(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0102382
(43) 공개일자 2023년07월07일

(51) 국제특허분류(Int. Cl.)
G06F 16/2452 (2019.01) *G06F 16/21* (2019.01)
G06F 16/22 (2019.01) *G06F 16/2453* (2019.01)
G06F 16/28 (2019.01) *G06N 20/00* (2019.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06F 16/24522 (2019.01)
G06F 16/211 (2019.01)
(21) 출원번호 10-2021-0192460
(22) 출원일자 2021년12월30일
심사청구일자 2021년12월30일

(71) 출원인
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
한욱신
경상북도 포항시 남구 청암로 77
강혁규
경상북도 포항시 남구 청암로 77
김현지
울산광역시 동구 월봉12길 50, C동 308호
(74) 대리인
특허법인이상

전체 청구항 수 : 총 18 항

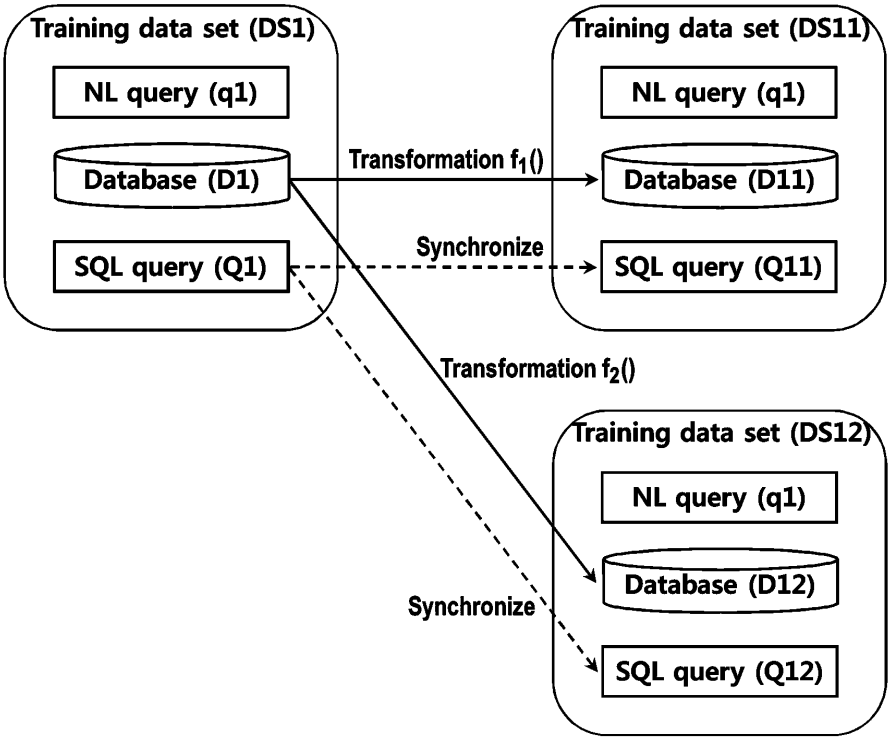
(54) 발명의 명칭 **자연어 처리 시스템의 학습 데이터셋 증강 방법**

(57) 요 약

소정의 신경망을 토대로 자연어 질의를 SQL 질의로 번역하는 번역기를 포함하는 자연어 처리 시스템에서 상기 신경망의 학습에 사용되는 학습 데이터셋을 증강시키는 방법이 제공된다. 학습 데이터셋 증강 방법은 제1 학습용 자연어 질의와, 상기 데이터베이스에 관한 데이터베이스 정보와, 상기 제1 학습용 자연어 질의에 상응하는 제1

(뒷면에 계속)

대 표 도 - 도4



학습용 SQL 질의를 포함하는 제1 초기 학습 데이터셋을 결정하는 단계; 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제1 신규 데이터베이스에 대한 제1 신규 데이터베이스 정보를 생성하는 단계; 상기 제1 스키마 변형 연산에 대응하는 제1 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제1 신규 데이터베이스에 대한 제1 신규 학습용 SQL 질의를 생성하는 단계; 및 상기 제1 학습용 자연어 질의와, 상기 제1 신규 데이터베이스 정보와, 상기 제1 신규 학습용 SQL 질의를 포함하는 제1 신규 학습 데이터셋을 결정하는 단계;를 포함한다.

(52) CPC특허분류
G06F 16/2228 (2019.01)
G06F 16/24534 (2019.01)
G06F 16/284 (2019.01)
G06N 20/00 (2021.08)
G06N 3/08 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711126201
과제번호	2018-0-01398-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발(R&D,정보화)
연구과제명	(SW 스타랩) 대화 가능하고 자동으로 튜닝하는 DBMS의 개발
기 여 율	1/1
과제수행기관명	포항공과대학교 산학협력단
연구기간	2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

소정의 신경망을 토대로 자연어 질의를 SQL 질의로 번역하는 번역기를 포함하는 자연어 처리 시스템에서, 상기 신경망의 학습에 사용되는 학습 데이터셋을 증강시키는 방법으로서,

제1 학습용 자연어 질의와, 상기 데이터베이스에 관한 데이터베이스 정보와, 상기 제1 학습용 자연어 질의에 상응하는 제1 학습용 SQL 질의를 포함하는 제1 초기 학습 데이터셋을 결정하는 단계;

제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제1 신규 데이터베이스에 대한 제1 신규 데이터베이스 정보를 생성하는 단계;

상기 제1 스키마 변형 연산에 대응하는 제1 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제1 신규 데이터베이스에 대한 제1 신규 학습용 SQL 질의를 생성하는 단계; 및

상기 제1 학습용 자연어 질의와, 상기 제1 신규 데이터베이스 정보와, 상기 제1 신규 학습용 SQL 질의를 포함하는 제1 신규 학습 데이터셋을 결정하는 단계;

를 포함하는 학습 데이터셋 증강 방법.

청구항 2

청구항 1에 있어서,

제2 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제2 신규 데이터베이스에 대한 제2 신규 데이터베이스 정보를 생성하는 단계;

상기 제2 스키마 변형 연산에 대응하는 제2 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제2 신규 데이터베이스에 대한 제2 신규 학습용 SQL 질의를 생성하는 단계; 및

상기 제1 학습용 자연어 질의와, 상기 제2 신규 데이터베이스 정보와, 상기 제2 신규 학습용 SQL 질의를 포함하는 제2 신규 학습 데이터셋을 결정하는 단계;

를 더 포함하는 학습 데이터셋 증강 방법.

청구항 3

청구항 2에 있어서,

제2 학습용 자연어 질의와, 상기 데이터베이스 정보와, 상기 제2 학습용 자연어 질의에 상응하는 제2 학습용 SQL 질의를 포함하는 제2 초기 학습 데이터셋을 결정하는 단계;

상기 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제3 신규 데이터베이스에 대한 제3 신규 데이터베이스 정보를 생성하는 단계;

상기 제1 SQL 질의 동기화 연산을 상기 제2 학습용 SQL 질의에 적용하여 상기 제3 신규 데이터베이스에 대한 제3 신규 학습용 SQL 질의를 생성하는 단계; 및

상기 제2 학습용 자연어 질의와, 상기 제3 신규 데이터베이스 정보와, 상기 제3 신규 학습용 SQL 질의를 포함하는 제3 신규 학습 데이터셋을 결정하는 단계;

를 더 포함하는 학습 데이터셋 증강 방법.

청구항 4

청구항 3에 있어서,

상기 제1 스키마 변형 연산을 상기 제1 신규 데이터베이스에 적용하여 상기 제1 신규 데이터베이스와 다른 스키

마를 가지는 제4 신규 데이터베이스에 대한 제4 신규 데이터베이스 정보를 생성하는 단계;

상기 제1 SQL 질의 동기화 연산을 상기 제1 신규 학습용 SQL 질의에 적용하여 상기 제4 신규 데이터베이스에 대한 제4 신규 학습용 SQL 질의를 생성하는 단계; 및

상기 제1 학습용 자연어 질의와, 상기 제4 신규 데이터베이스 정보와, 상기 제4 신규 학습용 SQL 질의를 포함하는 제4 신규 학습 데이터셋을 결정하는 단계;

를 더 포함하는 학습 데이터셋 증강 방법.

청구항 5

청구항 1에 있어서,

제2 학습용 자연어 질의와, 상기 데이터베이스 정보와, 상기 제2 학습용 자연어 질의에 상응하는 제2 학습용 SQL 질의를 포함하는 제2 초기 학습 데이터셋을 결정하는 단계;

상기 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제2 신규 데이터베이스에 대한 제2 신규 데이터베이스 정보를 생성하는 단계;

상기 제1 스키마 변형 연산에 대응하는 상기 제1 SQL 질의 동기화 연산을 상기 제2 학습용 SQL 질의에 적용하여 상기 제2 신규 데이터베이스에 대한 제2 신규 학습용 SQL 질의를 생성하는 단계; 및

상기 제2 학습용 자연어 질의와, 상기 제2 신규 데이터베이스 정보와, 상기 제2 신규 학습용 SQL 질의를 포함하는 제2 신규 학습 데이터셋을 결정하는 단계;

를 더 포함하는 학습 데이터셋 증강 방법.

청구항 6

청구항 1에 있어서, 상기 제1 스키마 변형 연산은 상기 데이터베이스의 스키마 구조를 변경하는 제1 변형 연산 및 상기 데이터베이스의 스키마 요소의 이름을 변경하는 제2 변형 연산 중에서 적어도 하나를 포함하는 학습 데이터셋 증강 방법.

청구항 7

소정의 신경망을 토대로 자연어 질의를 SQL 질의로 번역하는 번역기를 포함하는 자연어 처리 시스템에서, 상기 신경망을 학습시키는 방법으로서,

제1 학습용 자연어 질의와, 상기 데이터베이스에 관한 데이터베이스 정보와, 상기 제1 학습용 자연어 질의에 상응하는 제1 학습용 SQL 질의를 포함하는 제1 초기 학습 데이터셋을 결정하는 단계;

제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제1 신규 데이터베이스에 대한 제1 신규 데이터베이스 정보를 생성하는 단계;

상기 제1 스키마 변형 연산에 대응하는 제1 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제1 신규 데이터베이스에 대한 제1 신규 학습용 SQL 질의를 생성하는 단계;

상기 제1 학습용 자연어 질의와, 상기 제1 신규 데이터베이스 정보와, 상기 제1 신규 학습용 SQL 질의를 포함하는 제1 신규 학습 데이터셋을 결정하는 단계; 및

상기 제1 신규 학습 데이터셋을 사용하여 상기 신경망의 학습을 수행하는 단계;

를 포함하는 신경망 학습 방법.

청구항 8

청구항 7에 있어서,

제2 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제2 신규 데이터베이스에 대한 제2 신규 데이터베이스 정보를 생성하는 단계;

상기 제2 스키마 변형 연산에 대응하는 제2 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기

제2 신규 데이터베이스에 대한 제2 신규 학습용 SQL 질의를 생성하는 단계; 및

상기 제1 학습용 자연어 질의와, 상기 제2 신규 데이터베이스 정보와, 상기 제2 신규 학습용 SQL 질의를 포함하는 제2 신규 학습 데이터셋을 결정하는 단계;

를 더 포함하며, 상기 신경망의 학습을 수행하는 단계는 상기 제1 및 제2 신규 학습 데이터셋을 사용하여 수행되는 신경망 학습 방법.

