JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 18, No. 3 : 595~601 Mar. 2014

역순트리를 이용한 특이데이터 국소적 접근

임광철[1] · 설정자[2*]

Unusual data local access using inverse order tree

Kwang-cheol Rim[1] · Jung-ja Seol[2*]

[1]Department of Computer Engineering, Chosun University, Gwangju 230-7381, Korea
[2*]Department of Mathmatics, Chosun University, Gwangju 230-6610, Korea

요 약

스마트 정보통신시대에 데이터의 수는 기하급수적으로 증가하고 있다. 이에 데이터 발생지역과 발생상황을 실시간으로 파악하고 분석하는 것이 신속한 조치를 취하는 중요한 요소로 떠오르고 있다. 본논문에서는 분석자가 원하는 특성 데이터 발생지역의 국소적 판단을 하기 위하여 데이터 발생에 대한 값을 최하위 모듈에서부터 최상위 모듈까지 이어지는 루트를 역순으로 진행하면 데이터 발생과 동시에 분석이 가능하다. 먼저 군집분석에 대해 알아보고 군집원들의 합에 의한 분석법을 트리 구조에 병합하여 최하위 모듈부터 최상위 모듈까지 발생 특성값에 대해 수치로 치환하고 그 합을 도출하도록 설계하였다. 또한 특성값에 대한 가중치를 부여하여 원하는 값의 발생상황을 실시간으로 도출되도록 설계하였다.

ABSTRACT

With the advent of the Smart information-communication era, the number of data has increased exponentially. Accordingly, figuring out and analyzing in which area and circumstance the data has been created becomes one of the factors for prompt actions. In this paper identifies how to analyze the data by implementing a route from the lowest module to highest one in an inverse order for the part judgement for the particular data. The script first identifies cluster analisys, paralizes the analysis using the sum of each factors of the cluster with the tree structure, and finally transpose the answer into number. Also, it is designed to place priority on particular answer, thereafter, draws the wanted answer real-time.

키워드 : 역순트리, 가중트리, 10진트리, 빅데이터, 실시간분석

Key word : inverse order, weight tree, digit tree, Big data, real time analysis

접수일자 : 2014. 01. 20 심사완료일자 : 2014. 02. 11 게재확정일자 : 2014. 02. 26

* **Corresponding Author** Jung-ja Seol (E-mail:jzseol@korea.com, Tel:+82-62-230-7381)
Department of Mathmatics, Chosun University, Gwangju 230-6610, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2014.18.3.595 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

스마트 정보통신 시대로 접어들면서 범람하는 정보의 홍수와 이를 가공, 변환하여 새로운 정보로 생성하고 그로인한 여러 가지 결론을 도출 시키는 일이 계속하여 일어나고 있다. Big data라는 신조어가 만들어지고 Big data를 이해하고 이에 속한 data 의 흐름을 분석, 통계, 취합하여 보다나은 의사결정의 수단으로 사용되고 있다.

Big data의 생성은 새로운 경제성장을 위한 중요한 가치창출효과를 가져온다고 볼 수 있고 작은 정보들일지라도 모이면 거대한 흐름과 새로운 문화, 경제, 사회 흐름을 형성한다. 또한 소셜 네트워크 데이터로 글 퍼짐 관계와 특정주제에 대해 관심있는 그룹, 그리고 그룹간의 관계와 그룹내에 영향력 있는 사람 관련 분야에 대한 전반적인 내용을 분석 할 수 있다[1].

방대한 정보를 추출하여 다른 정보와 연계성을 파악하는 텍스트 마이닝 기술과 데이터의 긍정, 부정, 중립 등 선호도를 구별하는 오피니언 마이닝 등 페이스북이나 트위터에서 사용될 수 있는 관측개체를 몇 개의 그룹으로 나눔으로써 대상 집단에 대한 효율적인 활용과 아울러 앞으로 일어날 일을 예측할 수 있게 되었다.

