http://dx.doi.org/10.14400/JDC.2015.13.1.263

빅 데이터 환경에서 다중 속성 기반의 데이터 관리 기법

정윤수*, 김용태**, 박길철**
목원대학교 정보통신융합공학부*, 한남대학교 멀티미디어학부**

Multi-Attribute based on Data Management Scheme in Big Data Environment

Yoon-Su Jeong*, Yong-Tae Kim, Gil-Cheol Park****
Dept. of Information and Communication Convergence engineering, Mokwon University*
Dept. of Multimedia Engineering, Hannam, University**

요 약 IT 기술이 발달함에 따라 센서·모바일을 기반으로 사물에 정보를 담아 네트워크로 상호연계되는 유비쿼터스 정보기술이 발달하고 있다. 그러나 서버에 저장되어 있는 데이터를 손쉽게 사용하기 위한 보안 해결책이 미미한 상태이다. 본 논문에서는 빅 데이터 서비스에서 제공되고 있는 대용량 데이터를 사용자가 안전하게 처리하기 위해서 빅 데이터 서비스에 사용되는 데이터에 다중의 속성을 해쉬 체인 기법에 적용한 데이터 관리 기법을 제안한다. 제안 기법은 빅 데이터 서비스에 사용한 데이터의 종류, 기능, 특성에 따라 데이터의 속성을 분류하여 분류된 속성 정보를 해쉬 체인으로 묶어 데이터의 안전성을 향상시켰다. 또한, 제안 기법은 여러 지역에 분산된 데이터를 손쉽게 접근하기 위해서 데이터 속성 정보를 해쉬 체인의 연결 정보로 활용하여 빅 데이터의 접근 제어를 분산 처리하였다.

주제어 : 빅 데이터, 분산환경, 데이터 처리, 다중 속성

Abstract Put your information in the object-based sensors and mobile networks has been developed that correlate with ubiquitous information technology as the development of IT technology. However, a security solution is to have the data stored in the server, what minimal conditions. In this paper, we propose a data management method is applied to a hash chain of the properties of the multiple techniques to the data used by the big user and the data services to ensure safe handling large amounts of data being provided in the big data services. Improves the safety of the data tied to the hash chain for the classification to classify the attributes of the data attribute information according to the type of data used for the big data services, functions and characteristics of the proposed method. Also, the distributed processing of big data by utilizing the access control information of the hash chain to connect the data attribute information to a geographically dispersed data easily accessible techniques are proposed.

Key Words : Big Data, Distribution Environment, Data Process, Multi-attribute

* 이 논문은 2014년도 한남대학교 학술연구 조성비 지원에 의하여 연구되었음
Received 23 October 2014, Revised 25 November 2014
Accepted 20 January 2015
Corresponding Author: Yong-Tae Kim(Hannam University)
Email: ky7762@hnu.kr

ISSN: 1738-1916

1. 서론

최근 스마트폰의 IT 기술이 발달하면서 이질적인 환경에 존재하는 빅 데이터 서비스를 제공받는 사용자가 급하고 있는 상황이다. 특히, 소셜 네트워크의 대중화로 인하여 서버에 존재하는 여러 데이터를 누구나 손쉽게 다운로드할 수있어 빅 데이터의 마케팅 노력이 매우 활발하게 진행되고 있다. 그러나, 서비스되고 있는 빅 데이터의 데이터 량이 점점 증가하면서 사용자가 요청하는 데이터의 정확도 및 처리기술에 대한 요구사항이 증가하고 있는 추세이다[1,2].

서로 다른 네트워크 환경의 이기종 장치에 저장되어 있는 빅데이터는 스마트폰과 같은 장치를 이용해서 손쉽게 사용할 수 있다[3]. 과거에는 서버에서 대용량의 데이터를 주로 처리하였지만 최근에는 소규모 용량의 데이터를 주로 사용한다. 특히, 빅 데이터 서비스는 다양한 종류의 데이터를 생성, 수집, 분석, 표현하면서 다변화된 현재 사회를 더욱 정확하게 예측 가능하며 개인화된 현대 사회 구성원 마다 맞춤형 정보를 제공, 관리, 분석 또한 가능하다.