청구항 9

청구항 8에 있어서,

제2 학습용 자연어 질의와, 상기 데이터베이스 정보와, 상기 제2 학습용 자연어 질의에 상응하는 제2 학습용 SQL 질의를 포함하는 제2 초기 학습 데이터셋을 결정하는 단계;

상기 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제3 신규 데이터베이스에 대한 제3 신규 데이터베이스 정보를 생성하는 단계;

상기 제1 SQL 질의 동기화 연산을 상기 제2 학습용 SQL 질의에 적용하여 상기 제3 신규 데이터베이스에 대한 제3 신규 학습용 SQL 질의를 생성하는 단계; 및

상기 제2 학습용 자연어 질의와, 상기 제3 신규 데이터베이스 정보와, 상기 제3 신규 학습용 SQL 질의를 포함하는 제3 신규 학습 데이터셋을 결정하는 단계;

를 더 포함하며, 상기 신경망의 학습을 수행하는 단계는 상기 제1 내지 제3 신규 학습 데이터셋을 사용하여 수행되는 신경망 학습 방법.

청구항 10

청구항 9에 있어서,

상기 제1 스키마 변형 연산을 상기 제1 신규 데이터베이스에 적용하여 상기 제1 신규 데이터베이스와 다른 스키마를 가지는 제4 신규 데이터베이스에 대한 제4 신규 데이터베이스 정보를 생성하는 단계;

상기 제1 SQL 질의 동기화 연산을 상기 제1 신규 학습용 SQL 질의에 적용하여 상기 제4 신규 데이터베이스에 대한 제4 신규 학습용 SQL 질의를 생성하는 단계; 및

상기 제1 학습용 자연어 질의와, 상기 제4 신규 데이터베이스 정보와, 상기 제4 신규 학습용 SQL 질의를 포함하는 제4 신규 학습 데이터셋을 결정하는 단계;

를 더 포함하며, 상기 신경망의 학습을 수행하는 단계는 상기 제1 내지 제4 신규 학습 데이터셋을 사용하여 수행되는 신경망 학습 방법.

청구항 11

청구항 7에 있어서,

제2 학습용 자연어 질의와, 상기 데이터베이스 정보와, 상기 제2 학습용 자연어 질의에 상응하는 제2 학습용 SQL 질의를 포함하는 제2 초기 학습 데이터셋을 결정하는 단계;

상기 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제2 신규 데이터베이스에 대한 제2 신규 데이터베이스 정보를 생성하는 단계;

상기 제1 스키마 변형 연산에 대응하는 상기 제1 SQL 질의 동기화 연산을 상기 제2 학습용 SQL 질의에 적용하여 상기 제2 신규 데이터베이스에 대한 제2 신규 학습용 SQL 질의를 생성하는 단계; 및

상기 제2 학습용 자연어 질의와, 상기 제2 신규 데이터베이스 정보와, 상기 제2 신규 학습용 SQL 질의를 포함하는 제2 신규 학습 데이터셋을 결정하는 단계;

를 더 포함하며, 상기 신경망의 학습을 수행하는 단계는 상기 제1 및 제2 신규 학습 데이터셋을 사용하여 수행되는 신경망 학습 방법.

청구항 12

청구항 7에 있어서, 상기 제1 스키마 변형 연산은 상기 데이터베이스의 스키마 구조를 변경하는 제1 변형 연산 및 상기 데이터베이스의 스키마 요소의 이름을 변경하는 제2 변형 연산 중에서 적어도 하나를 포함하는 신경망 학습 방법.

청구항 13

자연어 질의에 상응하는 검색 결과를 제공하는 정보 검색 장치로서,

프로그램 명령들을 저장하는 메모리와; 상기 메모리에 접속되고 상기 메모리에 저장된 상기 프로그램 명령들을 실행하는 프로세서;를 구비하며,

상기 프로그램 명령들은 상기 프로세서에 의해 실행될 때 상기 프로세서로 하여금:

소정의 신경망을 학습시키는 동작;

상기 자연어 질의를 받아들이는 동작;

상기 신경망을 토대로 상기 자연어 질의를 SQL 질의로 번역하는 동작; 및

상기 SQL 질의를 사용하여 상기 자연어 질의에 상응하는 상기 검색 결과를 획득하는 동작;

을 수행하며, 상기 프로세서로 하여금 상기 신경망을 학습시키는 동작을 수행하게 하는 프로그램 명령들은 상기 프로세서로 하여금:

제1 학습용 자연어 질의와, 상기 데이터베이스에 관한 데이터베이스 정보와, 상기 제1 학습용 자연어 질의에 상응하는 제1 학습용 SQL 질의를 포함하는 제1 초기 학습 데이터셋을 결정하고;

제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제1 신규 데이터베이스에 대한 제1 신규 데이터베이스 정보를 생성하고;

상기 제1 스키마 변형 연산에 대응하는 제1 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제1 신규 데이터베이스에 대한 제1 신규 학습용 SQL 질의를 생성하고;

상기 제1 학습용 자연어 질의와, 상기 제1 신규 데이터베이스 정보와, 상기 제1 신규 학습용 SQL 질의를 포함하는 제1 신규 학습 데이터셋을 결정하고;

상기 제1 신규 학습 데이터셋을 사용하여 상기 신경망의 학습을 수행하게 하는 정보 검색 장치.

청구항 14

청구항 13에 있어서, 상기 프로세서로 하여금 상기 신경망을 학습시키는 동작을 수행하게 하는 프로그램 명령들은 상기 프로세서로 하여금:

제2 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제2 신규 데이터베이스에 대한 제2 신규 데이터베이스 정보를 생성하고;

상기 제2 스키마 변형 연산에 대응하는 제2 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제2 신규 데이터베이스에 대한 제2 신규 학습용 SQL 질의를 생성하고;

상기 제1 학습용 자연어 질의와, 상기 제2 신규 데이터베이스 정보와, 상기 제2 신규 학습용 SQL 질의를 포함하는 제2 신규 학습 데이터셋을 결정하는; 동작을 더 수행하게 하며,

상기 신경망의 학습을 수행하는 동작은 상기 제1 및 제2 신규 학습 데이터셋을 사용하여 수행되는 정보 검색 장치.

청구항 15

청구항 14에 있어서, 상기 프로세서로 하여금 상기 신경망을 학습시키는 동작을 수행하게 하는 프로그램 명령들은 상기 프로세서로 하여금:

제2 학습용 자연어 질의와, 상기 데이터베이스 정보와, 상기 제2 학습용 자연어 질의에 상응하는 제2 학습용 SQL 질의를 포함하는 제2 초기 학습 데이터셋을 결정하고;

상기 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제3 신

규 데이터베이스에 대한 제3 신규 데이터베이스 정보를 생성하고;

상기 제1 SQL 질의 동기화 연산을 상기 제2 학습용 SQL 질의에 적용하여 상기 제3 신규 데이터베이스에 대한 제3 신규 학습용 SQL 질의를 생성하고;

상기 제2 학습용 자연어 질의와, 상기 제3 신규 데이터베이스 정보와, 상기 제3 신규 학습용 SQL 질의를 포함하는 제3 신규 학습 데이터셋을 결정하는; 동작을 더 수행하게 하며,

상기 신경망의 학습을 수행하는 동작은 상기 제1 내지 제3 신규 학습 데이터셋을 사용하여 수행되는 정보 검색 장치.

청구항 16

청구항 14에 있어서, 상기 프로세서로 하여금 상기 신경망을 학습시키는 동작을 수행하게 하는 프로그램 명령들은 상기 프로세서로 하여금:

상기 제1 스키마 변형 연산을 상기 제1 신규 데이터베이스에 적용하여 상기 제1 신규 데이터베이스와 다른 스키마를 가지는 제4 신규 데이터베이스에 대한 제4 신규 데이터베이스 정보를 생성하고;

상기 제1 SQL 질의 동기화 연산을 상기 제1 신규 학습용 SQL 질의에 적용하여 상기 제4 신규 데이터베이스에 대한 제4 신규 학습용 SQL 질의를 생성하고;

상기 제1 학습용 자연어 질의와, 상기 제4 신규 데이터베이스 정보와, 상기 제4 신규 학습용 SQL 질의를 포함하는 제4 신규 학습 데이터셋을 결정하는; 동작을 더 수행하게 하며,

상기 신경망의 학습을 수행하는 동작은 상기 제1 내지 제4 신규 학습 데이터셋을 사용하여 수행되는 정보 검색 장치.

청구항 17

청구항 13에 있어서, 상기 프로세서로 하여금 상기 신경망을 학습시키는 동작을 수행하게 하는 프로그램 명령들은 상기 프로세서로 하여금:

제2 학습용 자연어 질의와, 상기 데이터베이스 정보와, 상기 제2 학습용 자연어 질의에 상응하는 제2 학습용 SQL 질의를 포함하는 제2 초기 학습 데이터셋을 결정하고;

상기 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제2 신규 데이터베이스에 대한 제2 신규 데이터베이스 정보를 생성하고;

상기 제1 스키마 변형 연산에 대응하는 상기 제1 SQL 질의 동기화 연산을 상기 제2 학습용 SQL 질의에 적용하여 상기 제2 신규 데이터베이스에 대한 제2 신규 학습용 SQL 질의를 생성하고;

상기 제2 학습용 자연어 질의와, 상기 제2 신규 데이터베이스 정보와, 상기 제2 신규 학습용 SQL 질의를 포함하는 제2 신규 학습 데이터셋을 결정하는; 동작을 더 수행하게 하며,

상기 신경망의 학습을 수행하는 동작은 상기 제1 및 제2 신규 학습 데이터셋을 사용하여 수행되는 정보 검색 장치.

청구항 18

청구항 13에 있어서,

상기 제1 스키마 변형 연산은 상기 데이터베이스의 스키마 구조를 변경하는 제1 변형 연산 및 상기 데이터베이스의 스키마 요소의 이름을 변경하는 제2 변형 연산 중에서 적어도 하나를 포함하며,

상기 제1 스키마 변형 연산이 상기 제1 변형 연산만을 포함할 때 상기 제1 SQL 질의 동기화 연산은 Null 연산인 정보 검색 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공신경망 학습 방법에 관한 것으로서, 보다 상세하게는, 인공신경망 학습에 필요한 학습 데이터셋

을 증강하는 방법에 관한 것이다.