Big data 의 발달은 순기능만 제공해주는 것은 아니다. 필요이상으로 많은 데이터들이 범람함으로 인하여 개인정보 유출과 이로 인한 보안 취약성이 급증할 수 있다. 최근 사이버공격이 표적공격으로 심화되어 사이버테러, 사이버전(war), 핵티비즘 등으로 집중공격이 매번 새로운 공격방법으로 접근하고 있다. 때문에 발생 data 의 관리는 중요한 부분을 차지하고 있다. Big data 관리에서 중요한 요소로 저장된 데이터로부터 필요한 정보를 신속히 획득 할 수 있어야 한다는 점이다[2].

이에 본고에서는 Big data로 인하여 발생하는 data들을 취합하지 않고 실시간으로 탐색하는 구조를 설계하고자 한다. Data 탐색의 기본 방식은 모든 data를 모아서 첫째항부터 끝항까지 검색하고 일치하는 data를 골라내는 방식이 기준으로 주어져 있다. 역순10진 트리 구조는 방향을 반대로 돌려서 최상위 모듈의 구성이 하위 모듈로부터 올라오게 설계를 하였다. 검색시간은 기하급수적으로 줄어들고 하위모듈에서 올라오는 data의 합으로 인하여 상위모듈의 성분을 결정하게 되므로 상위모듈에서 하위 모듈로 실시간 탐지가 가능하게 되었다.

먼저 그룹을 분류하고 분석할 수 있는 그룹분석이론에 대하여 살펴보고 역순 10진 트리구조와 각각의 성분에 가중치를 부여한 가중트리구조 그리고 그룹에 가중치를 부여한 모듈가중진트리에 대해 정의하였다. 마지막으로 역순트리구조를 활용하는 방안에 대해 간략히 언급하였다.

Ⅱ. 군집분석

관측대상에 대한 각각의 특징들이 표현되는 데이터들의 분류기준을 서로의 유사성에 의해 분류하고 분석하는 것을 군집분석(cluster analisys)라 한다.

군집분석의 종류는 크게 두 가지로 볼 수 있다.

관측대상간의 거리(유사성, similarity)를 기초로 해서 가까운 것들 끼리 군집으로 묶는 방법이 있고 데이터간의 분산을 이용하여 그룹간 분리 정도를 결정하는 방법이 있다.

본 고에서는 거리개념을 도입한 분석을 살펴보고 이를 실시간 데이터 검색으로 확장해서 알아본다.

군집분석을 실시하기 위하여 관측대상이 n개의 속성을 갖고 m개의 관측대상을 갖는다고 하면 각각의 m개의 관측대상마다 주관적인 척도를 줄 수 있고 이러한 척도에 대해 관측값이 변이되는 데이터들에 대한 측정값을 산정한다.

m개의 관측대상들중 임의의 두 관측대상을 x_r, x_s라 하면

$$x_r = (x_{r1}, x_{r2}, x_{r3}, \cdots, x_{rn})$$
$$x_s = (x_{s1}, x_{s2}, x_{s3}, \cdots, x_{sn})$$

로 나타낼 수 있고 이는 각 관측대상들을 n차원 벡터공간의 원소로 볼 수 있고 이에 대한 유사성의 척도를 각각의 대응하는 변량의 차가 작으면 두 대상을 유사하다고 판단하는 유클리드 거리와 각각 변량들마다 가중치를 적용하는 가중 유클리드거리로 표현할 수 있다.

$$d_{rs} = \sum_{k=1}^{n} (x_{rk} - x_{sk})^2, \text{ 유클리드 거리}$$

$wd_{rs} = \sum_{k=1}^{n} w_k (x_{rk} - x_{sk})^2$, $w_k (k = 1, \cdots, n)$는 가중치

주어진 관측대상을 벡터공간의 원소로 보고 두 벡터의 내적으로 유사성을 정의 할 수 도 있다.

$$x_r \cdot x_s = \sum_{k=1}^{n} x_{rk} x_{sk} \tag{1}$$

이고 이를 표준화하여 Pearson 적률상관계수를 구하면

$$x_r \cdot x_s = \sum_{k=1}^{n} \frac{(x_{rk} - \overline{x_r})(x_{sk} - \overline{x_s})}{\sqrt{\sum_{h=1}^{n} (x_{rh} - \overline{x_r})^2} \sqrt{\sum_{h=1}^{n} (x_{sh} - \overline{x_s})^2}},$$

$$\overline{x_r} = \frac{1}{n} \sum_{k=1}^{n} x_{rk} \tag{2}$$

만약 관측대상의 모든 성분들이 2진 data인 경우를 categorical data라 하며 동일 위치의 벡터성분의 연관성을 4가지의 형태의 자료로 분류 한다.

