http://dx.doi.org/10.14400/JDC.2015.13.9.201

빅 데이터의 처리속도 향상을 위한 확률기반 서브넷 선택 기법

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

Subnet Selection Scheme based on probability to enhance process speed of Big Data

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 SNS와 페이스북과 같은 서비스가 대중화되면서 마이크로블로그와 같은 작은 크기의 빅 데이터 사용이 증대되고 있다. 그러나, 현재까지 작은 크기의 빅 데이터의 탐색 결과의 정확성과 계산비용은 미해결 상태로 남아있다. 본 논문에서는 빅 데이터 환경에서 마이크로블러그와 같은 작은 크기의 텍스트 정보의 탐색 속도를 향상시키기 위한 확률기반의 서브넷 선택 기법을 제안한다. 제안 기법은 데이터의 속성 정보에 확률값을 부여하여 서브넷을 구성하여 데이터 탐색 속도를 높였다. 또한, 제안 기법은 분산된 데이터를 손쉽게 접근하기 위해서 서브넷을 구성하는 데이터 의확률값 간 연계 정보를 쌍으로 처리함으로써 데이터의 접근성을 향상시켰다. 실험결과, 제안 기법은 CELF 알고리즘보다 평균 6.8% 높은 탐지율을 보였으며, 처리시간은 평균 8.2% 단축시켰다.

주제어 : 빅 데이터, 데이터 속도, 확률, 서브넷 선택, 다중 속성

Abstract With services such as SNS and facebook, Big Data popularize the use of small size such as micro blogs are increasing. However, the problem of accuracy and computational cost of the search result of big data of a small size is unresolved. In this paper, we propose a subnet selection techniques based probability to improve the browsing speed of the small size of the text information from big data environments, such as micro-blogs. The proposed method is to configure the subnets to give to the attribute information of the data increased the probability data search speed. In addition, the proposed method improves the accessibility of the data by processing a pair of the connection information between the probability of the data constituting the subnet to easily access the distributed data. Experimental results showed the proposed method is 6.8% higher detection rates than CELF algorithm, the average processing time was reduced by 8.2%.

Key Words : Big Data, Data Speed, Probability, Subnet Selection, Multi-attribute

* 이 논문은 2015년도 한남대학교 학술연구조성비 지원에 의하여 연구되었음
Received 7 July 2015, Revised 29 August 2015
Accepted 20 September 2015
Corresponding Author: Gil-Cheol Park(Hannam University)
Email: gcpark@hnu.kr

ISSN: 1738-1916

1. 서론

빅 데이터는 과거 아날로그 환경에서 디지털 환경으로 변화하면서 다양한 형태의 데이터뿐만 아니라 문자와 영상 데이터를 모두 포함한다[1,2].

최근 인터넷이나 스마트폰을 사용하여 직접 제작한 UCC나 동영상 콘텐츠, 휴대전화와 SNS(Social Network Service)에서 생성되는 문자 등을 인터넷에서 이용하고 있으며, 데이터의 증가 속도, 형태와 질 측면에서 기존 서비스와 다른 양상을 보이고 있다[3,4,5]. 특히, 블로그나 SNS에서 유통되는 텍스트 정보는 내용을 통해 글을 쓴 사람의 성향뿐만 아니라 소통하는 상대방의 연결 관계까지도 분석이 가능하다. 사진이나 동영상 콘텐츠를 PC를 통해 이용하는 것은 이미 일반화되었고 방송 프로그램도 TV 수상기를 통하지 않고 PC나 스마트폰으로 보고 있다.

인터넷과 스마트폰에서 제공되는 서비스가 다양화 되면서 빅 데이터의 데이터 량이 점점 증가하고 사용자가 요청하는 데이터의 정확도 및 처리기술에 대한 요구사항도 현재 증가하고 있는 추세이다[3,5]. 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, 빅 데이터는 다양한 종류의 데이터를 생성, 수집, 분석, 표현하면서 개인화된 현재 사회 구성원 마다 맞춤형 정보를 제공, 관리, 분석하기 위한 처리속도가 매우 중요한 요구사항으로 부각되고 있다.