배 경 기 술

[0002] 자연어 질의를 SQL(Structured query language) 질의로 번역하는 것은 전문지식이 없는 사람이 관계형 데이터베이스에 질의하고 데이터를 획득할 수 있게 해주며, 관계형 데이터베이스를 구비하는 시스템의 범용성과 활용도를 높이는데 도움이 될 수 있다. 이에 따라 자연어 질의를 SQL로 번역하는 것은 자연어 처리 및 데이터베이스 분야에서 많은 관심을 받고 있다. 특히 최근에는 신경망을 기반으로 자연어 질의를 번역하는 방법에 대한 연구가 활발히 이루어지고 있다. 신경망을 기반으로 자연어 질의를 번역하고자 하는 경우에는, 신경망을 사전에 학습시키기 위해서 다량의 다양한 학습 데이터셋이 필요하다. 그런데, 신경망 모델 학습에 사용할 수 있는 학습 데이터가 부족한 경우가 많다. 복잡한 SQL 질의를 포함한 학습 데이터셋을 수집하는 것은 SQL 전문가의 노동력을 요구하는 것으로서, 시간과 비용이 많이 소요될 수 있는 어려운 문제이다. 이 문제를 해결하기 위해 다양한 학습 데이터 증강 방법들이 제시된 바 있다. 학습 데이터 증강 방법들은 크게 템플릿 기반 방식과 SQL을 자연어로 번역하는 SQL-to-Text 모델을 사용한 방식으로 나눌 수 있다.

[0003] 템플릿 기반 방식에 따르면, 기 정의된 자연어 질의와 SQL 쌍의 템플릿을 이용해서 합성 데이터(synthetic data)가 생성된다. 템플릿에는 데이터베이스 내의 데이터에 대한 참조 부분이 슬롯으로 비워져있다. 데이터베이스가 주어지면, 데이터베이스 내의 데이터를 임의로 선택하여 해당 슬롯을 채우는 방식으로 데이터를 생성한다. 이러한 방식은 사람이 정의한 한정된 수의 템플릿을 이용해 합성 데이터를 생성하기 때문에, 다양한 형태의 데이터를 수집하기 위해서는 다수의 템플릿을 정의하는 수작업이 요구된다. 또한, 데이터베이스의 도메인을 고려한 자연스러운 표현의 문장을 생성하는 것은 어려운 일이며, 이를 위해서는 도메인에 무관하게 적용 가능한 한정된 표현만을 채택하거나 데이터베이스의 도메인에 따라 새로운 템플릿을 만들어야 한다는 문제가 있다.

[0004] SQL-to-Text 모델을 사용한 방식은 SQL을 자연어로 번역하는 모델을 이용한다. 먼저, 다량의 SQL 질의와 대응되는 데이터베이스를 웹에서 크롤하거나 기 정의된 문법과 주어진 데이터베이스를 가지고 SQL 질의를 자동 합성한 뒤에, 각 SQL 질의에 대응되는 자연어 질의 문장을 모델을 사용해 생성하게 된다. 이러한 방식은 템플릿 기반 방식에 비해 다양한 자연어 질의 표현을 생성할 수 있다는 장점이 있지만, SQL-to-Text 모델을 학습시키기 위한 다양한 데이터를 수집 혹은 생성하는 것에 대한 어려움이 여전히 존재한다.

발명의 내용

해결하려는 과제

[0005] 이와 같이 종래의 학습 데이터 증강 방법들은 수작업으로 생성한 템플릿이나, 또 다른 모델 즉, SQL-to-Text 모델을 학습시키기 위한 자연어 학습 데이터를 추가로 필요로 한다. 아울러, 종래의 방법들은 데이터베이스 스키마 측면의 다양성 확장은 보장해주지 않는다.

[0006] 본 발명은 이와 같은 문제점을 해결하기 위한 것으로서, 수작업으로 생성한 템플릿을 요하지 않으면서, 자연어 학습 데이터의 추가가 없더라도 다양한 학습 데이터를 증강시킬 수 있게 해주는 학습 데이터 증강 방법을 제공하는 것을 기술적 과제로 한다.

과제의 해결 수단

[0007] 본 발명의 일 측면에 따르면, 소정의 신경망을 토대로 자연어 질의를 SQL 질의로 번역하는 번역기를 포함하는 자연어 처리 시스템에서 상기 신경망의 학습에 사용되는 학습 데이터셋을 증강시키는 방법이 제공된다. 학습 데이터셋 증강 방법은 제1 학습용 자연어 질의와, 상기 데이터베이스에 관한 데이터베이스 정보와, 상기 제1 학습용 자연어 질의에 상응하는 제1 학습용 SQL 질의를 포함하는 제1 초기 학습 데이터셋을 결정하는 단계; 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제1 신규 데이터베이스에 대한 제1 신규 데이터베이스 정보를 생성하는 단계; 상기 제1 스키마 변형 연산에 대응하는 제1 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제1 신규 데이터베이스에 대한 제1 신규 학습용 SQL 질의를 생성하는 단계; 및 상기 제1 학습용 자연어 질의와, 상기 제1 신규 데이터베이스 정보와, 상기 제1 신규 학습용 SQL 질의를 포함하는 제1 신규 학습 데이터셋을 결정하는 단계;를 포함한다.

[0008] 학습 데이터셋 증강 방법은 제2 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제2 신규 데이터베이스에 대한 제2 신규 데이터베이스 정보를 생성하는 단계; 상기 제2 스키마 변형 연산에 대응하는 제2 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제2 신규 데이터

베이스에 대한 제2 신규 학습용 SQL 질의를 생성하는 단계; 및 상기 제1 학습용 자연어 질의와, 상기 제2 신규 데이터베이스 정보와, 상기 제2 신규 학습용 SQL 질의를 포함하는 제2 신규 학습 데이터셋을 결정하는 단계;를 더 포함할 수 있다.

[0009] 학습 데이터셋 증강 방법은 제2 학습용 자연어 질의와, 상기 데이터베이스 정보와, 상기 제2 학습용 자연어 질의에 상응하는 제2 학습용 SQL 질의를 포함하는 제2 초기 학습 데이터셋을 결정하는 단계; 상기 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제3 신규 데이터베이스에 대한 제3 신규 데이터베이스 정보를 생성하는 단계; 상기 제1 SQL 질의 동기화 연산을 상기 제2 학습용 SQL 질의에 적용하여 상기 제3 신규 데이터베이스에 대한 제3 신규 학습용 SQL 질의를 생성하는 단계; 및 상기 제2 학습용 자연어 질의와, 상기 제3 신규 데이터베이스 정보와, 상기 제3 신규 학습용 SQL 질의를 포함하는 제3 신규 학습 데이터셋을 결정하는 단계;를 더 포함할 수 있다.

[0010] 학습 데이터셋 증강 방법은 상기 제1 스키마 변형 연산을 상기 제1 신규 데이터베이스에 적용하여 상기 제1 신규 데이터베이스와 다른 스키마를 가지는 제4 신규 데이터베이스에 대한 제4 신규 데이터베이스 정보를 생성하는 단계; 상기 제1 SQL 질의 동기화 연산을 상기 제1 신규 학습용 SQL 질의에 적용하여 상기 제4 신규 데이터베이스에 대한 제4 신규 학습용 SQL 질의를 생성하는 단계; 및 상기 제1 학습용 자연어 질의와, 상기 제4 신규 데이터베이스 정보와, 상기 제4 신규 학습용 SQL 질의를 포함하는 제4 신규 학습 데이터셋을 결정하는 단계;를 더 포함할 수 있다.

[0011] 상기 제1 스키마 변형 연산은 상기 데이터베이스의 스키마 구조를 변경하는 제1 변형 연산 및 상기 데이터베이스의 스키마 요소의 이름을 변경하는 제2 변형 연산 중에서 적어도 하나를 포함할 수 있다.

[0012] 본 발명의 다른 측면에 따르면, 소정의 신경망을 토대로 자연어 질의를 SQL 질의로 번역하는 번역기를 포함하는 자연어 처리 시스템에서 상기 신경망을 학습시키는 방법이 제공된다. 신경망 학습 방법은 제1 학습용 자연어 질의와, 상기 데이터베이스에 관한 데이터베이스 정보와, 상기 제1 학습용 자연어 질의에 상응하는 제1 학습용 SQL 질의를 포함하는 제1 초기 학습 데이터셋을 결정하는 단계; 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제1 신규 데이터베이스에 대한 제1 신규 데이터베이스 정보를 생성하는 단계; 상기 제1 스키마 변형 연산에 대응하는 제1 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제1 신규 데이터베이스에 대한 제1 신규 학습용 SQL 질의를 생성하는 단계; 상기 제1 학습용 자연어 질의와, 상기 제1 신규 데이터베이스 정보와, 상기 제1 신규 학습용 SQL 질의를 포함하는 제1 신규 학습 데이터셋을 결정하는 단계; 및 상기 제1 신규 학습 데이터셋을 사용하여 상기 신경망의 학습을 수행하는 단계;를 포함한다.

[0013] 신경망 학습 방법은 제2 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제2 신규 데이터베이스에 대한 제2 신규 데이터베이스 정보를 생성하는 단계; 상기 제2 스키마 변형 연산에 대응하는 제2 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제2 신규 데이터베이스에 대한 제2 신규 학습용 SQL 질의를 생성하는 단계; 및 상기 제1 학습용 자연어 질의와, 상기 제2 신규 데이터베이스 정보와, 상기 제2 신규 학습용 SQL 질의를 포함하는 제2 신규 학습 데이터셋을 결정하는 단계;를 더 포함할 수 있다. 이 경우, 상기 신경망의 학습을 수행하는 단계는 상기 제1 및 제2 신규 학습 데이터셋을 사용하여 수행될 수 있다.

[0014] 신경망 학습 방법은 제2 학습용 자연어 질의와, 상기 데이터베이스 정보와, 상기 제2 학습용 자연어 질의에 상응하는 제2 학습용 SQL 질의를 포함하는 제2 초기 학습 데이터셋을 결정하는 단계; 상기 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제3 신규 데이터베이스에 대한 제3 신규 데이터베이스 정보를 생성하는 단계; 상기 제1 SQL 질의 동기화 연산을 상기 제2 학습용 SQL 질의에 적용하여 상기 제3 신규 데이터베이스에 대한 제3 신규 학습용 SQL 질의를 생성하는 단계; 및 상기 제2 학습용 자연어 질의와, 상기 제3 신규 데이터베이스 정보와, 상기 제3 신규 학습용 SQL 질의를 포함하는 제3 신규 학습 데이터셋을 결정하는 단계;를 더 포함할 수 있다. 이 경우, 상기 신경망의 학습을 수행하는 단계는 상기 제1 내지 제3 신규 학습 데이터셋을 사용하여 수행될 수 있다.

[0015] 신경망 학습 방법은 상기 제1 스키마 변형 연산을 상기 제1 신규 데이터베이스에 적용하여 상기 제1 신규 데이터베이스와 다른 스키마를 가지는 제4 신규 데이터베이스에 대한 제4 신규 데이터베이스 정보를 생성하는 단계; 상기 제1 SQL 질의 동기화 연산을 상기 제1 신규 학습용 SQL 질의에 적용하여 상기 제4 신규 데이터베이스에 대한 제4 신규 학습용 SQL 질의를 생성하는 단계; 및 상기 제1 학습용 자연어 질의와, 상기 제4 신규 데이터베이스 정보와, 상기 제4 신규 학습용 SQL 질의를 포함하는 제4 신규 학습 데이터셋을 결정하는 단계;를 더 포함할

수 있다. 이 경우, 상기 신경망의 학습을 수행하는 단계는 상기 제1 내지 제4 신규 학습 데이터셋을 사용하여 수행될 수 있다.