$$a = \sum_{k=1}^{n} x_{rk} x_{sk}, \qquad (1,1)의\ 개수$$

$$b = \sum_{k=1}^{n} x_{rk} (1 - x_{sk}) \qquad (1,0)의\ 개수$$

$$c = \sum_{k=1}^{n} (1 - x_{rk}) x_{sk} \qquad (0,1)의\ 개수$$

$$d = \sum_{k=1}^{n} (1 - x_{rk})(1 - x_{sk}) \qquad (0,0)의\ 개수$$

이므로 $a + b + c + d = n$임을 알 수 있다.

본고에서는 이미 생성된 data가 아닌 생성될 data에 대한 상태파악을 기준으로 분석을 진행한다. 가중치를 적용하기 전의 각각의 벡터성분의 합을

$$S = \sum_{k=1}^{n} x_{rk} \tag{3}$$

라 하고 이를 상위 모듈에 전송하여 상위 모듈을 생성하는 방식으로 data를 완성한다.

Ⅲ. 역순트리

수형도는 나무 모양으로 데이터의 흐름을 표현하는 방법으로 주로 2진트리 구조를 사용한다. 본고에서는 가지가 10개인 10진트리구조를 제안하고 이에 대한 정의와 구조를 설계한다.

정의 - n항 수형도

모든 모듈의 차수가 n 이하인 루트 수형도를 n항 수형도라 하고 n=10 인 경우를 10진 수형도(digitary tree)라 한다.

정의 - 상태전이(state transmition)

n번째 모듈의 상태 즉 n번째 모듈의 총 성분들의 합을 상태라 하며 이 상태를 상위 모듈에 전해주는 것을 상태전이라 한다.

정의 - 상위 탐색

n-th 수형도의 최상위 모듈에서부터 반응 모듈로 내려가는 탐색을 상위 탐색이라 한다.

정의 - 상위 탐색시간

상위탐색으로 반응 모듈까지 상태를 검색하면서 거치는 모듈의 개수를 상위탐색시간이라 한다.

정의 - 차수

각각의 모듈에서 나오는 가지의 수를 차수라 하고 가지가 10개인 경우 10차라 한다.

정의 - 깊이

최상위 모듈에서 최하위 모듈까지 이르는 최단거리를 트리의 깊이 라 한다.

3.1. 역순10진트리

그림 1에서 보는바와 같이 각각의 모듈별로 10개의 가지를 가지고 있는 구조를 10진트리 구조라 하고 최상위 모듈에서 최하위 모듈까지의 데이터에 접근하는데 걸리는 시간은 어느 곳으로 접근하더라도 탐색시간이 3으로 나타난다. 이는 최상위 모듈에서 최하위 모듈까지의 거리를 계산하는데 깊이가 깊을수록 기하급수적으로 단축됨을 볼 수 있다. 깊이가 10인 트리구조를 살펴