빅 데이터와 관련된 최근 연구에서는 마이크로블로그(Microblog)의 텍스트 내용을 효과적으로 분석하는데 초점을 두고 있으며, 마이크로블러그는 실시간으로 생성되는 짧은 크기의 많은 마이크로블로그로 인하여 통신량과 계산 비용을 야기하는 문제점을 가지고 있다[6,7,8]. 만약 마이크로블로그에서 수집된 데이터가 보안 문제로 유출된다면, 거의 모든 사람들의 정보가 유출될 수 있는 문제가 있다[4].

본 논문에서는 빅 데이터 환경에서 사용자가 원하는 정보 즉, 마이크로블로그와 같은 작은 크기의 텍스트와 같은 데이터의 탐색 처리 속도를 향상시키기 위한 확률기반의 서브넷 선택 기법을 제안한다. 제안 기법은 서로 다른 유형의 마이크로블로그를 종류, 기능, 특징에 따라 속성을 부여하고 속성들에 대한 확률값에 따라 데이터를 서브넷으로 묶어 데이터 처리 속도를 향상시킨다. 또한, 제안 기법은 여러 지역에 분산된 데이터를 손쉽게 접근하기 위해서 서브넷 내 확률값이 높은 데이터의 속성 정보를 연계 정보로 연결하여 처리함으로써 데이터의 접근성을 향상시켰다.

이 논문의 구성은 다음과 같다. 2장에서는 빅데이터의 정의 및 특징을에 대해서 알아본다. 3장에서는 빅 데이터 처리 속도 향상을 위한 확률기반 서브넷 선택 기법을 제안하고, 4장에서는 제안 기법을 CELF 알고리즘과 비교 평가하고 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

2.1 빅데이터

빅 데이터는 정형 또는 비정형 데이터의 수용한계를 넘어서는 데이터로부터 정보를 추출 및 분석하는 기술을 의미한다[1,9,10].

빅 데이터의 분석 기술은 다양한 분야에서 사용되었던 데이터 마이닝, 자연 언어 처리, 패턴 인식 등과 같은 기술들이 있다. 이 기술들은 소셜 미디어와 같은 비정형 데이터 분석에 주목 받고 있다.

빅 데이터 정보는 다양한 분야에서 사용되고 있다. 예를들어, 블로그나 SNS에서 유통되는 텍스트 정보는 사용자의 성향과 연결 관계 분석에 사용되며, 영상 정보는 주요 도로와 공공건물, 아파트 엘리베이터 등에서 사용된다. 빅 데이터 정보는 민간 분야뿐만 아니라 공공 분야, 센서스(Census)를 비롯한 다양한 사회 조사, 국제자료, 의료보험, 연금 등 다양한 분야에서 만들어 지고 있다[5,11].

2.2 빅 데이터 특징

빅 데이터는 다양하고 방대한 규모의 데이터가 수십 테라바이트에서 수 페타바이트에 이르는 것이 특징이다[12,13]. 가트너는 2001년 연구보고서에서 데이터의 급성장에 따른 이슈와 기회를 일반적으로 3V, 데이터의 양(Volume), 데이터 입출력 속도(Velocity), 데이터 종류의 다양성(Variety)으로 정의하였다. 가트너의 3V 정의는 현재 널리 사용되고 있지만, 데이터와 그 사용방법에 있

어서 대상을 측정하고 경향을 예측하는 등의 일을 하기 위해 고밀도의 데이터로 구성된 기술적 통계를 활용하는 측면에서 경영정보학과 차이가 있다[4,14].