[0016] 본 발명의 또 다른 측면에 따르면, 자연어 질의에 상응하는 검색 결과를 제공하는 정보 검색 장치가 제공된다. 정보 검색 장치는 프로그램 명령들을 저장하는 메모리와, 상기 메모리에 접속되고 상기 메모리에 저장된 상기 프로그램 명령들을 실행하는 프로세서를 구비한다. 상기 프로그램 명령들은 상기 프로세서에 의해 실행될 때 상기 프로세서로 하여금: 소정의 신경망을 학습시키는 동작; 상기 자연어 질의를 받아들이는 동작; 상기 신경망을 토대로 상기 자연어 질의를 SQL 질의로 번역하는 동작; 및 상기 SQL 질의를 사용하여 상기 자연어 질의에 상응하는 상기 검색 결과를 획득하는 동작;을 수행한다. 상기 프로세서로 하여금 상기 신경망을 학습시키는 동작을 수행하게 하는 프로그램 명령들은 상기 프로세서로 하여금: 제1 학습용 자연어 질의와, 상기 데이터베이스에 관한 데이터베이스 정보와, 상기 제1 학습용 자연어 질의에 상응하는 제1 학습용 SQL 질의를 포함하는 제1 초기 학습 데이터셋을 결정하고; 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제1 신규 데이터베이스에 대한 제1 신규 데이터베이스 정보를 생성하고; 상기 제1 스키마 변형 연산에 대응하는 제1 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제1 신규 데이터베이스에 대한 제1 신규 학습용 SQL 질의를 생성하고; 상기 제1 학습용 자연어 질의와, 상기 제1 신규 데이터베이스 정보와, 상기 제1 신규 학습용 SQL 질의를 포함하는 제1 신규 학습 데이터셋을 결정하고; 상기 제1 신규 학습 데이터셋을 사용하여 상기 신경망의 학습을 수행하게 한다.

[0017] 상기 프로세서로 하여금 상기 신경망을 학습시키는 동작을 수행하게 하는 프로그램 명령들은 상기 프로세서로 하여금: 제2 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제2 신규 데이터베이스에 대한 제2 신규 데이터베이스 정보를 생성하고; 상기 제2 스키마 변형 연산에 대응하는 제2 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제2 신규 데이터베이스에 대한 제2 신규 학습용 SQL 질의를 생성하고; 상기 제1 학습용 자연어 질의와, 상기 제2 신규 데이터베이스 정보와, 상기 제2 신규 학습용 SQL 질의를 포함하는 제2 신규 학습 데이터셋을 결정하는; 동작을 더 수행하게 할 수 있다. 이 경우, 상기 신경망의 학습을 수행하는 동작은 상기 제1 및 제2 신규 학습 데이터셋을 사용하여 수행될 수 있다.

[0018] 상기 프로세서로 하여금 상기 신경망을 학습시키는 동작을 수행하게 하는 프로그램 명령들은 상기 프로세서로 하여금: 제2 학습용 자연어 질의와, 상기 데이터베이스 정보와, 상기 제2 학습용 자연어 질의에 상응하는 제2 학습용 SQL 질의를 포함하는 제2 초기 학습 데이터셋을 결정하고; 상기 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제3 신규 데이터베이스에 대한 제3 신규 데이터베이스 정보를 생성하고; 상기 제1 SQL 질의 동기화 연산을 상기 제2 학습용 SQL 질의에 적용하여 상기 제3 신규 데이터베이스에 대한 제3 신규 학습용 SQL 질의를 생성하고; 상기 제2 학습용 자연어 질의와, 상기 제3 신규 데이터베이스 정보와, 상기 제3 신규 학습용 SQL 질의를 포함하는 제3 신규 학습 데이터셋을 결정하는; 동작을 더 수행하게 할 수 있다. 이 경우, 상기 신경망의 학습을 수행하는 동작은 상기 제1 내지 제3 신규 학습 데이터셋을 사용하여 수행될 수 있다.

[0019] 상기 프로세서로 하여금 상기 신경망을 학습시키는 동작을 수행하게 하는 프로그램 명령들은 상기 프로세서로 하여금: 상기 제1 스키마 변형 연산을 상기 제1 신규 데이터베이스에 적용하여 상기 제1 신규 데이터베이스와 다른 스키마를 가지는 제4 신규 데이터베이스에 대한 제4 신규 데이터베이스 정보를 생성하고; 상기 제1 SQL 질의 동기화 연산을 상기 제1 신규 학습용 SQL 질의에 적용하여 상기 제4 신규 데이터베이스에 대한 제4 신규 학습용 SQL 질의를 생성하고; 상기 제1 학습용 자연어 질의와, 상기 제4 신규 데이터베이스 정보와, 상기 제4 신규 학습용 SQL 질의를 포함하는 제4 신규 학습 데이터셋을 결정하는; 동작을 더 수행하게 할 수 있다. 이 경우, 상기 신경망의 학습을 수행하는 동작은 상기 제1 내지 제4 신규 학습 데이터셋을 사용하여 수행될 수 있다.

발명의 효과

[0020] 본 발명의 일 실시예에 따르면, 신경망을 기반으로 한 번역기를 토대로 자연어 질의에 응답하여 검색 결과를 제공하는 자연어 처리 시스템에서, 자연어 학습 데이터의 추가가 없거나 제한적인 상황에서도 데이터베이스 스키마의 변형을 통해서 신경망의 학습에 필요한 학습 데이터셋을 다양하게 확장할 수 있게 된다. 이에 따라 자연어 처리 시스템의 신경망 학습에 필요한 학습 데이터셋을 다량 확보할 수 있고, 확보된 학습 데이터셋으로 신경망을 학습시킴으로써 질의어 번역의 정확도를 제고할 수 있게 된다.

[0021] 특히 본 발명의 학습 데이터 증강 방법은 종래의 학습 데이터 증강 방법과 달리 데이터베이스 스키마의 다양성을 확장시킬 수 있으며, 주어진 학습 데이터셋에 없는 새로운 형태의 SQL 질의를 자동으로 생성할 수 있다. 본

발명의 방법은 임의의 신경망 기반 자연어-SQL 번역기의 학습에 사용될 수 있으며, 기존의 템플릿 기반 증강 기법이나 SQL-to-Text 모델 기반 증강 기법과 함께 사용될 수도 있다. 주어진 데이터베이스 스키마 내에서 자연어 질의와 SQL 질의를 생성해내는 기존 기술과 달리, 본 발명의 방법은 데이터베이스 스키마를 변형하기 때문에, 학습 데이터 내의 데이터베이스 스키마의 다양성을 높일 수 있다.

[0022] 자연어-SQL 번역의 정확도가 향상됨에 따라, RDMBS에 저장된 정보에 접근하고자 하는 사용자는, 복잡한 테이블 스키마를 이해할 필요가 없이, 자신이 원하는 정보를 말 또는 텍스트 형태의 자연어로 질의를 하고 정확한 검색 결과를 획득할 수 있다. 이에 따라 데이터베이스 관리 시스템의 접근성이 대폭 향상될 수 있다.

[0023] 본 발명은 데이터베이스 관리 시스템의 자연어 인터페이스가 필요한 모든 분야에 적용될 수 있다. 예컨대 본 발명의 방법은 클라우드 서비스에서의 사용자 인터페이스에 적용될 수 있고, 모바일 디바이스 또는 자동차에서의 핸즈프리 장치에도 특히 유용할 수 있다.

도면의 간단한 설명

[0024] 도 1은 본 발명의 일 실시예에 따른 정보 검색 시스템의 블록도이다.

도 2는 도 1에 도시된 번역기 학습 장치의 블록도이다.

도 3은 본 발명의 일 실시예에 따른 학습 데이터셋 증강 방법을 보여주는 흐름도이다.

도 4는 본 발명의 일 실시예에 따른 학습 데이터셋 증강 과정의 일 예를 설명하기 위한 도면이다.

도 5는 데이터베이스 스키마 변형 연산의 예들을 정리한 표이다.

도 6은 데이터베이스 스키마 변형 연산에 따른 SQL 동기화 연산의 결정 방법의 일 예를 보여주는 흐름도이다.

도 7은 본 발명의 일 실시예에서 다수의 학습 데이터셋 증강 동작이 연속적으로 이루어지는 과정을 전체적으로 보여주는 흐름도이다.

도 8은 본 발명의 일 실시예에서 다수의 학습 데이터셋 증강 동작이 연속적으로 이루어지는 과정의 데이터 흐름을 보여주는 도면이다.

도 9는 본 발명의 다른 예시적 실시예에 따른 학습 데이터셋 증강 과정을 요약한 의사코드를 보여준다.

발명을 실시하기 위한 구체적인 내용

[0025] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 구성요소에 대해서는 유사한 참조부호를 사용하였다.

[0026] 제1, 제2, 등의 서수가 다양한 구성요소들을 설명하는 데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. "및/또는"이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0027] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0028] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0029] 달리 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속

하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0030] 이하, 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0031] 도 1은 본 발명의 일 실시예에 따른 정보 검색 시스템의 블록도이다. 정보 검색 시스템은 클라이언트로부터 자연어 질의를 받아들이고, 자연어 질의에 상응한 검색 결과를 클라이언트에 제공한다. 클라이언트는 자연어로 의사소통을 할 수 있는 이용자가 사용하는 장치, 예컨대 pc, 스마트폰과 같은 모바일 단말기, 차량 내 정보 단말기, 또는 여타의 데이터 처리장치일 수 있다. 또한, 클라이언트는 이와 같은 장치에서 동작하는 애플리케이션 프로그램을 지칭하는 것일 수도 있다. 그렇지만, 다른 응용예에서는 정보 검색 시스템이 외부의 클라이언트를 거치지 않고 사용자의 음성을 마이크를 통해 직접 인식하고, 이를 토대로 검색을 수행할 수도 있다. 이와 같은 경우, 자연어 질의는 인식된 음성이 될 수도 있고, 인식된 음성에서 텍스트로 변환된 자연어 텍스트가 될 수도 있다.

[0032] 도시된 실시예에서, 정보 검색 시스템은 질의 번역기(100)와, 관계형 데이터베이스 관리 시스템(RDBMS: 110)과, 데이터베이스(120)를 포함할 수 있다. 아울러, 정보 검색 시스템은 질의 번역기(100)를 학습시키기 위한 번역기 학습 장치(150)를 추가로 구비할 수 있다. 도면에는 질의 번역기(100)와 번역기 학습 장치(150)가 별도로 표시되어 있지만, 이는 기능적으로 도시된 것으로서 질의 번역기(100)와 번역기 학습 장치(150)는 하나로 합체될 수도 있다. 다른 실시예에서는, 번역기 학습 장치(150)가 정보 검색 시스템에 포함되는 것이 아니라 정보 검색 시스템의 외부에 마련될 수도 있다.

[0033] 질의 번역기(100)는 자연어 질의를 받아들이고, 자연어 질의를 RDBMS(110)가 인식할 수 있는 데이터베이스 질의, 예컨대 SQL 질의로 번역할 수 있다. 질의 번역기(100)는 심층신경망(Deep Neural Network)과 같은 신경망을 토대로 구축될 수 있다. RDBMS(110)는 데이터베이스(120)에 데이터를 저장하고, 저장된 데이터를 관리하며, 다른 응용 프로그램에게 조회 서비스를 제공할 수 있다. 특히, RDBMS(110)는 상기 SQL 질의를 받아들이고, 이를 파싱하여, SQL 질의에 상응하는 데이터를 데이터베이스(120)에서 획득할 수 있다. RDBMS(110)는 데이터베이스(120)에서 독출된 검색 결과 데이터를 상기 자연어 질의를 제출한 엔티티에게 제공할 수 있다.

