고 있다. MSCN은 여러 테이블에 걸친 조인 연산을 효과적으로 학습하면서 예측 성능을 높인다.

2. 트리 구조 모델 기반 비용 예측

칭화대학은 카디널리티 예측과 비용 예측을 동시에 수행할 수 있는 트리 구조 기반의 학습 모델을 제안했다[3]. 이 방법에서는 학습 데이터 생성기를 통해 데이터 셋의 다양한 조인 그래프에 따라 다양한 질의를 생성한다. 학습용 질의들에 대하여 질의 계획과 실제 수행한 결과의 비용 및 카디널리티 값을 토대로 모델을 학습한다. 모델은 크게 표현 레이어와 예측 레이어 두 개의 레이어로 구성되어 있다(그림 2).

입력된 질의 계획, 비용과 카디널리티를 사용하여 질의 조건, 질의 연산자, 데이터 샘플 등을 벡터로 인코딩하여 각각을 선형모델로 구성하고 이들을 연결하여 LSTM에 입력한다. LSTM을 포함한 이러한 구조들은 표현모델(Representation Model)을 구성하고 이러한 표현모델들은 질의 계획에 따라 트리 구조를 형성하여 표현 레이어(Representation Layer)를 구성한다.

표현 레이어의 출력값은 예측 레이어의 입력



출처 Reprinted from [3], CC BY-NC-ND 4.0.

그림 2 워크로드 기반 예측을 위한 트리구조 모델

이 되며, 선형모델인 예측 레이어는 실행 비용과 카디널리티를 출력하는 구조이다.

3. 워크로드 기반 지도 학습의 한계점

워크로드 기반의 지도 학습 모델은 많은 양의 질의를 통해 학습 데이터를 획득해야 하며, 이 과정은 긴 시간이 소요된다. 또한, 데이터 변경 시 해당 데이터에 적응하기 위해 다시 학습해야 하는데, 이때도 많은 학습 질의를 다시 실행해야 하는 문제가 발생한다. 따라서 새로운 데이터베이스나 워크로드에 일반화하기가 어려워, 기술의 적용이 고정된 데이터와 워크로드를 유지하는 특정 환경으로 제한되는 문제가 있다.

Ⅲ. 데이터 기반 비지도 학습 기술

워크로드 기반 지도 학습의 문제점인 학습 데이터 수집을 위한 높은 비용과 데이터 및 워크로드 변화에 대한 낮은 융통성 문제를 해결하기 위해 워크로드를 학습하는 대신 데이터 자체를 학습하는 비지도 학습이 연구되었다. 워크로드 기반 지도 학습 모델이 워크로드를 카디널리티에 매핑하는 함수를 학습하는 것이라면, 데이터 기반 비지도 학습은 각각의 튜플을 테이블에서 나타날 확률로 매핑하는 확률적 모델에 바탕을 두고 있다. 데이터에 대한 확률 분포를 학습하면, 이를 활용하여 카디널리티를 예측하거나 집계 질의 등의 처리 결과를 예측할 수 있게 된다(그림 3).

대표적인 데이터 기반 비지도 학습 모델로 DeepDB[4]를 소개한다. DeepDB는 데이터베이스를 샘플링한 후 속성과 테이블들 간 데이터 분포를 학습한다. 학습된 데이터 분포를 활용하여 카디널리티 예측이나 근사 집계 질의를 수행할 수 있다.

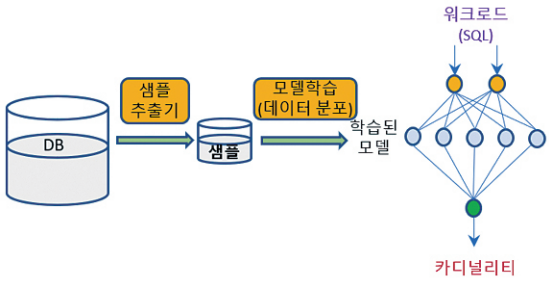
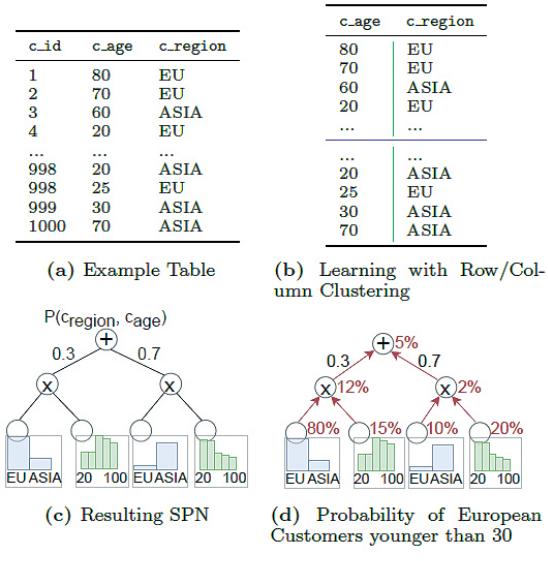