빅 데이터는 정치, 사회, 경제, 문화, 과학 기술 등 전 영역에 걸쳐서 사회와 인류에게 가치있는 정보를 제공할 수 있는 가능성을 제시하며 그 중요성이 부각되고 있다. 그러나, 빅데이터의 문제점은 바로 사생활 침해와 보안에 있다. 빅데이터는 개인들의 수많은 정보의 집합이다. 빅데이터를 수집, 분석할 때에 개인들의 사적인 정보까지 수집하여 관리하는 빅브라더의 모습이 될 수도 있다. 그리고 수집된 데이터가 보안 문제로 유출된다면, 거의 모든 사람들의 정보가 유출되는 것이기 때문에 사회적으로 큰 문제가 야기될 수 있다[4].

본 논문에서는 이질적인 환경에서 사용자가 빅 데이터 서비스를 요청할 경우 빅 데이터 서비스를 효율적으로 서비스받기 위한 데이터 관리 기법을 제안하다. 제안된 기법은 대용량 데이터를 사용자가 안전하게 처리하기 위해서 빅 데이터 서비스에 사용되는 데이터에 다중의 속성을 해쉬 체인 기법에 적용한다. 제안 기법은 빅 데이터 서비스에 사용한 데이터의 종류, 기능, 특성에 따라 데이터의 속성을 분류하여 분류된 속성 정보를 해쉬 체인으로 묶어 데이터의 처리 속도를 향상시켰다. 또한, 제안 기법은 여러 지역에 분산된 데이터를 손쉽게 접근하기 위해서 데이터 속성 정보를 해쉬 체인의 연결 정보로 활용하여 빅 데이터의 접근 제어를 분산 처리하였다.

이 논문의 구성은 다음과 같다. 2장에서는 빅데이터의 정의 및 특징을에 대해서 알아본다. 3장에서는 데이터 속성 정보를 이용한 데이터 처리 기법을 제안하고, 4장에서는 제안 기법의 성능평가를 분석하고 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

2.1 빅데이터

빅데이터란 과거 아날로그 환경에서 생성되던 데이터에 비해 그 규모가 방대하며 생성 주기가 짧고, 형태가 수치 데이터 뿐만 아니라 문자와 영상 데이터를 포함하는 대규모 데이터를 의미한다[1]. 최근 PC와 인터넷, 모바일 기기 등이 생활화 되면서 시간과 장소에 구애받지 않고 손쉽게 사이버 공간에서 사용 및 저장한 데이터가 기하급수적으로 증가하고 있다. 이 같은 현상은 사람과 기계, 기계와 기계가 서로 정보를 주고받는 사물지능통신(M2M, Machine to Machine)의 확산도 디지털 정보가 폭발적으로 증가하게 된 이유이다.

사용자가 직접 제작하는 UCC를 비롯한 동영상 컨텐츠, 휴대전화와 SNS(Social Network Service)에서 생성되는 문자 등은 데이터의 증가 속도뿐만 아니라, 형태와 질에서도 기존과 다른 양상을 보이고 있다. 특히, 블로그나 SNS에서 유통되는 텍스트 정보는 내용을 통해 글을 쓴 사람의 성향뿐만아니라 소통하는 상대방의 연결 관계까지도 분석이 가능하다. 또한 주요 도로와 공공건물은 물론 심지어 아파트 엘리베이터 안에까지 설치된 CCTV가 촬영하고 있는 영상 정보도 데이터로 저장되고 있다. 그리고, 민간 분야뿐만 아니라 공공 분야도 데이터를 양상 중인데 센서스(Census)를 비롯한 다양한 사회 조사, 국제자료, 의료보험, 연금 등의 분야에서 데이터가 생산되고 있다[5].

2.2 빅데이터 특징

빅데이터는 일반적으로 3V,데이터의 양(Volume), 데이터 생성 속도(Velocity), 형태의 다양성(Variety) 등의 특징을 가진다. 빅데이터의 다양하고 방대한 규모의 데

이터는 국가 경쟁력의 우위를 좌우하는 중요한 자원으로 활용되고 있지만 과거와 비교해 데이터의 양은 물론 질과 다양성 측면에서 패러다임의 전환이 필요하다[2,3].