[0034] 번역기 학습 장치(150)는 질의 번역기(100)를 구성하는 신경망을 학습시킬 수 있다. 번역기 학습 장치(150)는 기 확보된 학습 데이터셋을 토대로 질의 번역기(100)의 파라미터를 최적화하여 학습된 파라미터를 질의 번역기(100)에 제공할 수 있다. 아울러, 번역기 학습 장치(150)는, 상기 학습된 파라미터 이외에, 번역에 참조할 수 있는 번역 참조 데이터를 학습 과정에서 획득하여 질의 번역기(100)에 제공할 수 있다.

[0035] 도 2는 본 발명의 일 실시예에 따른 번역기 학습 장치(150)의 블록도이다. 도시된 실시예에서, 번역기 학습 장치(150)는 데이터 증강 엔진(152)와 학습 엔진(154)를 포함할 수 있다.

[0036] 데이터 증강 엔진(152)은 학습 데이터셋을 받아들이고, 상기 학습 데이터셋을 증강하여 그 양을 증대시킨다. 이에 따라, 데이터 증강 엔진(152)이 출력하는 증강된 학습 데이터셋은 학습 데이터셋 테이블(156)에 저장되어 있던 원래의 학습 데이터셋 이외에, 데이터 증강 엔진(152)에 의해 추가된 학습 데이터셋을 포함할 수 있다. 예시적인 실시예에 있어서, 학습 데이터셋은 자연어 질의와, 데이터베이스와, SQL 질의의 집합으로 구성될 수 있다. 상기 자연어 질의와, 상기 SQL 질의는 데이터베이스에 대한 정보와 함께 학습 데이터셋 테이블(156)에 저장되어 있을 수 있다.

[0037] 상기 데이터베이스는 RDBMS(110)에 의해 관리되는 데이터베이스(120)일 수 있다. 그렇지만, 변형된 실시예에서는, 상기 데이터베이스가 RDBMS(110)에 의해 관리되는 데이터베이스(120)를 전체적으로 또는 부분적으로 복제한 것일 수도 있다. 또 다른 실시예에서는 상기 데이터베이스가 RDBMS(110)에 의해 관리되는 데이터베이스(120)의 스키마 정보만을 의미할 수 있다. 이와 같은 변형 가능성을 고려하여, 본 명세서에 첨부되는 청구범위에서는 학습 데이터셋에 포함되는 데이터베이스를 '데이터베이스 정보'로 칭하기로 하고, 본 명세서에서는 간략하게 '데이터베이스'로 칭하기로 한다.

[0038] 본 발명의 예시적인 실시예에 따르면, 데이터 증강 엔진(152)은 새로운 자연어 질의 데이터를 포함하는 학습 데이터를 증강하는 것이 아니라, 데이터베이스 스키마 변형 연산을 통해서 새로운 학습 데이터를 생성할 수 있다. 즉, 새로운 자연어 질의 데이터를 토대로 학습데이터를 증강하는 종래의 학습 데이터 증강 방법과 달리, 본 발명의 예시적인 실시예에 따른 데이터 증강 엔진(152)은 데이터베이스 스키마의 다양성을 확장시킴으로써 주어진

학습 데이터셋에 없는 새로운 형태의 SQL 질의를 자동으로 생성할 수 있다. 이 때, 자연어-SQL 번역을 위한 초기 학습 데이터셋만이 주어진다고 가정할 수도 있다. 이에 따라 데이터 증강 엔진(152)은, 새로운 자연어 질의 데이터를 포함하는 추가적인 학습 데이터가 외부에서 공급되지 않아도, 새로운 학습 데이터를 생성할 수 있다.

[0039] 예시적인 실시예들에 있어서, 데이터 증강 엔진(152)의 증강 동작은 스키마 변형 연산과 SQL 동기화 연산을 포함할 수 있다. 상기 스키마 변형 연산은 변형 연산 테이블(158)에 저장된 스키마 변형 연산의 종류를 토대로, 그리고 소정의 실행 지시 알고리즘에 따라서 수행될 수 있다. 상기 SQL 동기화 연산은 변형 연산 테이블(158)에 저장된 동기화 연산 정의 정보를 토대로 수행될 수 있다. 구체적인 증강 동작은 아래에서 구체적으로 설명한다.

[0040] 학습 엔진(154)은 데이터 증강 엔진(152)으로부터 증강된 학습 데이터셋을 받아들이고, 증강된 학습 데이터셋을 토대로 질의 번역기(100)를 학습시키거나 학습 상태를 갱신시킬 수 있다. 위에서 언급한 바와 같이, 학습이 진행됨에 따라 또는 학습이 완료된 후, 학습 엔진(154)은 학습된 파라미터와 번역 참조 데이터를 질의 번역기(100)에 제공할 수 있다.

[0041] 도 3은 본 발명의 일 실시예에 따른 학습 데이터셋 증강 방법을 보여주는 흐름도이다. 도 4는 본 발명의 일 실시예에 따른 학습 데이터셋 증강 과정의 일 예를 설명하기 위한 도면이다.

[0042] 도 3과 도 4를 참조하면, 데이터 증강 엔진(152)은 먼저 초기 학습 데이터셋을 받아들인다(제200단계). 예컨대, 도 5에 도시된 바와 같이, 첫 번째 초기 학습 데이터셋(DS_1)은 자연어 질의(q_1)와, 데이터베이스(D_1)와, SQL 질의(Q_1)를 포함할 수 있다.

[0043] 이어서, 데이터 증강 엔진(152)은 데이터 증강을 위해 상기 데이터베이스(D_1)에 적용할 스키마 변형 연산 $f_1()$을 결정할 수 있다(제210단계). 그리고, 데이터 증강 엔진(152)은 데이터베이스(D_1)에 스키마 변형 연산 $f_1()$를 적용하여 새로운 데이터베이스(D_{11})를 생성할 수 있다(제220단계).

[0044] 데이터 증강 엔진(152)은 스키마 변형 연산 $f_1()$에 대응하는 SQL 질의 동기화 연산이 무엇인지 확인하고, 확인된 SQL 질의 동기화 연산을 SQL 질의(Q_1)에 적용하여, 새로운 데이터베이스(D_{11})에 대한 SQL 질의(Q_{11})를 생성할 수 있다(제230단계).

[0045] 이와 같은 과정을 통해 도출되는 새로운 학습 데이터셋(DS_{11})은 최초의 자연어 질의(q_1)와, 새로운 데이터베이스(D_{11}), 그리고 새로운 SQL 질의(Q_{11})를 포함하게 된다. 데이터 증강 엔진(152)은 새로운 학습 데이터셋(q_1, D_{11}, Q_{11})을 증강에 의해 추가된 학습 데이터셋(DS_{11})으로서 저장할 수 있다(제240단계).

[0046] 도 4에 도시된 바와 같이, 위와 같은 데이터 증강 동작은 1회성으로 종료되는 것이 아니라, 다른 스키마 변형 연산에 대해서, 그리고 다른 초기 학습 데이터셋에 대해서 반복적으로 수행되어, 학습 데이터를 새로운 학습 데이터셋을 생성할 수 있다. 즉, 첫 번째 스키마 변형 연산 $f_1()$을 토대로 새로운 학습 데이터셋을 생성한 후에, 데이터 증강 엔진(152)은 두 번째 스키마 변형 연산 $f_2()$을 토대로 새로운 학습 데이터셋을 생성할 수 있다.

[0047] 데이터 증강 엔진(152)은 초기 학습 데이터셋(DS_1) 내에 있는 데이터베이스(D_1)에 적용할 두 번째 스키마 변형 연산 $f_2()$을 결정한 후(제210단계), 데이터베이스(D_1)에 두 번째 스키마 변형 연산 $f_2()$를 적용하여 새로운 데이터베이스(D_{12})를 생성할 수 있다(제220단계). 이어서, 데이터 증강 엔진(152)은 두 번째 스키마 변형 연산 $f_2()$에 대응하는 SQL 질의 동기화 연산이 무엇인지 확인하고, 확인된 SQL 질의 동기화 연산을 SQL 질의(Q_1)에 적용하여, 새로운 데이터베이스(D_{12})에 대한 SQL 질의(Q_{12})를 생성할 수 있다(제230단계). 이에 따라, 새로운 학습 데이터셋(DS_{12})는 최초의 자연어 질의(q_1)와, 새로운 데이터베이스(D_{12}), 그리고 새로운 SQL 질의(Q_{12})를 포함하게 된다. 데이터 증강 엔진(152)은 새로운 학습 데이터셋(q_1, D_{11}, Q_{11})을 증강에 의해 추가된 학습 데이터셋(DS_{12})으로서 저장할 수 있다(제240단계).

[0048] 도 5는 데이터베이스 스키마 변형 연산의 예들을 정리한 표이다. 데이터베이스 스키마 변형 연산들은 데이터베이스 스키마 구조를 변형시키는 연산들과, 데이터베이스 스키마 요소의 이름을 변경하는 연산들을 포함한다.

[0049] 데이터베이스 스키마 구조를 변형시키는 연산들은 테이블 조인, 테이블 분해, 테이블 추가, 및 컬럼 추가를 위

한 연산들을 포함한다. 테이블 조인은 고유키-외래키 조인 관계에 있는 두 개의 테이블을 키를 이용하여 하나로 합치는 것을 의미할 수 있다. 테이블 분해는 단일 테이블을 두 개의 테이블로 분해하는 것을 의미할 수 있다. 이때, 분해 후의 두 개의 테이블은 기존 테이블의 고유키를 공유하며, 고유키가 아닌 다른 속성들은 두 개의 테이블에 나뉘어 들어갈 수 있다. 테이블 추가는 데이터베이스 내에 새로운 테이블을 추가하는 것을 의미할 수 있다. 컬럼 추가는 데이터베이스 내의 임의의 테이블에 새로운 컬럼을 추가하는 것을 의미할 수 있다.

[0050] 데이터베이스 스키마 요소의 이름을 변경하는 연산들은, 이름 변경 방법에 따라, 접두사 추가, 접두사 삭제, 데이터 타입 추가, 약어 확장, 두문자어 확장, 동의어로 변경에 의한 이름 변경 연산들을 포함한다. 접두사 추가는 예컨대 컬럼 이름에 테이블 이름을 접두사로 추가하는 것을 의미할 수 있다. 접두사 삭제는 예컨대 'ContactName'을 'Name'으로 변경하는 것과 같이 컬럼 이름의 첫 번째 토큰을 삭제하는 것을 의미할 수 있다. 데이터 타입 추가는 컬럼 이름에 데이터 타입을 접두사로 추가하는 것을 의미할 수 있다. 약어 확장은 컬럼이나 테이블의 약어를 풀어서 표시하는 것을 의미할 수 있다. 두문자어 확장은 컬럼이나 테이블 이름의 두문자어를 풀어서 표시하는 것을 의미할 수 있다. 동의어로 변경은 컬럼이나 테이블 이름의 각 단어를 동의어로 변경하는 것을 의미할 수 있다.