그림 3 데이터 기반 비지도 학습 모델

DeepDB는 데이터의 분포를 학습하였기에 학습 워크로드를 별도로 실행할 필요성을 제거하였다. 또한 데이터 셋의 변경이 발생할 경우, 해당 데이터에 대한 부분만 학습하므로 모델 업데이트가 빠르게 수행될 수 있는 장점이 있다.

DeepDB에서는 단일 테이블에 대한 질의를 위해 SPN(Sum Product Network) 모델을 활용한다(그림 4). SPN 모델은 테이블을 반복하여 독립적인 칼럼 클러스터와 행 클러스터로 나눈다. 칼럼 클러스터는 조



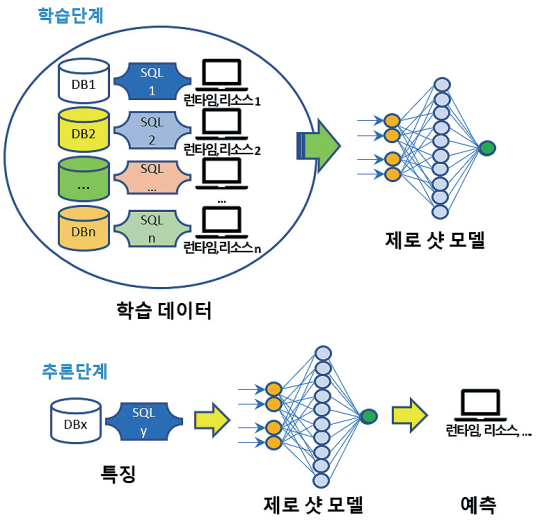
출처 Reprinted from [4], CC BY-NC-ND 4.0.

그림 4 테이블-SPN 모델

건을 처리하는 부분이 되므로 Product 노드로 연결되며, 행 클러스터는 결합되는 특성으로 Sum 노드로 연결된다.

DeepDB는 단일 테이블 지원을 다중 테이블로 확장하고 다른 중요한 관계형 데이터베이스 특징을 지원하기 위해 RSPN(Relational SPN)으로의 확장을 제공한다. 이러한 확장에는 Null 값 처리 지원, 데이터 분포 외에 가능한 한 정확한 값을 지원하기 위한 방법으로 값-빈도 저장, 속성 간 함수 종속성 반영을 지원한다. 또한, 다중 테이블 간 관계를 학습하기 위해 모든 외래키-키 관계에 대해 완전한 아우터 조인(Full Outer Join)에 대한 RSPN을 학습한다. 그리고 RSPN들을 연결하여 임의의 조인 연산을 지원할 수 있게 하였다.

DeepDB와 같은 데이터 기반 비지도 학습 방법은 데이터의 분포를 학습함으로써 다양한 서로 다른 워크로드를 지원할 수 있는 장점이 있다. 그러나 데이터의 스키마가 커질 경우, 가능한 모든 조인을 학습하려면 매우 방대한 저장 공간이 필요하므로 확장성에서 다소 제한이 있을 수 있다. 또한 런타임 환경, 자원 경합 등의 워크로드에 대한 정보가 필요한 질의는 지원하기에 한계가 있다는 제약이 있다.