빅데이터는 분산처리방식과 같은 기술을 활용해서 과거에 비해 대규모 고객정보를 빠른 시간 안에 분석하는 것이 가능해졌다. 트위터와 인터넷에서 생성되는 기업 관련 검색어와 댓글을 분석해 자사의 제품과 서비스에 대한 고객 반응을 실시간으로 파악해 즉각적인 대처를 수행할 수 도 있다.

빅데이터에서는 소프트웨어나 하드웨어도 오픈 소스 형태의 하둡(Hadoop)이나 분석용 패키지인 R 과 분석병렬처리기술, 클라우드 컴퓨팅 등을 활용하기 때문에 기존의 비싼 스토리지와 데이터베이스에 기반한 고비용의 데이터웨어하우스를 구축하지 않아도 효율적인 시스템 운용이 가능하다[4,6].

3. 다중속성 기반 데이터 관리기법

이 절에서는 빅 데이터 서비스에 사용되는 데이터를 종류, 기능, 특성에 따라 속성값을 부여하여 계층적으로 데이터를 분류하고, 분류된 데이터의 속성정보를 사용자가 안전하게 서비스 받을 수 있도록 해쉬 체인에 적용하여 데이터를 분산 처리하도록 한다.

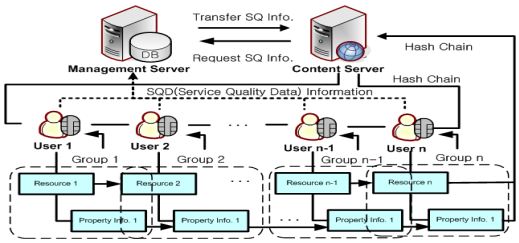
3.1 개요

빅 데이터 서비스에서 사용되고 있는 데이터의 종류는 서비스 환경에 따라 다양하게 사용되고 있다. 제안 기법에서는 빅 데이터 서비스를 제공받는 사용자의 원활한 서비스를 위해서 빅 데이터에서 제공되는 기본 적인 데이터는 사전에 서버에 등록되고 서비스되어야 한다. 빅 데이터를 제공하는 서비스 환경에서는 데이터를 구성하고 관리하여야 한다.

빅 데이터 환경에서 사용되는 데이터는 다양하며 복잡한 특성을 가진다. 이러한 데이터의 특성을 빅 데이터 서비스에 반영하기 위해서는 데이터에 다양한 속성을 반영해야 한다.

제안 기법은 [Fig. 1]처럼 대규모의 데이터와 데이터의 속성정보를 분산 처리 및 저장 관리할 수 있도록 저장될 파일을 블록 단위로 데이터를 나누어 분산된 서버에 저장할 수 있도록 구성한다.

[Fig. 1]에서 제안 모델은 빅 데이터와 그 데이터에 속한 속성값을 계층적으로 구성한다. 데이터와 자원은 각각 해쉬체인을 통해 데이터를 하나의 그룹으로 묶어 데이터의 다양한 속성에 따른 계층적 서비스를 수행하도록 한다.



[Fig. 1] System Constructure of Proposed Scheme

[Fig. 1]처럼 제안 기법은 장애가 발생할 경우 장애 복구 능력을 갖추는 보조 네임노드(secondary namenode)가 이중 해쉬를 이용하여 사용자가 데이터에 쉽게 접근하는 특징이 있다. 또한, 악의적인 데이터노드가 네임노드인척 가장할 경우, 데이터의 종류, 기능, 특성에 따라 데이터를 해쉬 체인으로 묶어 데이터에 높은 처리량을 지원한다. 데이터 노드가 공격자에게 노출되었을 때 발생하는 보안 취약점을 해결하기 위해서 데이터의 속성정보를 해쉬 체인의 연결 정보로 활용하여 빅 데이터의 접근 제어를 분산 처리한다.

3.2 용어 정의

<Table 1>은 제안 기법에서 사용하는 용어에 대한 설명이다.

〈Table 1〉 Notations

Notation	Definition
$\overrightarrow{D}$	Data Information
d_i	i^{th} Data
$\overline{d}$	Group of all property value related to p_j
U_i	i^{th} User
S_i	i^{th} Server
$E()$	Encryption
$D()$	Decryption
h_i	Information adopt to pairs of data and property value to hashchain function
$H()$	one-way hash chin

3.3 계층적 속성 기반 데이터 처리 기법

이 절에서는 사용자에게 제공하는 빅 데이터의 사용 목적에 따라 데이터의 속성정보(종류, 기능, 특성)을 부여하여 계층적으로 데이터를 관리할 수 있도록 데이터를 해쉬체인에 적용한 기법을 제안한다.