[0051] 도 6은 데이터베이스 스키마 변형 연산에 따른 SQL 동기화 연산의 결정 방법의 일 예를 보여주는 흐름도이다.

[0052] 도 5에 요약된 예시적인 데이터베이스 스키마 변형 연산들은 크게 기존 SQL 질의를 변형시키지 않는 연산들과, 기존 SQL 질의를 변형시킬 수 있는 연산들로 구분할 수 있다. 기존 SQL 질의를 변형시키지 않는 연산들에는 새로운 테이블이나 컬럼의 추가가 포함될 수 있다. 기존 SQL 질의를 변형시킬 수 있는 연산들은 데이터베이스 스키마 이름의 변경, 테이블 분해, 테이블 조인 등을 포함할 수 있다.

[0053] 도 6을 참조하면, 일 실시예에 있어서는, SQL 동기화 연산을 결정함에 있어서, 기존 SQL 질의를 변형시키지 않는 연산에 대해서는, 변형된 데이터베이스 D'에 대해서 데이터베이스 스키마 변형 연산 이전의 SQL 질의를 그대로 유지할 수 있다(제300단계, 제310단계).

[0054] 한편, SQL 질의를 변형시키는 연산에 대해서는, 각 연산별로 SQL 동기화 연산이 정의될 수 있다(제320단계). 먼저, 데이터베이스 스키마 이름의 변경의 경우에는, 기존 이름과 새로운 이름에 대한 매핑 테이블을 생성하여 유지하고, SQL 질의의 데이터베이스 스키마 참조 부분을 새로운 이름으로 변경한다. 테이블 분해의 경우, 데이터베이스 스키마 변형 연산 이후의 SQL 질의가 분해 이전의 테이블(T)를 참조한다면, 분해 이전 테이블(T)에 대한 참조부를 분해 이후의 두 테이블(T1, T2)를 고유키로 조인한 다음, 참조하는 테이블로 변경할 수 있다. 테이블 조인의 경우, SQL 질의를 관계대수로 표현한 뒤 관계대수의 동치 관계를 이용하여 변형할 수 있다. 동치 관계의 예로는 조인을 셀렉션(Selection), 프로젝션(Projection) 연산 내부로 푸시하는 것 등을 들 수 있다. 다음으로, 관계대수로 표현된 SQL 질의에 테이블 조인에 사용된 고유키-외래키 조인 연산이 있다면 해당 참조부를 조인된 새로운 테이블에 대한 참조로 변경한다. 만일 SQL 질의가 조인된 두 개의 테이블 중 하나의 테이블만 참조한다면, 마찬가지로 참조부를 새로운 테이블에 대한 참조로 변경하되 해당 테이블의 고유키 컬럼이 NULL이 아닌 조건을 추가한다.

[0055] 도 7은 본 발명의 일 실시예에서 다수의 학습 데이터셋 증강 동작이 연속적으로 이루어지는 과정을 전체적으로 보여주는 흐름도이다. 도 8은 본 발명의 일 실시예에서 다수의 학습 데이터셋 증강 동작이 연속적으로 이루어지는 과정의 데이터 흐름을 보여주는 도면이다. 여기서는, 초기 학습 데이터셋(DSj)이 m개 존재하고, 각 초기 학습 데이터셋(DSj)에 대하여 적용할 수 있는 스키마 변형 연산 $f_i()$이 n개 존재한다고 가정한다.

[0056] 도 3과 도 4를 참조하여 설명한 바와 같이, 데이터 증강 엔진(152)은 먼저 초기 학습 데이터셋(DS_j)을 순차적으로 받아들인다(제400단계). 각각의 초기 학습 데이터셋(DS_j)은 자연어 질의(q_j)와, 데이터베이스(D_j)와, SQL 질의(Q_j)를 포함할 수 있다.

[0057] 데이터 증강 엔진(152)은 첫 번째 초기 학습 데이터셋(DS_1: q_1, D_1, Q_1)에 있는 데이터베이스(D_1)에 첫 번째 스키마 변형 연산 $f_1()$을 적용하고, 첫 번째 스키마 변형 연산 $f_1()$에 대응하는 SQL 질의 동기화 연산을 수행함으로써, 새로운 학습 데이터셋(DS_{11}: q_1, D_{11}, Q_{11})을 생성하여 저장할 수 있다(제410단계).

[0058] 이어서, 데이터 증강 엔진(152)은 첫 번째 초기 학습 데이터셋(DS_1: q_1, D_1, Q_1)에 있는 데이터베이스(D_1)에 두 번째 스키마 변형 연산 $f_2()$을 적용하고, 두 번째 스키마 변형 연산 $f_2()$에 대응하는 SQL 질의 동기화 연산을 수

행함으로써, 새로운 학습 데이터셋(DS_{12}: q_1, D_{12}, Q_{12})을 생성하여 저장할 수 있다.

[0059] 데이터 증강 엔진(152)은 첫 번째 초기 학습 데이터셋(DS_1: q_1, D_1, Q_1)에 n개의 스키마 변형 연산 $f_i()$을 순차적으로 적용함으로써 데이터 증강 동작을 수행한다. 최종적으로, 데이터 증강 엔진(152)은 첫 번째 초기 학습 데이터셋(DS_1: q_1, D_1, Q_1)에 있는 데이터베이스(D_1)에 n-번째 스키마 변형 연산 $f_n()$을 적용하고, n-번째 스키마 변형 연산 $f_n()$에 대응하는 SQL 질의 동기화 연산을 수행함으로써, 새로운 학습 데이터셋(DS_{1n}: q_1, D_{1n}, Q_{1n})을 생성하여 저장할 수 있다(제420단계).

[0060] 데이터 증강 엔진(152)은 이와 같은 방식으로 m개의 초기 학습 데이터셋에 대하여 순차적으로 데이터 증강 동작을 수행할 수 있다. 즉, 데이터 증강 엔진(152)은 m-번째 초기 학습 데이터셋(DS_m: q_m, D_m, Q_m)에 있는 데이터베이스(D_m)에 첫 번째 스키마 변형 연산 $f_1()$을 적용하고, 첫 번째 스키마 변형 연산 $f_1()$에 대응하는 SQL 질의 동기화 연산을 수행함으로써, 새로운 학습 데이터셋(DS_{m1}: q_m, D_{m1}, Q_{m1})을 생성하여 저장할 수 있다(제410단계).

[0061] 데이터 증강 엔진(152)은 m-번째 초기 학습 데이터셋(DS_m: q_m, D_m, Q_m)에 n개의 스키마 변형 연산 $f_i()$을 순차적으로 적용함으로써 데이터 증강 동작을 수행할 수 있다. 최종적으로, 데이터 증강 엔진(152)은 m-번째 초기 학습 데이터셋(DS_m: q_m, D_m, Q_m)에 있는 데이터베이스(D_m)에 n-번째 스키마 변형 연산 $f_n()$을 적용하고, n-번째 스키마 변형 연산 $f_n()$에 대응하는 SQL 질의 동기화 연산을 수행함으로써, 새로운 학습 데이터셋(DS_{mn}: q_m, D_{mn}, Q_{mn})을 생성하여 저장할 수 있다(제430단계).

[0062] 이와 같이 m개의 초기 학습 데이터셋(DSj) 각각에 대하여 n개의 스키마 변형 연산 $f_i()$을 순차적으로 적용함으로써, m×개의 새로운 학습 데이터셋을 생성할 수 있다. 또한, 새로이 생성된 m×개의 학습 데이터셋 각각에 대하여 다시 n개의 스키마 변형 연산 $f_i()$을 순차적으로 적용함으로써, 새로운 학습 데이터셋을 추가적으로 생성할 수 있다. 도 7과 도 8에 도시된 재귀적인 학습 데이터 증강 동작은 사전에 정해진 종료 조건이 달성될 때까지 계속될 수 있다. 이와 같은 과정을 통해서, 학습 데이터셋의 숫자는 비약적으로 증가할 수 있다.

[0063] 도 9는 본 발명의 다른 예시적 실시예에 따른 학습 데이터셋 증강 과정을 요약한 의사코드(pseudocodes)를 보여준다.

[0064] 스키마 변형을 통한 학습 데이터 증강을 효율적으로 진행하기 위하여, 데이터베이스에 대한 스키마 변형 연산을 한 번만 수행하고, 상기 데이터베이스에 대한 모든 SQL 질의를 수정하는 과정을 한꺼번에 수행할 수 있다. 이 경우, 데이터베이스(D)와, 학습 데이터셋 내에 존재하는 SQL 질의의 집합인 Q_{SQL}, 데이터베이스(D)에 적용할 스키마 변형 연산(f_D)이 입력으로 주어진다고 볼 수 있다. 주어진 스키마 변형 연산(f_D)에 대하여, SQL 동기화 연산(f_{sql})을 구한다. 데이터베이스(D)에 스키마 변형 연산(f_D)을 적용하여 새로운 데이터베이스(D')을 구하고, SQL 질의의 집합(Q_{SQL})에 속하는 모든 SQL 질의에 대해서 동기화 연산(f_{sql})을 수행하여 변형된 SQL 질의를 얻을 수 있다.

[0065] 도 1에 도시된 정보 검색 시스템은 프로세서와 메모리를 구비하는 범용 데이터 처리 장치에 의해 구현될 수 있다. 도 10은 본 발명의 일 실시예에 따른 정보 검색 시스템의 물리적 구성 예를 보여준다. 정보 검색 시스템은 적어도 하나의 프로세서(500), 메모리(510), 및 저장 장치(520)를 구비할 수 있다. 정보 검색 시스템의 구성요소들은 버스(bus)에 의해 연결되어 데이터를 교환할 수 있다.

[0066] 프로세서(500)는 메모리(510) 및/또는 저장 장치(520)에 저장된 프로그램 명령들을 실행할 수 있다. 프로세서(500)는 적어도 하나의 중앙 처리 장치(central processing unit, CPU), 그래픽 처리 장치(graphics processing unit, GPU), 또는 본 발명에 따른 방법을 수행할 수 있는 여타의 프로세서를 포함할 수 있다. 메모리(510)는 예컨대 RAM(Random Access Memory)와 같은 휘발성 메모리와, ROM(Read Only Memory)과 같은 비휘발성 메모리를 포함할 수 있다. 메모리(510)는 저장 장치(520)에 저장된 프로그램 명령들을 로드하여, 프로세서(500)에 제공함으로써 프로세서(500)가 이를 실행할 수 있도록 할 수 있다. 저장 장치(520)는 프로그램 명령들과 데이터를 저장하기에 적합한 기록매체로서, 예컨대 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM(Compact Disk Read Only Memory), DVD(Digital Video Disk)와 같은 광 기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-Optical Media), 플래시 메모리나 EPROM(Erasable Programmable ROM) 또는 이들을 기반으로 제작되는 SSD와 같은 반도체 메모리를 포함

할 수 있다.