출처 Reproduced from [6], CC BY 4.0

그림 5 데이터베이스를 위한 제로샷 모델 구조

Ⅳ. 제로샷(Zero-shot) 학습 기반 모델

자연어 처리를 위한 제로샷 학습을 가능하게 한 GPT-3[6]의 등장으로 기존의 많은 데이터를 학습하여 새로운 데이터를 생성할 수 있는 생성형 인공지능 및 전이 학습 기술이 최근 큰 인기를 끌고 있다. GPT와 같은 파운데이션 모델의 아이디어를 활용하여 많은 데이터베이스를 학습하여 일반화된 데이터베이스를 표현할 수 있는 모델이 있다면, 이는 다양한 데이터 특징을 갖는 테이블들로 구성된 아직 경험하지 못한 새로운 데이터베이스를 일반화할 수 있다. 제로샷 학습 방법[5]은 여러 데이터베이스를 대상으로 다양한 워크로드를 서로 다른 런타임 환경에 대하여 학습하고 제로샷 모델을 생성한다. 예를 들어, 질의 비용 모델을 위한 제로샷 모델을 그림 5에 나타냈다. 학습된 모델을 통해 새로운 데이터베이스와 질의 계획을 입력하면 런타임 비용에 대한 예측을 수행하는 비용 예측기로서 동작한다.

이때 고려해야 할 것은 다양한 데이터베이스들 사이에서 동일한 의미를 갖지만, 형식상 서로 다른 특성들(속성 이름, 속성 타입 등이 서로 다른 경우 등)을 동일하게 표현하는 방법이 필요하다. 제로샷 학습 방법[5]에서는 물리 연산자, 조건문, 테이블 이름, 칼럼 등을 그래프 노드로 인코딩하는 그래프 인코딩 방법을 제안했다. 또한, 데이터베이스 간 형식상 차이를 해소하고 동일하게 특징을 인식할 수 있도록 공통화된 표현방법으로 각 그래프 노드에 대해 원핫(One-hot) 인코딩과 튜플 수, 페이지 수 같은 일반화된 특징을 활용하였다.

학습은 각 그래프 노드의 특징들을 벡터로 인코

딩한 후, 그래프 노드들로 이루어진 DAG(Direct Acyclic Graph) 형태를 상향식(Bottom-up) 메시지 패싱을 통해 통합되는 형태로 루트 노드까지 학습하며 도달하는 형태로 이루어진다.

제로샷 학습이 충분히 새로운 데이터베이스를 일반화할 수 있으려면 얼마나 많은 학습이 필요한지를 결정하는 것은, 교차 검증을 통해 예측 성능이 수용 가능할 때까지 점차 늘리는 방법을 사용하였다. 테스트에 따르면 19개의 서로 다른 데이터베이스를 학습하고, 학습한 모델을 통해 다른 1개의 데이터베이스를 테스트하였을 때 충분히 만족할만한 성능을 보였다.

제로샷 학습 방법은 워크로드 기반 학습 방법과 데이터 기반 학습 방법을 혼합한 형태를 가짐으로 인해, 런타임 비용 문제 외에도 물리적 데이터베이스 설계, 런타임 파러미터 튜닝, 트랜잭션 스케줄링 등 다양한 분야에 적용할 수 있다. 하나의 예로 물리적 설계 문제 중 하나인 인덱스 설정 문제를 들 수 있다. 제로샷 모델은 인덱스 설정 문제를 해결하기 위해 학습 시 랜덤하게 특정 인덱스 셋을 생성하여 학습 질의를 수행하여 학습을 수행한다. 제로샷 모델은 인덱스를 포함한 질의에 대하여 물리 연산자가 변경되므로 이를 학습할 수 있게 되어, 특정 속성에 대한 인덱스 존재 여부에 따른 실행 비용을 비교할 수 있게 된다.

V. 데이터베이스 시스템을 위한 ETRI 인공지능 활용 기술

한국전자통신연구원은 기업의 데이터 분석과 비즈니스 의사 결정의 적시성을 확보하기 위한 노력의 일환으로, 머신러닝을 기반으로 한 DBMS 근사 질의 처리 엔진 기술을 개발하고 있다. 근사 질의 처리 기술 분야의 최신 동향을 살펴보면, 이전의 샘플링 또는 요약 기반 접근 방식에서 최근 발전한 머신러닝 기술을 활용하여 정확도와 성능을 향상시키는 방향으로 진화하고 있다. 이러한 노력은 데이터 분석과 의사 결정 프로세스에 새로운 가능성을 제공하며 산업 현장에서 더 나은 결과를 이끌어내고자 함이다.

머신러닝을 활용한 근사 질의 처리 기술은 모델 학습 방식에 따라 크게 워크로드 기반 모델 방식과 데이터 기반 모델 방식으로 구분되는데, 워크로드 기반 모델은 정확한 질의와 해당 결과를 학습하는 방식으로, 이를 위해 충분한 양의 질의 이력이 필요한 제한 사항이 존재하게 된다.