2.3 빅 데이터 탐지 알고리즘

현재까지 연구된 빅 데이터 탐지 알고리즘은 마이크로블로그의 텍스트 내용을 분석하는데 관심을 갖고 있다[15]. CELF 알고리즘은 빅 데이터 탐지 알고리즘 중 마이크로블로그의 텍스트를 탐지하는 가장 대표적인 알고리즘이다[6]. CELP 알고리즘은 인간의 청각 특성을 이용해서 아날로그 음성 신호를 디지털 데이터로 변환하는 부호화 방식으로서 1984년 미국 AT&T가 개발하였다. CELF 알고리즘은 모든 이벤트를 가능한 모두 탐지하기 위해서 서브넷을 선택한다. 그러나, CELF 알고리즘은 음성 신호의 입력 데이터로부터 출력까지의 지연 시간이 비교적 길다는 단점이 있고, CELF 알고리즘은 최적의 서브넷을 선택하는 부분에서 mixed-integer 최적화 문제를 해결하지 못하였다.

[7]은 자원 제약에 주요 이벤트들을 온라인에서 탐지하는 방법을 제안하였다. 그러나 이 방법은 작은 마이크로블로그와 같은 데이터를 효율적으로 탐지하기 위해서 작은 서브넷을 선택하고 모니터링해야 하는 문제가 있다. [8]에서는 데이터 탐색에 최대 영향을 미치는 노드의 서브넷이 선택되고, 영향은 하나의 노드에서 다른 노드의 확률값의 진행정도를 고려하고 있다. 비록 이 기법이 최대 영향을 가지는 서브넷이 많은 이벤트들이 참여하는 것을 의미하는 것이 아닐 지라도 진행하는 확률이 크게 영향을 미치는 부분에서 특징이 있다.

3. 확률기반 서브넷 선택기법

이 절에서는 빅 데이터 서비스에서 사용되는 데이터의 종류, 기능, 특성에 따라 데이터의 속성값을 확률적으로 부여함으로써 데이터의 탐색 정확도를 향상시켰으며, 데이터의 확률값에 따라 유사 데이터를 서브넷으로 묶음으로써 데이터 처리속도를 향상시켰다. 서브넷으로 분류된 데이터는 속성값 이외에 서로 연계할 수 있는 정보들을 쌍으로 묶어 이질적인 환경에서도 관련 데이터를 처리할 수 있도록 하였다.

3.1 개요

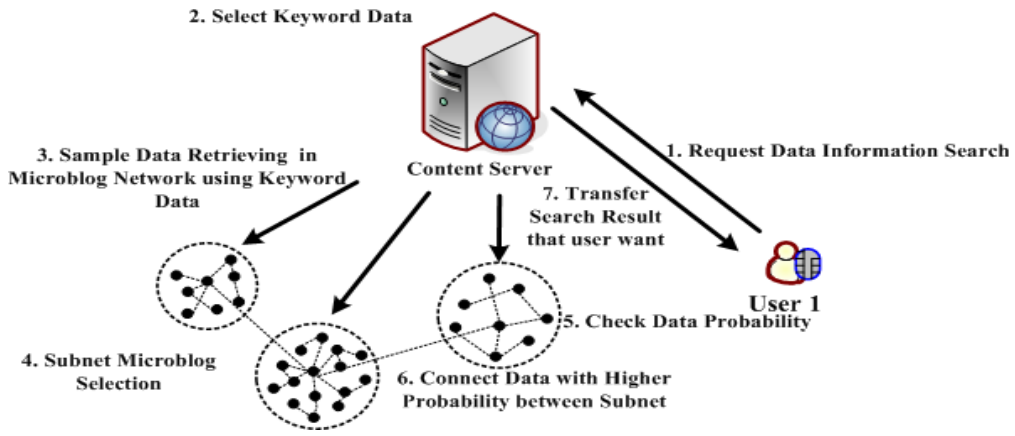
빅 데이터는 다양한 종류의 데이터가 다양한 환경에서 사용되고 있다. 특히, 데이터의 양이 작은 마이크로블로그와 같은 데이터는 하루에도 수많은 주제로 생성되고 삭제되고 있다. SNS와 페이스북과 같은 서비스에서는 데이터를 정확하게 탐색하여 서비스를 제공받는 것이 중요하다. 본 논문에서는 사용자가 원하는 데이터를 빅 데이터 환경에서 정확하면서도 빠르게 찾기 위해서 데이터의 속성정보에 확률정보를 부여하여 동일 정보를 가지는 데이터를 서브넷으로 구성하여 사용자가 원하는 데이터와 가장 부합되는 서브넷을 선택할 수 있도록 하는 것을 목적으로 한다.