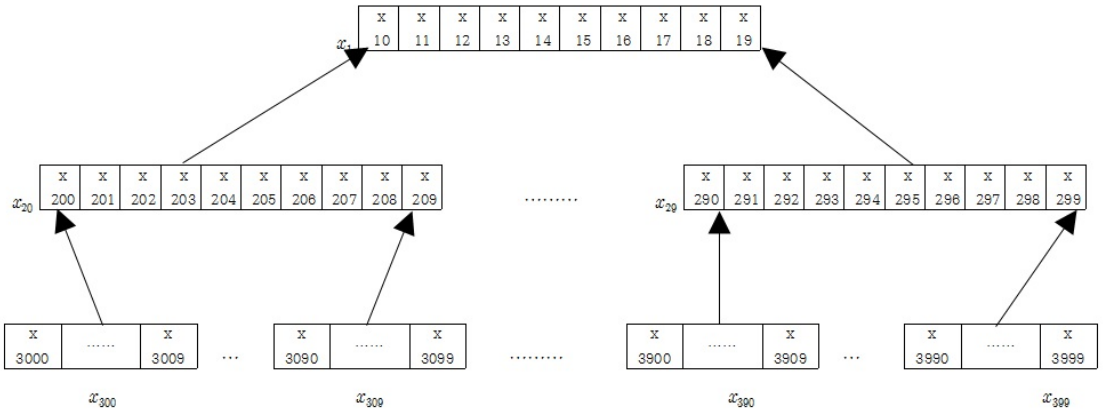

그림 1. 깊이3인 10진트리
Fig. 1 digit tree of deep 3

보면 최하위 모듈의 수는 10^{10}개인데 반해 최상위 모듈에서 최하위 모듈까지 탐색해 들어가는데 걸리는 탐색시간은 $\log 10^{10} = 10$으로 지수에 종속됨을 볼 수 있다.

각각 모듈에서의 성분의 움직임은 아래 모듈의 성분의 합으로 표현한다. 그림1에서 최하위 모듈인 x30의 성분의 합은 $\sum_{k=0}^{9} x30k$이고 이는 x20의 첫 번째 성분으로 상태전이 된다. 또한 x20의 두 번째부터 10번재 성분은 각각 x31의 성분의 합과 x39의 성분의 합으로 상태전이 된다.

x20의 성분의 합은 또다시 x1의 첫 번째 성분으로 상태전이 된다. 즉 최하위 모듈인 x30에서 최상위 모듈인 x1까지 거리가 3으로, 상위탐색에 의해 최하위 모듈까지 가는데 거리 3이 됨을 볼 수 있다.

깊이가 3인 10진트리의 성분은 다음식을 따른다.

$$x_1 = (x_{10},\ x_{11},\ x_{12},\ x_{13},\ x_{14},\ x_{15},\ x_{16},\ x_{17},\ x_{18},\ x_{19})$$

$$\begin{aligned} x_{10} &= \sum_{k=0}^{9} x_{20k} \\ &= x_{200} + x_{201} + \cdots + x_{208} + x_{209} \\ &= \sum_{j=0}^{9} x_{300j} + \sum_{j=0}^{9} x_{301j} + \cdots + \sum_{j=0}^{9} x_{309j} \\ &= \sum_{k=0}^{9}\sum_{j=0}^{9} x_{30kj} \end{aligned}$$

$$\begin{aligned} x_{19} &= \sum_{k=0}^{9} x_{29k} \\ &= x_{290} + x_{291} + \cdots + x_{298} + x_{299} \\ &= \sum_{j=0}^{9} x_{390j} + \sum_{j=0}^{9} x_{391j} + \cdots + \sum_{j=0}^{9} x_{399j} \\ &= \sum_{k=0}^{9}\sum_{j=0}^{9} x_{39kj} \end{aligned} \quad (4)$$

가 됨을 볼 수 있다. 이를 깊이가 10인 10진트리로 확장해 보면 최상위 모듈의 성분들은 아래 모듈들의 모든 성분의 합으로 표현할 수 있다.

$$x_{1n} = \sum_{k=0}^{9}\sum_{j=0}^{9}\sum_{a=0}^{9}\sum_{b=0}^{9}\sum_{c=0}^{9}\sum_{d=0}^{9}\sum_{e=0}^{9}\sum_{f=0}^{9}\sum_{g=0}^{9} x_{100nkjabcdefg},$$

$$n = 0 \sim 9 \quad (5)$$

3.2. 가중치를 적용한 10진트리

최하위 모듈의 성분에 각각 가중치를 부여하거나 각각의 모듈에 가중치를 부여하는 트리구조를 살펴보자. 역순10진트리 구조에서 최하위 모듈 각각의 성분에 특정한 가중치를 부여한 구조를 가중10진트리라 하고 성분이 아닌 모듈에 가중치를 부여한 구조를 모듈가중10진트리라 하자. 그림 2에서와 같이 최하위 모듈에 속하는 10개의 성분에 각각 가중치 $m_i (i = 0 \sim 9)$를 부여하