3.3.1 데이터 속성 정보 생성 과정

이 절에서는 데이터의 사용목적에 따라 데이터의 속성 정보를 생성하여 이기종 장치에 저장된 데이터를 계층적으로 관리할 수 있도록 데이터의 속성 정보를 생성하는 과정을 3단계로 구성한다.

- 단계 1 : 빅 데이터 중 사용자가 서비스를 원하는 데이터 정보 $\vec{D}$를 식 (1)처럼 생성한다. 여기서 d_i는 빅 데이터 서비스 중 서비스를 제공받고자 하는 데이터를 의미한다.

$$\vec{D} = (d_1,\ d_2,\ ...,\ d_n) \tag{1}$$

- 단계 2 : 데이터 정보 $\vec{D}$ 중 서비스에 사용되는 데이터 특성에 따라 데이터 d_i에 식 (2)처럼 속성값을 생성한다.

$$d_j = (p_1,\ p_2,\ ...,\ p_n) \tag{2}$$

여기서 p_j은 데이터 특성 값을 의미하며 j는 집합 Z의 원소($j \in Z$)이다. p_j와 관계가 있는 모든 특성 값들의 집합을 식 (3)처럼 나타낸다.

$$\bar{d} = \{dp_i \in Z \mid d_j \sim dp_i\},\ \ 1 \le i \le n,\ 1 \le j \le n \tag{3}$$

- 단계 3 : 데이터 정보 $\vec{D}$와 p_j와 관계가 있는 모든 특성 값들의 속성 집합 $\bar{d}$을 식 (5)처럼 해쉬 함수에 적용하여 데이터의 속성 정보 PI_i를 생성한다.

$$H_U : \{0,1\} \rightarrow Z_N \tag{4}$$

$$PI_i = H(\vec{D}, \bar{d}),\ \ 1 \le i \le n \tag{5}$$

여기서, $H_i : \{0,1\} \rightarrow Z_N$는 서버가 사용하는 안전한 해쉬함수를 의미하고, 데이터의 속성 정보 PI_i는 데이터 정보를 복구 및 조회하기 위해 사용된다.

3.3.2 데이터 접근제어 과정

이 과정은 사용자 U_i가 서버 S_i에게 빅 데이터 서비스를 요구할 경우 서버가 사용자가 원하는 정보만을 추출하여 서비스하기 위한 과정이다.

- 단계 1 : 사용자 U_i는 빅 데이터 서비스를 제공받기 원하는 데이터를 서버에게 요청한다.

- 단계 2 : 서버 S_i는 데이터베이스에 저장되어 있는 정보 중 사용자가 요청한 데이터 정보 $\vec{D}$를 수집한다.

$$Gathering\ \vec{D} = (d_1,\ d_2,\ ...,\ d_n) \tag{6}$$

- 단계 3 : 서버 S_i는 수집한 데이터 정보 $\vec{D}$에 대해서 속성 정보 d_i를 부여하고 해당 데이터의 종류, 기능, 특성에 따라서 속성 집합 $\bar{d}$를 생성한다. 서버는 데이터 정보 $\vec{D}$와 속성 집합 $\bar{d}$를 식 (7)처럼 해쉬 함수 $H()$에 적용하여 데이터의 속성 정보 PI_i를 생성한다.

$$PI_i = H(\vec{D}, \bar{d}),\ \ 1 \le i \le n \tag{7}$$

- 단계 4 : 서버 S_i는 사용자 U_i에게 빅 데이터와 함께 데이터의 속성 정보를 식 (8)처럼 전달한다.

$$Data' = \sum_{i=1}^{n} H(Data,\ PI_i) \tag{8}$$

- 단계 5 : 서버 S_i는 사용자 U_i에게 전달한 식 (8)의 정보를 실시간으로 모니터링하며, 사용쟈는 전달받은 데이터를 속성 정보 레벨에 맞게 사용한다.