제안 기법에서 사용되는 데이터는 다양하고 복잡한 특성을 가지고 있기 때문에 빅 데이터를 구성하는 데이터는 손쉽게 구성하고 관리하도록 사전에 서버에 등록되고 서비스하는 것으로 가정한다. 이 때, 데이터는 데이터의 종류와 특성에 맞게 다양한 속성을 반영한다.

제안 기법은 [Fig. 1]처럼 대규모의 데이터와 데이터의 속성정보를 분산 처리 및 저장 관리할 수 있도록 서브넷으로 데이터를 나누어 분산처리 할 수 있도록 구성한다. 서브넷의 크기는 데이터의 유사 속성정보의 상관관계에 따라 달라질 수 있다.

[Fig. 1]은 제안 모델의 전체 프로세스를 보여주고 있다. [Fig. 1]처럼 제안 기법은 전체 7 단계로 구성되며 빅 데이터와 그 데이터에 속한 속성값은 서브넷을 구성하는 정보로 활용되며, 서브넷을 구성하는 데이터의 가장 높은 확률 정보는 다른 서브넷을 구성하는 가장 높은 확률 정보와 서로 연계할 수 있도록 연계정보를 (확률값, 서브넷정보) 쌍으로 구성하여 서비스를 수행한다. [Fig. 1]의 7단계의 세부동작은 다음과 같다.

- 1단계 : 사용자는 컨텐츠 서버에게 데이터 정보 검색을 요청한다.
- 2단계 : 컨텐츠 서버는 사용자로부터 요청된 정보를 찾기 위해 정보를 기반으로 해서 키워드 데이터를 선택한다.
- 3단계 : 컨텐츠 서버는 키워드 데이터를 사용하여 마이크로블러그내 데이터에서 키워드 데이터와 유사한 샘플 데이터를 검색한다.
- 4단계 : 검색된 샘플 데이터를 중심으로 서브넷을 구성한다.



[Fig. 1] Overall Process of Proposed Scheme

- 5단계 : 마이크로블러그내 데이터와 유사한 샘플 데이터를 유사도 평가를 통해 데이터들의 확률값을 체크한다.
- 6단계 : 확률값을 기준으로 서브넷을 구성하고 있는 데이터 중 확률값이 높은 데이터를 중심으로 연결정보를 만든다.
- 7단계 : 확률값이 가장 높은 데이터를 선택하여 사용자가 원하는 정보인지를 판별한 후 동일하다면 사용자에게 정보를 전달한다.

3.2 용어 정의

<Table 1>은 제안 기법에서 사용하는 용어에 대한 설명이다.

〈Table 1〉 Notations

Notation	Definition
N	The number of user
U_i	i^{th} User
$\hat{U}$	User of $\hat{U} \subseteq U$
D_i	i^{th} Dataset
p_i	i^{th} Data Property
DP_i	i^{th} Data Probability included in i^{th} Dataset

3.3 연계정보를 이용한 서브넷 선택 기법

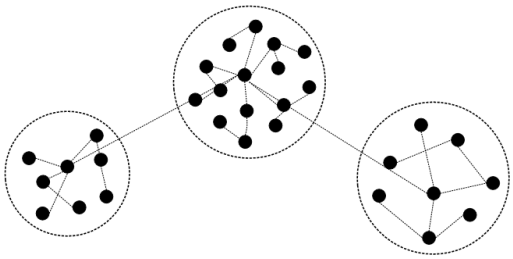
이 절에서는 빅 데이터의 사용목적에 따라 데이터의 속성정보(종류, 기능, 특성)을 확률값으로 표현하여 유사정도에 따라 서브넷을 구성하여 구성된 서브넷을 서로 연계할 수 있는 연계정보를 확률값과 쌍으로 구성하여 서비스를 제공하기 위한 서브넷 선택 기법을 제안한다.