[0067] 상기 프로그램 명령들은 프로세서(500)에 의해 실행될 때 프로세서(500)로 하여금 본 발명에 의한 정보 검색 방법, 특히, 신경망 학습 데이터셋 증강 방법을 구현하는데 필요한 동작을 수행하게 할 수 있다. 예컨대, 상기 프로그램 명령들은 프로세서(500)에 의해 실행될 때 프로세서(500)로 하여금: 소정의 신경망을 학습시키는 동작; 상기 자연어 질의를 받아들이는 동작; 상기 신경망을 토대로 상기 자연어 질의를 SQL 질의로 번역하는 동작; 및 상기 SQL 질의를 사용하여 상기 자연어 질의에 상응하는 상기 검색 결과를 획득하는 동작;을 수행하게 할 수 있다. 상기 프로세서(500)로 하여금 상기 신경망을 학습시키는 동작을 수행하게 하는 프로그램 명령들은 상기 프로세서(500)로 하여금: 제1 학습용 자연어 질의와, 상기 데이터베이스에 관한 데이터베이스 정보와, 상기 제1 학습용 자연어 질의에 상응하는 제1 학습용 SQL 질의를 포함하는 제1 초기 학습 데이터셋을 결정하고; 제1 스키마 변형 연산을 상기 데이터베이스에 적용하여 상기 데이터베이스와 다른 스키마를 가지는 제1 신규 데이터베이스에 대한 제1 신규 데이터베이스 정보를 생성하고; 상기 제1 스키마 변형 연산에 대응하는 제1 SQL 질의 동기화 연산을 상기 제1 학습용 SQL 질의에 적용하여 상기 제1 신규 데이터베이스에 대한 제1 신규 학습용 SQL 질의를 생성하고; 상기 제1 학습용 자연어 질의와, 상기 제1 신규 데이터베이스 정보와, 상기 제1 신규 학습용 SQL 질의를 포함하는 제1 신규 학습 데이터셋을 결정하고; 상기 제1 신규 학습 데이터셋을 사용하여 상기 신경망의 학습을 수행하게 할 수 있다.

[0068] 이상에서는 학습 데이터셋을 증강함에 있어 데이터베이스 스키마 변형 연산과 이에 따른 SQL 질의 동기화를 통해 데이터베이스 스키마의 다양성을 확장시키는 것을 위주로 설명하였지만, 자연어 질의의 숫자를 증가를 통한 학습 데이터셋 증강을 완전히 배제하는 것은 아니다.

[0069] 위에서 언급한 바와 같이 본 발명의 실시예에 따른 장치와 방법은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 프로그램 또는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의해 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어 분산 방식으로 컴퓨터로 읽을 수 있는 프로그램 또는 코드가 저장되고 실행될 수 있다.

[0070] 상기 컴퓨터가 읽을 수 있는 기록매체는 롬(rom), 램(ram), 플래시 메모리(flash memory) 등과 같이 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치를 포함할 수 있다. 프로그램 명령은 컴파일러(compiler)에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터(interpreter) 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

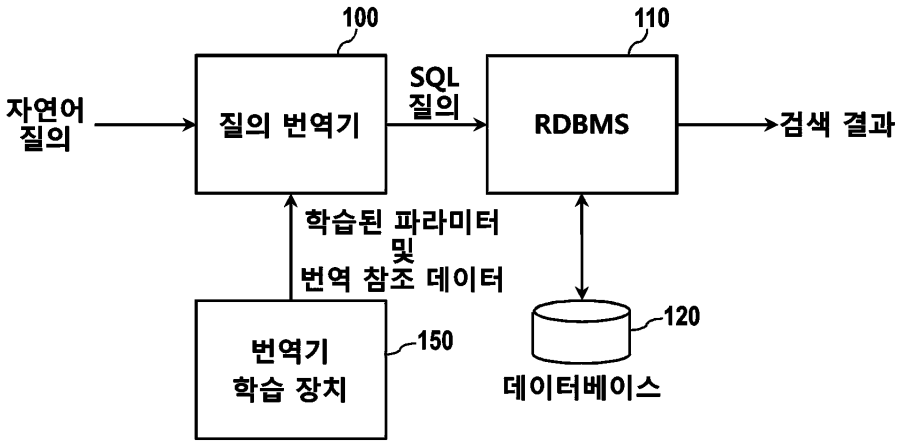
[0071] 본 발명의 일부 측면들은 장치의 문맥에서 설명되었으나, 그것은 상응하는 방법에 따른 설명 또한 나타낼 수 있고, 여기서 블록 또는 장치는 방법 단계 또는 방법 단계의 특징에 상응한다. 유사하게, 방법의 문맥에서 설명된 측면들은 또한 상응하는 블록 또는 아이템 또는 상응하는 장치의 특징으로 나타낼 수 있다. 방법 단계들의 몇몇 또는 전부는 예를 들어, 마이크로프로세서, 프로그램 가능한 컴퓨터 또는 전자 회로와 같은 하드웨어 장치에 의해(또는 이용하여) 수행될 수 있다. 몇몇의 실시예에서, 가장 중요한 방법 단계들의 하나 이상은 이와 같은 장치에 의해 수행될 수 있다.

[0072] 실시예들에서, 프로그램 가능한 로직 장치(예를 들어, 필드 프로그래머블 게이트 어레이)가 여기서 설명된 방법들의 기능의 일부 또는 전부를 수행하기 위해 사용될 수 있다. 실시예들에서, 필드 프로그래머블 게이트 어레이는 여기서 설명된 방법들 중 하나를 수행하기 위한 마이크로프로세서와 함께 작동할 수 있다. 일반적으로, 방법들은 어떤 하드웨어 장치에 의해 수행되는 것이 바람직하다.

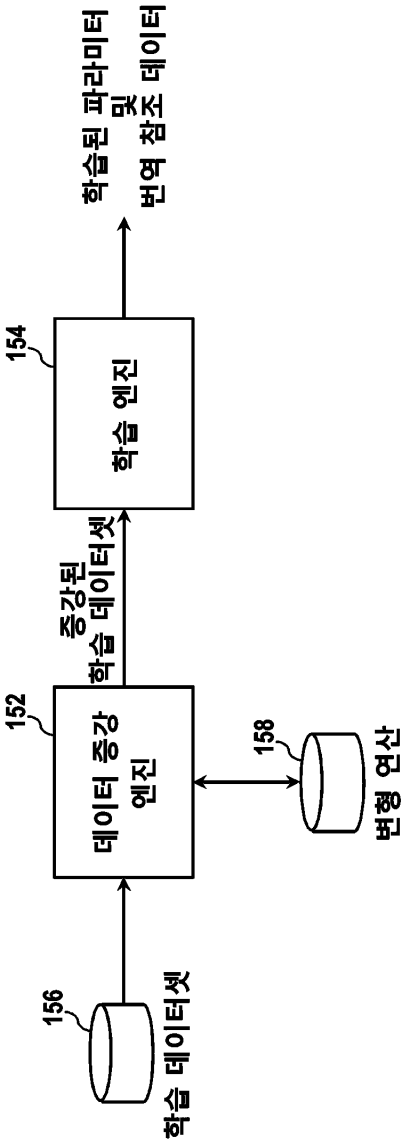
[0073] 이상에서 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

도면

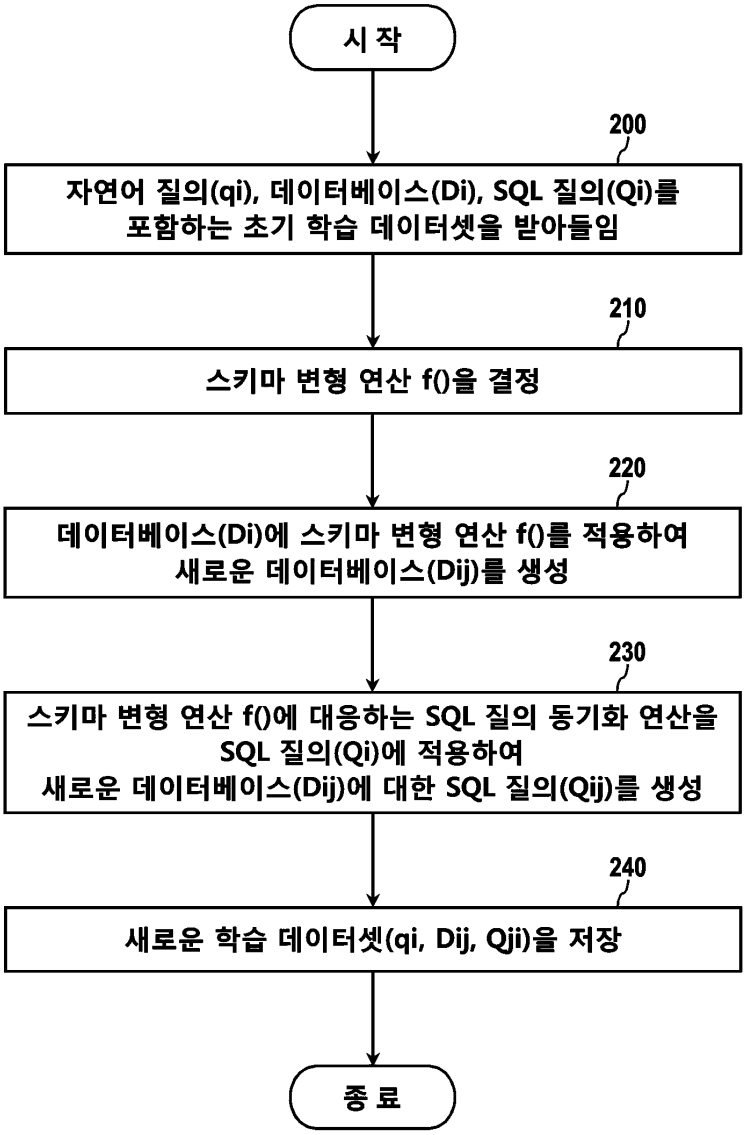
도면1


100
110
자연어 질의
질의 번역기
SQL 질의
RDBMS
검색 결과
학습된 파라미터 및 번역 참조 데이터
번역기 학습 장치
150
120
데이터베이스

