빅데이터 프로비넌스 표준화를 위한 현황 조사

정동원*†, 온병원*, 정현준**, 이석훈*†, 하수욱***, 이강찬***

A Survey for Big Data Provenance Standardization

Dongwon Jeong*† , Byung-Won On*, Hyun-Jun Jung**, Sukhoon Lee*† , Su-Wook Ha***, and Kangchan Lee***

요 약

이 논문에서는 빅데이터 프로비넌스(Provenance) 표준화 현황을 분석한다. 지금까지 빅데이터에 대한 연구가 매우 활발하게 진행됐으며 최근 몇 년 사이 여러 표준화 기구에서 빅데이터 표준화를 깊이 있게 다루고 있다. 빅데이터 표준화의 범위를 정의하기란 쉽지 않으며 고전적인 데이터 표준화와의 경계가 모호한 경우도 많다. 이 논문에서는 빅데이터 프로비넌스와 관련한 표준화에 초점을 두고 데이터 프로비넌스 표준화 현황을 분석한다. 또한 기존의 데이터 프로비넌스 표준과 빅데이터 프로비넌스의 차별성을 정의하고 향후 표준화 방향을 제시한다.

Abstract

This paper analyzes the standardization status for the big data provenance. Until now, much research of big data has been conducted, and several standard organizations have been deeply handled the big data standardization. It is hard to define the scope of big data standardization, and the standardization boundary between the traditional data and the big data is unclear. This paper focuses on standardization about big data provenance, and analyzes a standardization status of data provenance. It also defines differences between traditional data provenance standards and big data provenance, and suggests future standardized works.

Key words
big data, provenance, standardization

Ⅰ. 서 론

데이터 프로비넌스는 데이터의 출처, 연혁, 데이터 변경, 정보에 대한 메타데이터를 의미한다. 최근 양(Volume), 다양성(Variety), 속도(Velocity), 정확성(Veracity), 가치(Value) 등에 특징을 지닌 빅데이터가 화두에 오르며 빅데이터 영역에서도 프로비넌스를 표준화하고자 하는 움직임들이 일어나고 있다.

이 논문은 기존의 데이터 프로비넌스 표준들을 기술하고 데이터 프로비넌스와 빅데이터 프로비넌

* 군산대학교 소프트웨어융합공학과
** 고려대학교 컴퓨터전파통신공학과
*** 한국전자통신연구원
† 공동책임저자

- 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (R0166-16-1011, 빅데이터 시스템 연동 표준개발)

스의 차이점을 정의한다. 이후 정의된 차이로 인해 발생하는 빅데이터 프로비넌스의 요구사항을 기술한다.

II. 표준화 현황

2.1 PROV-DM

PROV-DM은 W3C에서 개발한 표준으로 데이터나 어떤 것의 조각을 포함하는 엔티티, 활동, 사람에 대한 정보를 의미한다[1].

PROV-DM은 단순히 데이터 모델을 표현하므로 실질적인 도메인에서 프로비넌스를 정의하고자 할 때 도메인 특성에 따라 적절한 모델링을 지원하지 못한다. 이는 빅데이터 프로비넌스 모델링에서 빅데이터의 특징이 충분히 반영되지 않았다.

2.2 ISO/IEC 8000-120 Data Quality - Provenance

ISO/IEC 8000-120은 프로비넌스의 속성값 쌍과 데이터 세트들에 대한 정보를 표현하고 교환하기 위한 개념적 모델을 설명한다[2].

하지만 이 표준은 프로비넌스 단순한 이벤트로 표현하기 때문에, 빅데이터 요소에는 적절하지 않다. 그 이유는 이 표준에서 정의한 데이터 요소들이 빅데이터의 데이터 세트를 위한 프로비넌스를 정의하지 못하기 때문이다.

2.2 ITU-T Y.3600 Big Data

ITU-T Y.3600은 클라우드 컴퓨팅 기반의 빅데이터를 위한 요구사항, 역량, 유스케이스 등을 제공한다[3].

이 표준은 클라우드 서비스 관점에서 빅데이터의 마켓이나 포털의 모델과 유사하며, 프로비넌스의 구체적 요소와 같은 부분은 다루지 않는다.

III. 데이터 프로비넌스와 빅데이터 프로비넌스

<표 1>은 데이터 프로비넌스와 빅데이터 프로비넌스를 각각 정의하고 차이점을 보인다.

표 1. 데이터 프로비넌스와 빅데이터 프로비넌스의 차이점

항목	정의
데이터 프로비넌스	데이터 집약 처리 시스템에서 데이터의 출처, 생성, 전파 등을 탐지하는 제반기술 데이터 및 데이터 오브젝트의 연혁과 유도로 구성. 필요에 따라 데이터 소유권도 포함 데이터 검증, 데이터 디버깅, 데이터 감시, 데이터 품질, 데이터 신뢰성을 위해 사용 데이터베이스 관리시스템, 워크플로우 관리시스템, 분산 시스템에서 활용
빅데이터 프로비넌스	기존 프로비넌스 시스템에서 빅데이터 크기로터 발생하는 요소 고려 데이터 은닉성/안전성/프라이버시 보호 고려 분산 저장되어 있는 데이터 통합과 질의어의 최적화 고려 정형, 반정형, 비정형 데이터 등 이질적인 메타데이터 관리를 위한 상호운용성 문제 고려 빅데이터 분석 시스템의 성능 분석을 위해 프로비넌스 데이터의 특정 위치 추적 및 기록

IV. 결론 및 향후 연구

빅데이터 프로비넌스 표준은 기존 데이터 프로비넌스가 충분히 담아내지 못하는 빅데이터만의 특징들을 담아 명세 되어야 한다. 이를 위하여 이 논문은 데이터 프로비넌스 표준들의 현황을 조사하고 빅데이터 프로비넌스를 표준화하기 위한 차이점을 기술하였다.

향후로는 빅데이터 프로비넌스의 표준화를 위하여 빅데이터 프로비넌스의 유스케이스 및 요구사항을 도출하고 이를 관리할 수 있는 프레임워크를 개발할 것이다.

참 고 문 헌

[1] Paul Groth and Luc Moreau, "PROV-Overview", W3C Working Group Note, April 2013.

[2] ISO TC 184/SC 4, ISO 8000-120 Data quality - Provenance, 2009.

[3] ITU-T Y.3600 Big data - Cloud Computing based Requirements and Capabilities