3.3.1 샘플링 데이터 검색 과정

인터넷이나 스마트폰에 의해 생성되는 빅 데이터는 매일 수십만건이 생성되지만 생성되는 빅 데이터를 예측하기 위해서 네트워크상에 존재하는 모든 빅 데이터를 처리하는 것은 불가능하다.

빅 데이터의 통신로드와 계산로드를 줄이기 위해서 본 논문에서는 데이터의 속성정보에 확률값을 부여하여 검색 데이터와의 가장 부합되는 유사 정보를 검색하도록 서브넷을 구성하도록 한다. 이 과정은 매 순간 생성되는 데이터를 샘플 데이터셋으로 수집 관리 할 수 있는 특징을 가진다.

확률기반으로 유사성이 높은 데이를 구성하는 서브넷들은 서브넷을 대표하는 가장 높은 확률값을 가진 데이터와 연계정보를 통해 연계한다. 이 때, 서브넷을 구성하는 데이터는 확률값에 따라 계층적으로 구성된다.



[Fig. 2] Subnet Construction based Probability

3.3.2 서브넷 선택 과정

서브넷을 선택하기 위해서 우선 먼저 제안 기법에서는 빅 데이터의 수많은 데이터 중 확률값이 높은 속성 정보를 가진 데이터를 샘플링하여 데이터셋(dataset)을 만든다. 이 때, 데이터셋은 정확도와 속도를 높이기 위해서 확률값이 일정 수준 아래에 있는 데이터, 즉 threshold(P<0.3) 보다 적은 데이터는 필터링한다. 여기서, threshold를 0.3을 기준값으로 설정한 이유는 제안 기법이 확률기반으로 서브넷을 구성하기 때문에 0.3보다 적은 threshold는 데이터의 수가 낮아 서브넷 생성 과정에서 비효율적인 서브넷 생성이 이루어지기 때문이다.

필터링 과정이 끝나면 서브넷을 선택하기 위한 샘플링 데이터셋의 크기가 줄어들게 된다. 제안 기법에서는 확률값이 높은 속성정보를 추출하기 위해서 통신비용과 계산비용을 최소화할 수 있는 데이터를 탐지하기 위한 데이터셋을 선택하여 서브넷을 생성한다.

3.3.3 서브넷 생성 과정

이 절에서는 데이터를 속성정보에 따라 확률값을 부여하여 데이터를 계층적으로 구성하여 서브넷을 생성·관리한다.

제안기법에서는 빅 데이터 서비스를 제공받으려는 사용자가 식 (1)처럼 N명이라고 가정한다. 빅 데이터 서비스를 제공받으려는 N 명의 사용자 U중 컨텐츠 서버로부터 정확한 데이터를 탐지하여 수신하는 사용자 U_i는 식 (2)처럼 $\hat{U} \subseteq U$와 같다.

$$U = \{u_1,\ u_2,\ ...,\ u_N\}\ ,\ i \in [1,N] \quad (1)$$

$$U_i = \{\hat{U} \subseteq U \mid i \in [1,N]\} \quad (2)$$

식 (2)에서 사용자 $U_i(i \in [1,N])$는 식 (3)처럼 데이터 d를 N개 샘플링하여 데이터셋(dataset) D_i를 만들기 위 식 (4)처럼 데이터셋을 $D_i(i \in [1,N])$ 로 설정하며, 샘플링되는 데이터셋은 $D_i \subseteq D$와 같다.

$$D = \{d_1,\ d_2,\ ...,\ d_N\}\ ,\ i \in [1,N] \quad (3)$$

$$D_i = \{D_i \subseteq D \mid i \in [1,N]\} \quad (4)$$

여기서, d_i는 빅 데이터 서비스 중 서비스를 제공받고자 하는 데이터를 의미한다.