4. 성능 평가

이 절에서는 서버 S_i와 사용자 U_i 사이의 통신 범위에서 안전한 통신이 이루어진다는 가정한다. 제안 기법의 성능평가는 처리 지연시간, 통신 오버헤드, 통신 복잡도와 계산 복잡도 등으로 평가한다.

4.1 환경설정

제안 기법의 성능 평가를 위해 각각의 사용자 U_i가 서

비스를 요청할 경우 시간 t는 0으로 설정하고 threshold th는 {1, 3, 5}로 설정한다. 서버 S_i는 사용자 U_i가 빅 데이터 서비스를 요청할 경우 데이터와 함께 데이터 정보 $\vec{D}$를 데이터와 함께 전달한다고 가정한다.

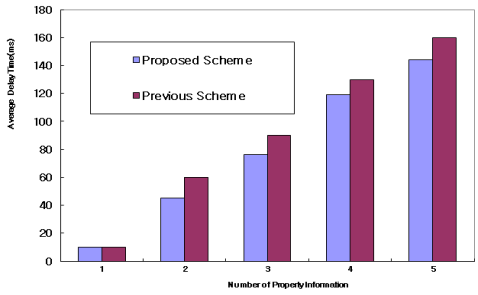
〈Table 2〉 Simulation Setting

Parameter	Setting
Simulation area	500m × 500m
Number of users	l = {100, 500, 1000}
Number of Data	d={10,000, 50,000}
Number of Property	p = {1, 2, 3, 4, 5}
Simularity threshold	th = {1, 3, 5}
Transmission of smartphone	20m
Data generation interval	0.01 ms

4.2 성능분석

4.2.1 속성 수에 따른 처리 지연시간

[Fig. 2]는 속성 수에 따른 데이터 처리 지연시간을 나타내고 있다.

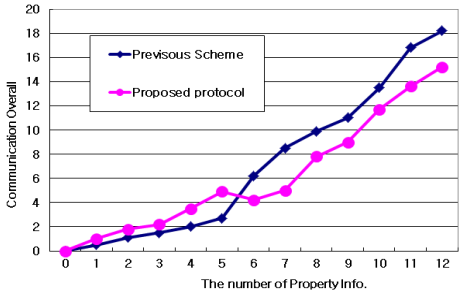


[Fig. 2] Average Delay Time thourgh the Number of Property Info.

[Fig. 2]의 결과, 속성수가 적을수록 기존기법과 제안기법의 차이는 없었지만, 데이터에 속한 속성수가 많은 제안 기법이 데이터에 속한 속성수가 없는 경우에 비해 처리 지연 시간이 7.3% 향상되었다. 이 같은 결과는 빅 데이터 환경에서 많은 데이터를 사용자가 처리할 경우 사용자가 원하는 데이터만을 추출하여 처리하기 때문에 나타난 결과이다.

4.2.2 통신 오버헤드

[Fig. 3]은 속성 수에 따른 서버와 사용자간 통신 오버헤드를 나타내고 있다. [Fig. 3]의 결과, 속성수가 낮을 경우 서버와 사용자간 통신 오버헤드는 제안 기법이 기존 기법 보다 오버헤드가 평균 9.7% 높았지만 속성수가 증가할수록 제안기법이 기존기법보다 통신 오버헤드가 11.8% 낮았다. 이 같은 결과는 사용자가 요청한 데이터의 양이 적을 경우 제안기법은 데이터를 해쉬함수를 통해 계층화 하기 때문에 데이터 처리 시간이 초기에 일정 시간 증가하기 때문에 나타난 결과이다. 따라서, 제안기법은 데이터의 양이 작을 때보다 데이터 양이 많을수록 통신 오버헤드가 낮아지는 결과를 얻었다.



[Fig. 3] Communication Overall thourgh the Number of Property Info.

