한국컴퓨터정보학회 동계학술대회 논문집 제25권 제1호 (2017. 1)

Predicate Tree를 이용한 질의 처리 최적화

송병후*, 김상영*, 송준석O, 김경태*, 윤희용**
O*성균관대학교 정보통신대학 전자전기컴퓨터공학과
**성균관대학교 소프트웨어대학 소프트웨어학과
e-mail:{by911129, impsoft, alskpo}@skku.eduO*, kyungtaekim76@gmail.com*, youn7147@skku.edu**

Optimization of Query Processing Using Predicate Tree

Byung-Hoo Song*, Sang-Young Kim*, Jun-Seok SongO, Kyung-Tae Kim*, Hee-Yong Youn**
O*Dept. of Electrical and Computer Engineering, Sungkyunkwan University
**Dept. of Software, Sungkyunkwan University

● 요 약 ●

본 논문은 Predicate Tree를 이용한 질의 최적화를 서술한다. 인터넷 등의 보급으로 데이터는 급등했으며 이러한 대용량 데이터를 처리하기 위해서는 적적한 모델이 필요하다. 시멘틱 웹은 컴퓨터가 해독할 수 있는 데이터의 형태로 데이터를 저장하는 것을 말하며, RDF는 시멘틱 웹에서 중요한 역할을 한다. RDF는 유동성과 데이터의 규모가 크며 그래프 모델을 통한 질의 처리는 데이터가 커짐에 따라 성능이 저하된다. 본 논문에서는 이러한 시멘틱 웹의 포맷인 RDF를 제안하는 기법인 Predicate Tree를 이용하여 데이터를 저장하고 처리한다.

키워드: RDF, SPARQL, Query, Predciate Tree

I. Introduction

시멘틱 웹(Sementic Web)은 컴퓨터가 해독할 수 있는 웹 데이터에 대한 의미적 정보를 뜻하는 용어로 Metadata를 이용하여 보다 지능적인 시스템을 구축해준다. RDF(Resource Description Framework)는 W3C(World Wide Web Construction)에서 표준안으로 제정한 언어로, RDF는 시멘틱 웹에서 정보를 저장하는 데이터 모델 형식으로 웹 Resource에 대한 Metadata를 표현한다.[1] RDF는 주어(Subject), 서술어(Predicate), 목적어(Object) 이렇게 구성되며, 이러한 데이터 셋은 인터넷 사용의 일반화로 인해서 데이터가 폭발적으로 증가하고 있으며 유동성이 크기 때문에 데이터를 처리하는 것에 지장이 있다. 이에 이러한 유동적인 대용량 데이터를 처리하기 위한 질의 처리 및 저장 방식에 관한 연구가 진행이 되고 있다.[2, 3]

SPARQL은 그래프 기반 질의 언어로 표현력에 있어 여타 온톨로지 Query에 비해서 우수하다.[4] SPARQL의 이점은 그래프 패턴을 매칭시키는 것에 있다. 따라서 이러한 방대한 양의 데이터를 처리하는 것에 적합하다. 따라서, 본 논문에서는 Predicate Tree를 이용하여 데이터를 저장하고 SPARQL 언어를 이용하여 질이 처리를 최적화에 대한 연구를 서술한다.

II. Preliminaries

1. Related works

1.1 RDF

RDF는 주어, 서술어, 목적어로 구성된 Triple 구조의 데이터 포맷으로 웹상의 Resource의 정보를 표현하기 위한 XML(eXtensible Markup Language) 규격이다. 주어와 서술어는 Resource의 URI를 의미하며 목적어는 URI 또는 Literal 값을 가진다. RDF 모델은 주로 그래프 모델로 다음과 같다.