제안기법에서는 데이터 D_i 중 서비스에 사용되는 데이터 특성에 따라 데이터 D_i에 식 (5)처럼 속성값을 부여한다. 이때 확률값이 높은 데이터는 데이터 특성 따라 생성된 데이터 D_i는 서브넷을 구성하여 속성 정보 p_i와 함께 컨텐츠 서버에 전달하여 저장된다. 컨텐츠 서버에 저장된 (D_i, p_i) 정보는 대규모 빅 데이터를 계층적으로 분산 저장함으로써 데이터 관리 및 추출이 손쉬워진다.

$$D_i = (p_1,\ p_2,\ ...,\ p_n)\ ,\ i \in [1,N] \quad (5)$$

여기서 p_i은 데이터 특성 값을 의미하며 i는 집합 Z의 원소($i \in Z$)이다. p_i와 관계가 있는 모든 특성 값들의 집합을 식 (6)처럼 나타낸다.

$$\bar{d} = \{dp_i \in Z \mid d_j \sim dp_i\},\quad 1 \le i \le n,\ 1 \le j \le n \quad (6)$$

데이터셋에 포함된 데이터 확률은 $DP_i(i \in [1,N])$로 나타내며, 데이터 특성 값이 부여된 데이터 $\bar{d}$는 데이터 특성값에 따라 식 (7)처럼 이진 확률정보를 부여한다.

$$P_i = \begin{cases} 1 & \text{if } d_i\ participated \in dataset \\ 0 & otherwise \end{cases} \quad (7)$$

여기서 데이터 확률은 $DP_i = \Pr(P_i=1)$이거나 $1-DP_i = \Pr(P_i=0)$로 나타내며, 데이터 확률 평가는 $|D_i| / |D|$ 로 평가한다. $|\cdot|$는 데이터셋의 크기를 의미한다.

컨텐츠 서버 CS는 데이터 확률정보에 의해 구성된 서브넷의 정보 중 데이터 확률이 가장 높은 정보 $\vec{D}$에 대해서 속성 정보 d_i를 부여하고 해당 데이터의 종류, 기능, 특성에 따라서 속성 집합 $\bar{d}$를 생성한 후 식 (8)처럼 해쉬 함수 $H()$에 적용하여 데이터의 연계 정보 CI_i를 생성한다.

$$CI_i = H(\vec{D}, \bar{d}),\quad 1 \le i \le n \quad (8)$$

컨텐츠 서버 CS는 사용자 U_i에게 데이터 연계정보와 함께 데이터를 식 (9)처럼 구성하여 전달한다.

$$\sum_{i=1}^{n} H(Data,\ CI_i) \quad (9)$$

이 때, 사용자 U_i는 식 (10)의 정보를 실시간으로 모니터링하며 사용한다.

4. 성능 평가

제안 기법의 성능평가는 통신비용과 계산비용 등으로 CELF 알고리즘과 비교 평가한다.

4.1 환경설정

<Table 2>처럼 성능 평가 기준은 [12]을 근거로 하여 설정한 수치들이다. 제안 기법의 성능 평가를 위해 각각의 사용자 U_i가 서비스를 요청할 경우 서브넷의 수는 {1, 3, 5, 10}로 설정하고, 선택된 데이터 수는 {250, 500, 1000, 2000}으로 설정하고, 속성수는 {1, 2, 3, 4, 5}로 설정한다. 컨텐츠 서버 CS_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
Number of Subnet	s={1, 3, 5, 10}
Number of Selected Data	d = {250, 500, 1000, 2000}
Number of Property	p = {1, 2, 3, 4, 5}
N_{Det}	Input Parameter for Detection