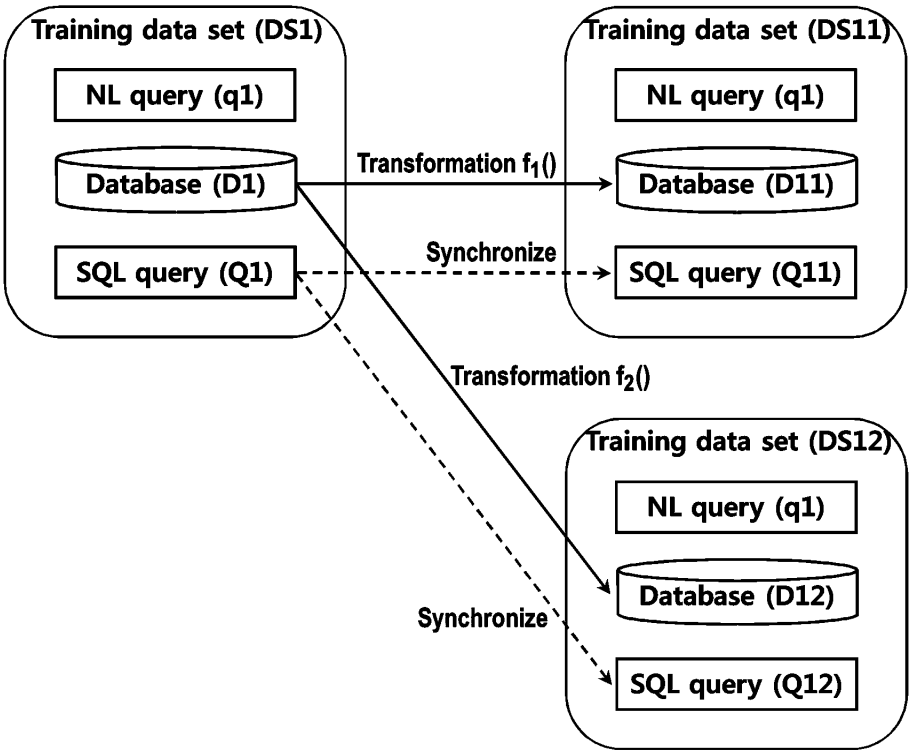
도면2


학습된 파라미터 및 번역 참조 데이터
154
학습 엔진
증강된 학습 데이터셋
152
데이터 증강 엔진
158
변형 연산
156
학습 데이터셋

도면3


시 작
200
자연어 질의(qi), 데이터베이스(Di), SQL 질의(Qi)를
포함하는 초기 학습 데이터셋을 받아들임
210
스키마 변형 연산 f()을 결정
220
데이터베이스(Di)에 스키마 변형 연산 f()를 적용하여
새로운 데이터베이스(Dij)를 생성
230
스키마 변형 연산 f()에 대응하는 SQL 질의 동기화 연산을
SQL 질의(Qi)에 적용하여
새로운 데이터베이스(Dij)에 대한 SQL 질의(Qij)를 생성
240
새로운 학습 데이터셋(qi, Dij, Qji)을 저장
종 료

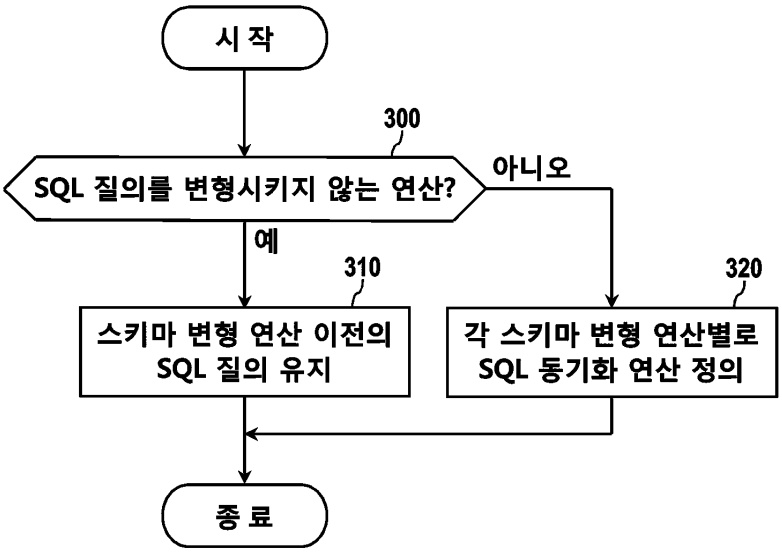
도면4


Training data set (DS1)
NL query (q1)
Database (D1)
SQL query (Q1)
Transformation $f_1()$
Synchronize
Transformation $f_2()$
Synchronize
Training data set (DS11)
NL query (q1)
Database (D11)
SQL query (Q11)
Training data set (DS12)
NL query (q1)
Database (D12)
SQL query (Q12)

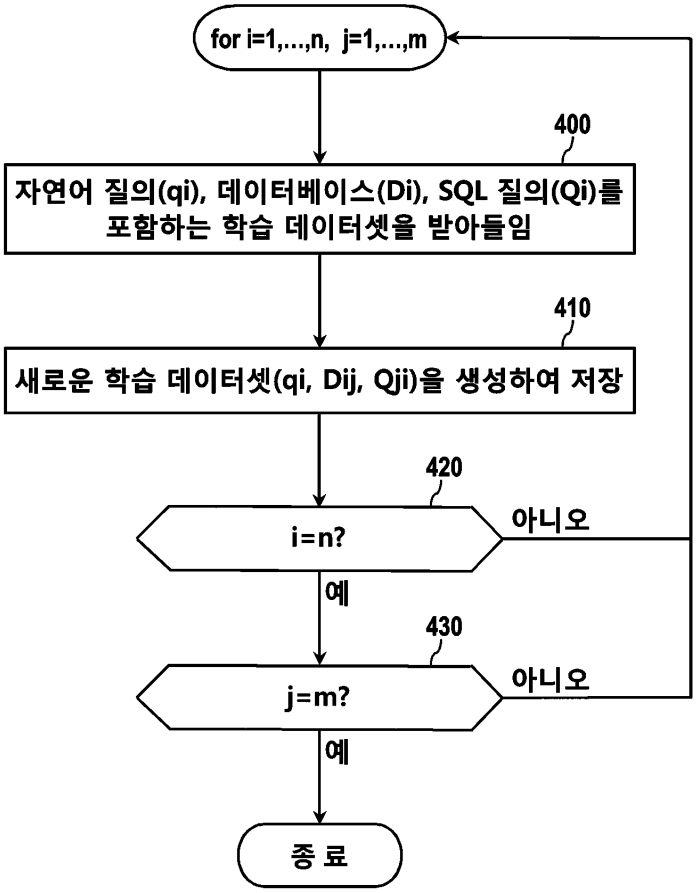
도면5

변형 대상	항목	설명
데이터베이스 스키마 구조	테이블 조인	고유키-외래키 조인 관계에 있는 두 개의 테이블을 키를 이용하여 하나로 합침
	테이블 분해	단일 테이블을 두 개의 테이블로 분해함, 분해 후의 두 개의 테이블은 기존 테이블의 고유키를 공유하며, 고유키가 아닌 다른 속성들은 두 개의 테이블에 나뉘어 들어감
	테이블 추가	데이터베이스 내에 새로운 테이블을 추가
	컬럼 추가	데이터베이스 내의 임의의 테이블에 새로운 컬럼을 추가
데이터베이스 스키마 요소의 이름	접두사 추가	컬럼 이름에 테이블 이름을 접두사로 추가
	접두사 삭제	컬럼 이름의 첫 번째 토큰을 삭제
	데이터 타입 추가	컬럼 이름에 데이터 타입을 접두사로 추가
	약어 확장	컬럼/테이블의 약어를 풀어서 작성
	두문자어 확장	컬럼/테이블 이름의 두문자어를 풀어서 작성
	동의어로 변경	컬럼/테이블 이름의 각 단어를 동의어로 변경

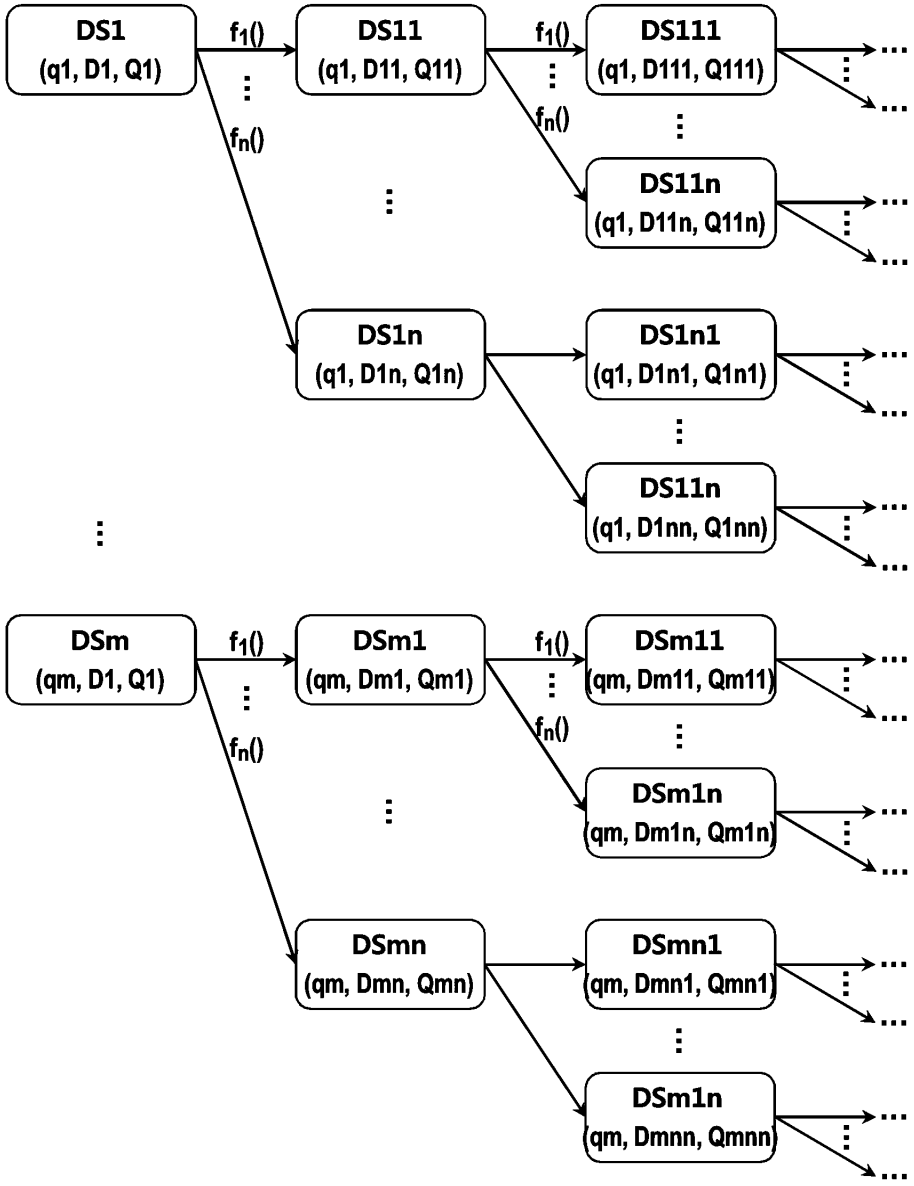
도면6



도면7


for i=1,...,n, j=1,...,m
400
자연어 질의(qi), 데이터베이스(Di), SQL 질의(Qi)를 포함하는 학습 데이터셋을 받아들임
410
새로운 학습 데이터셋(qi, Dij, Qji)을 생성하여 저장
420
i=n?
아니오
예
430
j=m?
아니오
예
종 료

도면8



도면9

Algorithm 1: DATABASEAUGMENTATION

Input: Database D, a set of SQL queries Q_{SQL} on D, One of predefined database transformation rules f_D

Output: Transformed database D', Synchronized SQL queries Q'_{SQL}

$f_{sql} \leftarrow$ GETSYNCHRONIZATIONRULE(f_D)
$Q'_{SQL} \leftarrow \emptyset$
$D' \leftarrow f_D(D)$
foreach $q_{sql} \in Q_{SQL}$ **do**
  $Q'_{SQL} \leftarrow Q'_{SQL} \cup f_{sql}(q_{sql}, D, D')$
return D', Q'_{SQL}

도면10