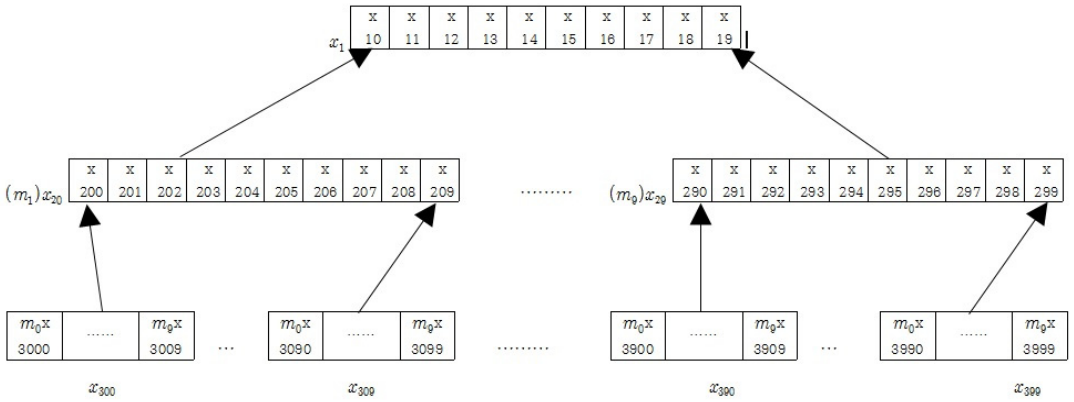

그림 2. 깊이가 3인 가중트리
Fig. 2 weight of deep 3

여 각각의 성분에 크기를 달리 하는 것을 가중10진 트리라 하고 그림 2에서와 같이 차상위 모듈에 군집분석에 의해 분류하고 각각의 그룹의 중요도에 따른 가중치 $m_i(i=0\sim 9)$를 각 모듈의 성분합에 곱하여 상태전이 시키는 방법을 모듈가중10진트리라 한다.

3.2.1. 가중10진트리

최하위 성분의 원소를 결정지을 때 각각 성분의 중요도에 따라 성분출현에 대한 가중치를 부여하는 방법으로 그림 2의 x_{30}의 성분각각에 다음과 같이 가중치를 부여한다.

$$x_{30} = (m_0x_{3000},\ m_1x_{3001},\ m_2x_{3002},\ \cdots,\ m_9x_{3009})$$

이의 합 $\sum_{j=0}^{9} m_j x_{300j}$는 x_{20}의 첫 번째 성분으로 상태전이 된다. 마찬가지로 x_{20}의 두 번째부터 10번째 성분은 각각 x_{31}와 x_{39}의 성분합으로 상태전이 된다. x_{20}의 성분의 합은 x_1의 첫 번째 성분으로 상태전이 된다. 깊이가 3인 역순가중10진트리의 성분은 다음과 같이 결정된다.

$$\begin{aligned} x_{10} &= \sum_{k=0}^{9} x_{20k} \\ &= x_{200}+x_{201}+\ \cdots\ +x_{208}+x_{209} \\ &= \sum_{j=0}^{9} m_j x_{300j}+\sum_{j=0}^{9} m_j x_{301j}+\ \cdots\ +\sum_{j=0}^{9} m_j x_{309j} \\ &= \sum_{k=0}^{9}\sum_{j=0}^{9} m_j x_{30kj} \\ &\ \ \cdot \\ x_{19} &= \sum_{k=0}^{9} x_{29k} \\ &= x_{290}+x_{291}+\ \cdots\ +x_{298}+x_{299} \\ &= \sum_{j=0}^{9} m_j x_{390j}+\sum_{j=0}^{9} m_j x_{391j}+\ \cdots\ \sum_{j=0}^{9} m_j x_{399j} \\ &= \sum_{k=0}^{9}\sum_{j=0}^{9} m_j x_{39kj} \end{aligned} \tag{6}$$

가중10진트리는 찾고자 하는 성분에 특성을 부여하여 최하위 성분에 출현빈도가 적더라도 강한특성값을 갖는다면 상위 모듈에 그 특성이 반영되게 설계하였다.