4.2 성능분석

4.2.1 속성 정보 수에 따른 데이터 탐지율

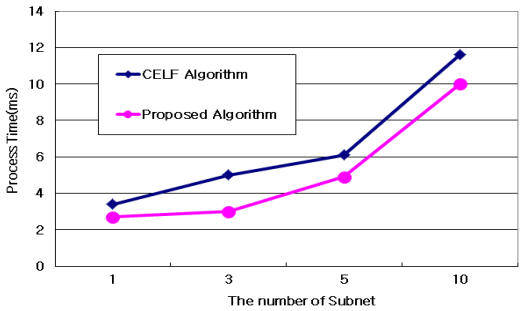
<Table 3>는 사용자가 요청한 데이터와 유사한 정보를 확률기반으로 서브넷을 구성한 데이터의 속성 정보 수에 따른 데이터 탐지율을 CELF 알고리즘과 제안기법을 비교분석한 결과를 나타내고 있다. <Table 3>처럼 서브넷 수는 1, 3, 5, 10으로 구성되도록 한 후 확률이 높은 데이터의 수를 각각 250, 500, 1,000, 2,000 으로 선택되도록 하여 사용자가 요청한 데이터와 일치되는 탐지율을 CELF 알고리즘과 비교한 결과 제안 알고리즘이 평균 6.8% 높은 탐지율을 보였다. 그러나, 서브넷 수가 3으로 설정하고 선택된 데이터 수를 500으로 설정하였을 경우에는 평균 2.9%로 CELF 알고리즘과 거의 차이를 보이지 않았다. 이 같은 결과는 서브넷 수가 3일 경우가 데이터 탐지율이 가장 높게 나타나기 때문이다.

〈Table 3〉 Detected Rate through the number of Property based Probability Info. within Subnet

Algorithm	# of subnet	# of selected data	Detected rate through the number of Property Info.
CELF Algorithm	1	250	65.2%
	3	500	61.0%
	5	1000	57.5%
	10	2000	44.8%
Proposed Algorithm	1	250	70.1%
	3	500	64.8%
	5	1000	61.7%
	10	2000	55.6%

4.2.2 서브넷 수에 따른 처리시간

[Fig. 3]은 서브넷 수에 따른 사용자가 요청한 데이터의 검색 처리 시간을 나타내고 있다.



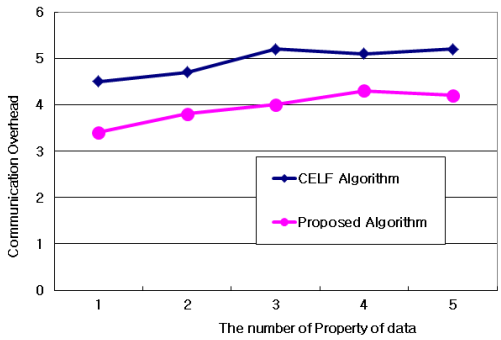
[Fig. 3] Process Time through the Number of Subnet

[Fig. 3]의 실험결과, 서브넷 수가 증가할수록 데이터 정확도를 나타내는 확률정보의 값이 높게 나타난 제안 알고리즘이 CELF 알고리즘보다 평균 8.2% 처리시간이 단축되었다. 이 같은 결과는 제안 기법에 사용되는 알고리즘이 데이터간 연계정보를 통해 데이터를 검색하여 탐지하기 때문에 나타난 결과이다. 따라서, 제안기법은 서브넷 수가 증가할수록 정확한 데이터를 찾는 처리 시간이 CELF 알고리즘보다 짧아지는 결과를 얻었다.

4.2.3 데이터 속성수에 따른 통신 오버헤드

[Fig. 4]는 데이터 속성 수에 따른 데이터의 통신 오버헤드를 나타내고 있다. [Fig. 4]의 결과, CELF 알고리즘은 속성 수가 증가할수록 통신 오버헤드가 비례적으로

증가하였지만 제안 기법의 경우 데이터의 속성수가 3개 미만일 경우에는 CELF 알고리즘과 동일하게 통신오버헤드가 비례적으로 증가하였지만, 3개 이상일 경우에는 통신 오버헤드가 일정하게 나타났다. 이 같은 결과는 제안기법에서 데이터의 속성에 따른 확률정보를 쌍으로 데이터를 탐색하도록 설정하였고, 서브넷을 통해 데이터의 확률정보가 높은 데이터간 서브넷을 연결하였기 때문에 서브넷의 수가 증가할수록 탐색 속도가 높아져 통신 오버헤드가 증가하지 않는 결과를 얻었다.