반면, 데이터 기반 모델 방식은 데이터 자체를 학습하는 방식으로, 질의 유형에 일반화되는 특징을 갖고 있으며, 질의 이력을 많이 확보해야 하는 부담을 해결할 수 있어서, 한국전자통신연구원은 데이터 중심 모델을 기반으로 하면서 질의 고려를 통해 성능을 향상시키는 방향으로 연구를 진행하고 있다.

그림 6에서와 같이 DBMS 근사 질의 처리 엔진은 다양한 DBMS에 적용 가능하도록 별도의 미들웨어 프레임워크 형태로 구성된다. 먼저 원시 DBMS의 데이터를 학습하여 머신러닝 모델을 구축한 다음, 이를 활용하여 직접 근사 결과를 추론하거나 합성 데이터인 데이터 시놉시스를 생성하여

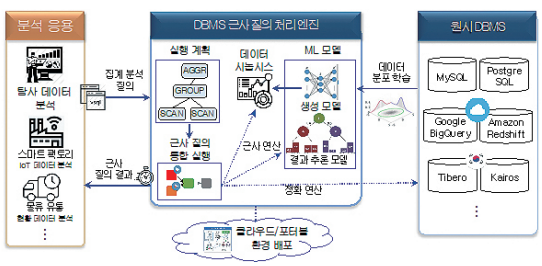


그림 6 ETRI DBMS 근사 질의 처리 엔진 기술 개념도

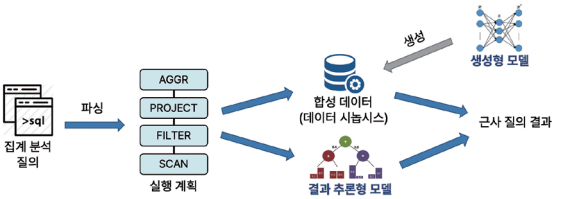
근사 결과를 제공하게 된다. 데이터 시놉시스는 원본 데이터의 특성을 보존하면서 축약된 형태로, 기존 샘플 데이터와는 달리 더 간결하게 정보를 포함할 수 있다.

한국전자통신연구원에서 개발하고 있는 시스템은 근사 질의 처리 엔진을 포함하여 정확한 질의 처리와 통합 실행을 지원하며, 활용성을 극대화하기 위해 원시 DBMS와 별도로 클라우드 및 포터블 환경 배포도 고려하여 개발하고 있다.

머신러닝 모델은 질의 결과 추론형과 시놉시스 생성형으로 구분할 수 있으며, 두 가지 모델은 상호 보완적인 역할을 한다. 결과 추론형 모델은 빠른 수행 시간으로 결과를 즉시 산출하며, 반면에 시놉시스 생성형 모델은 다양한 사용자 정의 함수를 지원하여 다양한 질의 유형을 처리할 수 있다. 이러한 방식으로 머신러닝 모델을 활용해 다양한 요구 사항에 대응하면서 질의 처리의 효율성과 다양성을 확보할 수 있다.

질의 결과 추론형 모델은 근사 질의 결과를 즉시 출력하는 모델로서, 데이터를 학습 가능한 형태로 인코딩하여 모델 학습을 수행한다. 이 모델은 테이블, 속성 등 데이터의 수에 따라 모델 수가 증가하며, 그룹, 조인 등이 포함된 복잡한 질의에 대한 지원을 위해서는 모델 수가 기하급수적으로 늘어날 수 있다. 따라서 테이블과 속성 간의 상관관계를 활용하여 모델을 조합하여 조인트 모델을 구축하는 방향으로 연구를 진행하고 있다.

데이터 시놉시스 생성형 머신러닝 모델의 핵심은 시놉시스를 생성할 때 원본 데이터와 유사한 값을 가지면서 더 작고 효과적인 시놉시스를 만드는 것이다. 또한, 원본 데이터의 변화를 반영하는 연속 학습 기술과 모델 경량화 기술을 개발하고 있으며, 이를 통해 모델 학습 성능을 향상시키고 클라우드 및 포터블 환경에서의 근사 질의 시스템을 지원하고자 한다.