4.2.3 복잡도

그림 4는 데이터 수에 따른 제안기법의 통신 복잡도와 계산 복잡도를 나타내고 있다. 서버 S_i와 사용자 U_i 사이에서 요구되는 처리 비용은 다음과 같이 계산한다. n은 사용자의 수이고 r은 통신 범위를 의미한다. 속성 정보의 데이터 속성 정보는 사용자 U_i 당 $n \times r = \log_2 \frac{N}{M} + \log_2 M = \log_2 N$이다. 서버 S_i는 $n \times r = (2*\frac{N}{M}) + \frac{N}{M} * (2*\log_2 M) = 2*\frac{N}{M}(\log_2 m + 1) - 1$이다. 통신 비용은 $2*\log_2 \frac{N}{M} + \log_2 M = 2*\log_2 N - \log_2 M$이다. 여기서 M과 N은 통신범위내 사용자 U_i와 서버 S_i가 사용자 U_i의 데이터 속성 정보를 처리하는 수이고 통신범위 내의 사용자들은 데이터의 속성 정보를 모두 공유되는 것

으로 나타낸다.

5. 결론

최근 휴대폰 기술의 발달로 인하여 빅 데이터 서비스가 활용이 점점 증가하고 있다. 본 논문에서는 대용량의 빅 데이터 서비스를 사용자가 끊김없이 받을 수 있는 데이터 관리 기법을 제안하였다. 제안 기법은 속성 수에 따른 데이터 처리 시간은 7.3% 향상시켰으며, 통신오버는 11.8% 향상시켰다. 또한, 데이터의 종류, 기능, 특성에 따라 데이터의 통신 오버헤드 및 계산 오버헤드는 데이터의 속성수가 증가할 경우와 데이터 수가 많을 경우 기존 기법보다 오버헤드가 낮게 나타났다. 향후 연구로 본 연구의 결과를 기반으로 빅데이터 시스템에 실제 적용할 계획이다.

ACKNOWLEDGMENTS

This paper has been supported by 2014 Hannam University Research Fund.

REFERENCES

[1] J. Manyika and M. Chui, “Big data: The next frontier for innovation, competition, and productivity”, McKinsey Global Institute, pp. 1. 2011.

[2] P. Russom, “Big Data Analytics”, TDWI Research Fourth Quarter, pp. 6. 2011.

[3] Y. C. Jung. “Big Data revolution and media policy issues”, KISDI Premium Report, Vol. 12, No. 2, pp. 1-22.2012.

[4] S. Y. Son, “Big data, online marketing and privacy protection”, KISDI Premium Report, Vol. 13, No. 1, pp.1-26.2013

[5] H. Amur, J. Cipar, V. Gupta, G. R. Ganger, M. A. Kozuch, and K. Schwan, “Robust and flexible power-proportional storage”, In SoCC '10: Proceedings of the 1st ACM symposium on Cloud computing, pp. 217-228.2010

[6] J. Leverich and C. Kozyrakis. “On the energy (in)efficiency of hadoop clusters”. SIGOPS Oper. Syst. Rev., 44(1):61-65. 2010

정 윤 수(Jeong, Yoon Su)

- 2000년 2월 : 충북대학교 대학원 전자계산학 이학석사
- 2008년 2월 : 충북대학교 대학원 전자계산학 박사
- 2009년 8월 ~ 2012년 2월 : 한남대학교 산업기술연구소 전임연구원
- 2012년 3월 ~ 현재 : 목원대학교 정보통신공학과 조교수
- 관심분야 : 센서 보안, 암호이론, 정보보호, Network Security, 이동통신보안
- E-Mail : bukmunro@gmail.com

김 용 태(Kim, Yong Tae)

- 1984년 2월 : 한남대학교 계산통계학과 학사
- 1988년 2월 : 숭실대학교 전자계산학과 석사
- 2008년 2월 : 충북대학교 전자계산학과 박사
- 2002년 12월 ~ 2006년 2월 : (주)가림정보기술 이사
- 2010년 10월 ~ 현재 : 한남대학교 멀티미디어학부 교수
- 관심분야 : 모바일 웹서비스, 정보 보호, 센서 웹, 모바일 통신보안
- E-Mail : ky7762@hannam.ac.kr

박 길 철(Park, Gil Cheol)

- 1983년 2월 : 한남대학교 계산통계학과 학사.
- 1986년 2월 : 숭실대학교 전자계산학과 석사.
- 1998년 2월 : 성균관대학교 전자계산학과 박사.
- 1998년 8월. ~ 현재 : 한남대학교 멀티미디어학부 교수
- 2005년 2월 : 한국정보기술학회 이사 멀티미디어 분과 위원장
- 관심분야 : Multimedia And Mobile Communication, Network Security
- E-Mail : gcpark@hnu..kr