[Fig. 4] Communication Overhead thourgh the Number of Property of data

5. 결론

최근 인터넷이나 스마트폰을 통해 사용되는 데이터의 증가로 인하여 빅 데이터의 중요성이 증대되고 있다. 본 논문에서는 빅 데이터의 데이터 속성정보(종류, 기능, 특성)에 따라 데이터에 확률값을 부여하여 서브넷을 구성한 후 서브넷간 유사정도에 따라 서브넷을 서로 연계하여 데이터의 정확도와 계산비용을 줄일 수 있는 서브넷 선택 기법을 제안한다. 실험 결과, CELF 알고리즘보다 제안 기법이 평균 6.8% 높은 탐지율을 보였으며, CELF 알고리즘보다 평균 8.2% 처리시간이 단축되었다. 향후 연구로 본 연구의 결과를 기반으로 빅 데이터 시스템에 실제 적용할 계획이다.

ACKNOWLEDGMENTS

This paper has been supported by 2015 Hannam University Research Fund.

REFERENCES

[1] H. Hu, Y. Wen, T. S. Chua, X. Li, “Toward Scalable Systems for Big Data Anaqlytics: A Technology Tutorial”, IEEE Access, vol. 2, pp. 652-687, 2014.

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

[3] V. Gadepally, J. Kepner. “Big data dimensional analysis”, 2014 IEEE High Performance Extreme Computing Conference(HPEC) pp. 1-6, Sep. 2014.

[4] Y. Demchenko, C. De Laat, P. Membrey, “Defining architecture components of the Big data Ecosystem”, 2014 International conference on Collaboration Technologies and Systems(CTS), pp.104-112, May, 2014.

[5] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers, “Big Data: The Next Frontier for Innovation, Competition and Productivity”, Mckinsey Global Institute, pp. 1-137. 2011.

[6] P. Shen, Y. Zhou, K. Chen, “A Probability based Subnet Selection Method for Hot Event Detection in Sina Weibo Microblogging”, 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1410-1413, Aug. 2013.

[7] K. Chen, Y. Zhou, H. Zha, J. He, P. Shen, X. Yang, “Cost-Effective Node Monitoring for Online Hot Event Detection in Sina Weibo”, In Proceedings of the 22nd international conference on World Wide Web, ACM. pp. 107-108, April. 2013.

[8] D. Kempe, J. Klenberg, E. Tardos, “Maximizing the spread of influence through a social netowrk”, In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 137-146, Aug. 2003.

[9] K. M. P. Shrivastba, M. A. Rizvi, S. Singh, "Big Data Privacy Based on Differential Privacy a Hope for Big Data", 2014 International conference on Computational Intelligence and Communication Networks, pp. 776-781. Nov. 2014.

[10] A. Katal, M. Wazid, R. H. Goudar, "Big data: Issues, challenges, tools and Good practices ", 2013 Sixth International Conference on Contemporary Computing(IC3), pp. 404-409, Aug. 2013.

[11] Y. C. Jung. "Big Data revolution and media policy issues", KISDI Premium Report, Vol. 12, No. 2, pp. 1-22, 2012.

[12] S. H. Kim, N. U. Kim, t. M. Chung, "Attribute Relationship Evaluation Methodology for Big Data Seucrity", 2013 International Conference on IT Convergence and Security(ICITCS), pp. 1-4, Dec. 2013.

[13] S. Y. Son, "Big data, online marketing and privacy protection", KISDI Premium Report, Vol. 13, No. 1, pp.1-26, 2013.

[14] J. T. Kim, B. J. Oh, J. Y. Park, "Standard Trends for the BigData Technologies", 2013 Electronics and Telecommunications Trends, Vol. 28, No. 1, pp. 92-99, 2013.

[15] M. Paryasto, A. Alamsyah, B. Rahardjo, Kuspriyanto, "Big-data security management issues", 2014 2nd International Conference on Information and Communication Technology(ICoICT), pp. 59-63, May, 2014.

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@mokwon.ac.kr

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