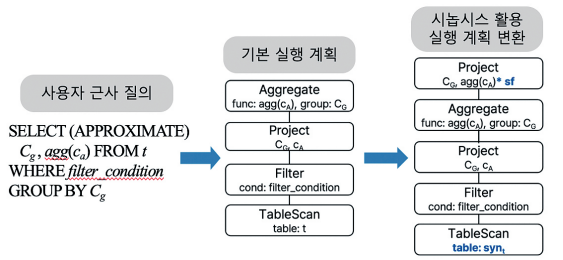


출처 Reproduced with permission from [7].

그림 7 머신러닝 모델을 사용한 근사 질의 처리 과정

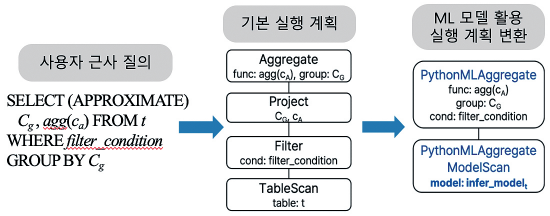
그림 7[7]은 머신러닝 모델을 사용한 근사 질의 처리 과정을 개략적으로 나타낸 것으로, 입력 질의는 질의 엔진에서 파싱(Parsing)되고 초기 실행 계획으로 변환된다. 만약 집계 질의의 대상 테이블에서 훈련된 특정한 머신러닝 모델이 존재한다면, 해당 질의는 질의에서 지정된 원본 테이블에 액세스하지 않고 근사 질의 처리될 수 있다. 근사 질의 처리에 사용될 수 있는 머신러닝 모델로 시놉시스 생성 모델과 결과 추론형 모델이 존재하는데, 이러한 모델을 일관된 방식으로 질의 처리에 활용하기 위해서는 질의 실행 계획에서 각 모델을 활성화하기 위한 근사 질의 변환 규칙이 각각 필요하다.

시놉시스 생성 모델 방법은 생성된 시놉시스 데이터를 근사 질의 처리에 활용하는 것으로, 시놉시스 데이터는 미리 생성될 수도 있고 런타임에서 생



출처 Reproduced with permission from [7].

그림 8 시놉시스 데이터 사용을 위한 질의 변환



출처 Reproduced with permission from [7].

그림 9 추론 모델 사용을 위한 질의 변환

성될 수도 있다. 그림 8과 같이 일반적인 질의 실행 계획을 시놉시스 데이터를 사용하여 근사 질의를 처리하는 실행 계획으로 변환할 수 있다. 변환된 시놉시스 활용 질의 실행 계획에서는 원래 테이블 대신 시놉시스 데이터 테이블을 액세스하여 근사 집계 값을 계산하는 방식으로 질의를 수행한다. 이때 원본 테이블과 시놉시스 데이터 테이블의 크기가 다르기 때문에 스케일 팩터(Scale Factor)를 곱하기 위한 새로운 프로젝션 노드가 추가된다.

결과 추론형 모델 방법은 질의 결과를 직접 예측하여 근사 질의 처리에 활용하는 방법이다. 예를 들어, DeepDB[4]의 SPN 모델을 사용하여 질의 결과를 예측할 수 있다. 이러한 훈련된 추론 모델을 사용하기 위해 그림 9와 같이 일반적인 실행 계획을 추론 모델을 사용하도록 변환하여 근사 질의 처리를 수행할 수 있다. 시놉시스 활용 실행 계획과 달리, 실행 계획이 머신러닝 모델을 실행하는 새로운 물리적 실행 노드를 포함하는 형태로 변환된다. 이 방법은 추론 모델로부터 직접 근사 집계 값을 제공할 수 있지만 특정한 집계 함수만을 지원할 수 있다[7].

VI. 결론

데이터베이스 시스템은 정보기술 환경에서 중요한 역할을 수행해왔다. 최근의 인공지능 기술의 발전을 위한 학습 데이터를 제공하는 기능을 담당해왔고, 현재 인공지능 기술을 이용한 데이터베이스 시스템 개선 방안이 연구되고 있다. 본고에서는 최근의 머신러닝 기술을 활용하여 데이터베이스 시스템 성능을 높이는 연구 동향에 대하여 살펴보았다.