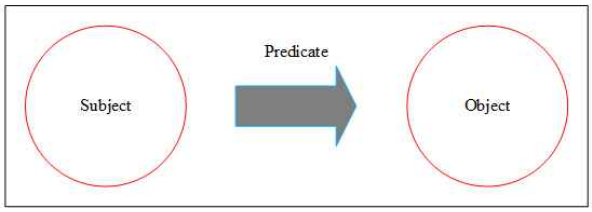

Fig. 1. RDF Graph

III. The Proposed Scheme

본 논문에서는 기존의 Graph 형식의 데이터 처리가 아닌 Predicate Tree를 제안한다. Predicate Tree는 기존의 주어에서 서술어를 통해 목적어로 가는 그래프 방식이 아닌 서술어를 기반으로 데이터 셋을 모델링한다. 서술어를 부모노드로 가지고 왼쪽노드는 주어 오른쪽노드는 목적어로 구성된다. Predicate Tree 모델은 다음과 같다.

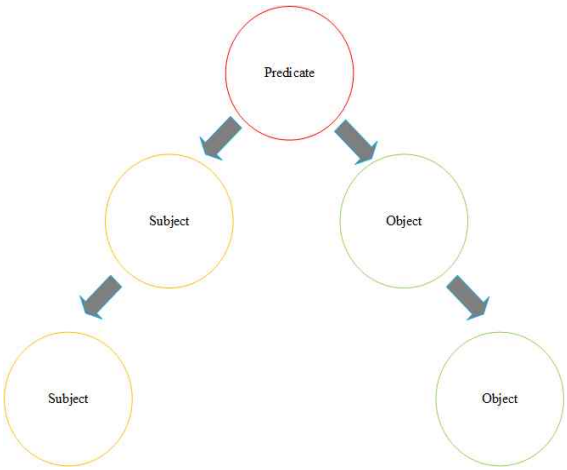

Fig. 2. Predicate Tree 구조

제안한 모델을 이용하여 데이터 셋을 저장하고 처리할 경우 기존의 방식인 주어, 서술어, 목적어를 비교하는 방식이 아닌 서술어만을 이용한 비교를 통하여 처리속도의 향상이 된다. 그림 3은 SPAQRL 언어를 이용하여 기존의 기법인 RDF-3x, Allegro를 1285개의 RDF 데이터 셋을 이용한 비교를 나타낸다. 성능 비교를 위해 이용 질의는 다음과 같다.

Query1: SELECT ?article WHERE {?article http://www.w3.org/1999/02/22-rdf-syntax-ns]type http://localhost/vocabulary/bench/Article}

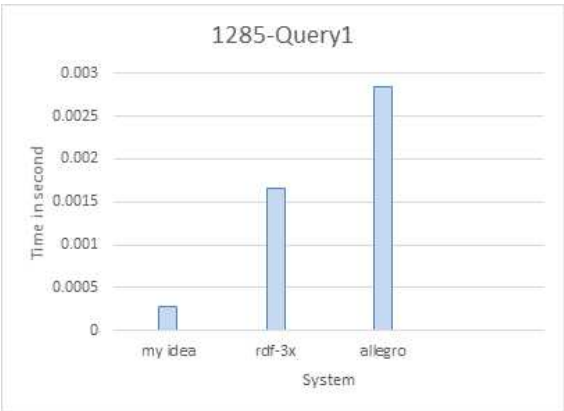

Fig. 3. Query 1 처리 속도

IV. Conclusions

본 논문에서는 제안한 Predicate Tree를 SPARQL 언어를 이용하여 질의 처리에 대한 최적화를 하였다. 추후 연구로는 다양한 데이터 셋을 이용한 성능을 검증하고자 한다.

Acknowledgment

본 연구는 Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.B0717-16-0070), Science and Technology (2016R1A6A3A11931385), the second Brain Korea 21 PLUS의 일환으로 수행되었음.

References

[1] Klyne, Graham, and Jeremy J. Carroll. "Resource description framework (RDF): Concepts and abstract syntax." (2006).

[2] Zneika, Mussab, et al. "RDF Graph Summarization Based on Approximate Patterns." International Workshop on Information Search, Integration, and Personalization. Springer International Publishing, 2015.

[3] Shen, Xuchuan, et al. "A graph-based RDF triple store." 2015 IEEE 31st International Conference on Data Engineering. IEEE, 2015.

[4] Buil-Aranda, Carlos, et al. "A preliminary investigation into SPARQL query complexity and federation in Bio2RDF." Alberto Mendelzon International Workshop on Foundations of Data Management. 2015.