워크로드를 기반으로 머신러닝 모델을 학습하고 학습된 모델을 활용하여 카디널리티 예측, 실행비용 예측 등을 수행하는 워크로드 기반 지도 학습 방식은 가볍고 빠른 장점이 있는 반면, 데이터나 워크로드가 변할 경우, 매우 많은 양의 학습 데이터를 수집하고 재학습해야 하는 단점을 갖고 있다. 이에 반해 데이터 자체를 학습하여 데이터에 대한 확률적 분포를 학습하는 데이터 기반 비지도 학습 방법은 데이터의 변화에 빠르게 대응할 수 있는 장점이 있다. 그러나, 큰 데이터베이스에 대한 확장성이 부족하고 워크로드 정보가 요구되는 태스크에는 적용할 수 없는 문제가 있다. 마지막으로, 파운데이션 모델에 기반한 제로샷 기반 방식은 데이터와 워크로드를 함께 학습하여 다양한 태스크에 적용할 수 있는 방법으로 소개하였다.

인공지능 기술의 발전과 더불어 이를 데이터베이스 시스템에 적용하는 기술은 앞으로도 계속해서 함께 성장할 것으로 기대되고 있다. 현재 적용되는 기술은 (1) 최근 데이터 경향이나 특정 도메인의 데이터를 반영하기 위한 퓨샷 학습법의 적용, (2) 데이터 독립적인 제로샷 모델 이외에 태스크 독립적인 학습 방법의 고려, (3) 데이터베이스의 다양한 특징 중 어떤 것을 선택할지에 대한 문제의 해결, (4) 데이터베이스 시스템의 다양한 구성요소에 따라 서로 다른 최적 모델을 융합하고 관리하는 방법이나 통합된 모델의 도출 등이 있다. 이상과 같은 연구 방향으로 계속 성장할 것으로 전망한다.

용어해설

카디널리티(Cardinality) 집합에서 집합의 요소 수를 나타냄. DBMS에서는 질의를 수행할 때 중간 결과 또는 최종 결과 집합의 요소 수를 지칭함

제로샷 학습(Zero-shot Learning) 많은 학습 데이터를 이용하여 학습한 모델을 사용하여, 학습 데이터에 포함되지 않은 새로운 클래스의 데이터에 대한 예측이 가능하도록 학습하는 기술. 특정 분야의 학습 데이터로 학습된 모델을 재학습 없이 다른 분야의 데이터에 대해 추론이 가능한 기술

근사 질의 처리(Approximate Query Processing) 정확한 질의를 수행하기 위해서는 많은 자원과 시간이 소요되므로, 샘플링/요약정보/머신러닝 등을 활용하여 질의 결과를 근사하여 적시에 실행하는 기술

약어 정리

CNN	Convolutional Neural Network
DBMS	DataBase Management System
GPT	Generative Pre-trained Transformer
LSTM	Long Short Term Memory
MSCN	Multi-Set Convolutional Network
RNN	Recurrent Neural Network
RSPN	Relational Sum Product Network
SPN	Sum Product Network

참고문헌

[1] G. Li and X. Zhou, "Machine learning for data management: A system view," in Proc. IEEE ICDE, (Kuala Lumpur, Malaysia), May 2022, pp. 3198-3201.

[2] A. Kipf et al., "Learned cardinalities: Estimating correlated joins with deep learning," in Proc. CIDR, (Asilomar, CA, USA), Jan. 2019, http://cidrdb.org/cidr2019/papers/p101-kipf-cidr19.pdf

[3] J. Sun and G. Li, "An An end-to-end learning-based cost estimator," Proc. VLDB Endow, vol. 13, no. 3, 2019, pp. 307-319, https://doi.org/10.14778/3368289.3368296

[4] B. Hilprecht et al., "DeepDB: Learn from data, not from queries!," Proc. VLDB Endow, vol. 13, no. 7, 2020, pp. 992-1005, https://doi.org/10.14778/3384345.3384349

[5] B. Hilprecht et al., "One model to rule them all: Towards zero-shot learning for databases," in Proc. CIDR, (Mineola, NY, USA), Jan. 2022, https://www.cidrdb.org/cidr2022/papers/p16-hilprecht.pdf

[6] T. Brown et al., "Language models are few-shot learners," NeurIPS 2020, vol. 33, 2020, pp. 1877-1901.

[7] T. Lee et al., "Exploiting machine learning models for approximate query processing," in Proc. Big Data, (Osaka, Japan), Jan. 2022, pp. 6752-6754.