3.2.2. 모듈가중 10진트리

동일한 성분의 구성이라도 각각의 모듈을 형성하는데 서로 다른 성질의 군집을 이룰 수 있다. 군집을 구성하는 방법은 일반적인 유클리드 군집분석으로 접근하기 힘든 상황들이 연출된다.

이에 사용자의 전문적인 지식과 생각이 부여된 전문가 시스템에 의한 군집분석이 이루어 져야한다. 이후 각각의 군집들의 가중치를 달리 표현해서 분석하는 방

법을 모듈가중 10진트리 라 한다.

본고에서는 단순 유클리드 거리를 이용하여 모듈가중10진트리의 구조를 살펴보기로 하자. 깊이가 3인 10진트리구조에서 두 번째 단계인 x_{20}은 x_{300}, … ,x_{309}까지의 성분의 합이 x_{20}각각의 성분으로 상태전이 되었고 x_{21}의 각각의 성분은 x_{310}, … ,x_{319} 성분의 합이 상태전이 된 것이다. 이는 x_{30}부터 x_{39}까지 거리 1인 유클리드 군집분류에 의해 x_{20}으로 상태전이 됨을 볼 수 있다. 이후 x_{20}에는 가중치 m_0를 x_{21}에는 가중치 m_1을 부여하면 첫 번째 모듈인 x_1에서 가중치의 영향을 받은 성분이 각각 생성된다.

$$
\begin{aligned}
x_{10} &= m_0 \sum_{k=0}^{9} x_{20k} \\
&= m_0 \sum_{k=0}^{9} \sum_{j=0}^{9} x_{30kj} \\
&\cdot \\
x_{19} &= m_9 \sum_{k=0}^{9} x_{29k} \\
&= m_9 \sum_{k=0}^{9} \sum_{j=0}^{9} x_{39kj}
\end{aligned} \tag{7}
$$

Ⅳ. 결 론

현대사회는 과거에 비해 어마어마한 양의 data가 발생하고 이러한 data 들을 분석, 가공 후 필요한 정보를 생성하는 대는 시간과 공간의 제약이 따른다. 최근 big data에 대한 여러 연구 성과와 실사용 예들이 속속 나오고 있는 실정이지만 대부분이 이미 발생된 data를 취합하여 제 가공하는 방식으로 진행된다. 모든 data는 취합하는 순간 이미 지나버린 정보로 취급된다. 다시 말해 data를 가공하기 위하여 하는 작업 자체는 실시간 대응은 이룰 수 없다고 볼 수 있다.

본 논문에서 살펴본바는 먼저 군집분석에 대하여 정의를 하고 군집분석상 거리 1인 유클리드 거리를 이용하여 하위 모듈을 분류하고 진행하였다. 역순10진트리를 설계하고 수학적인 흐름을 통해 최상위 모듈까지의 발생 빈도합을 도출하였고 가중10진트리구조는 각각의 발생data에 가중치를 부여하는 방식과 발생모듈에 가중치를 부여하는 방식으로 발생 빈도합을 수학적으로 계산하였다. 이는 기존 알려진 트리구조의 설계에 비하여 지역적접근 면에서 실시간으로 발생을 제어할 수 있다는 것을 살펴보았다. 역순10진 트리와 역순가중10진트리는 발생하는 data를 실시간으로 빈도를 확인하고 어떠한 장소, 어떠한 그룹에서 일어나는 이벤트인지를 바로 확인할 수 있는 장점이 있다.

예를 들어 감기와 관련된 단어를 최하위 성분으로 구성한다면

$$
x_{3000} \rightarrow \text{감기}, x_{3001} \rightarrow \text{인플루엔자}, x_{3002} \rightarrow \text{열}, x_{3003} \rightarrow \text{기침}, x_{3004} \rightarrow \text{컨디션}, \cdots
$$

와 같이 감기와 관련어를 최하위 성분으로 배열하고 발생하는 매순간 각 성분값이 1 로 활성화 된다. 이후 “감기”, “열” 이라는 단어가 최하위 성분에 도출 되면 x_{3000}, 과 x_{3002}가 1이 되므로 x_{200}=2 로 상태전이 된다. 이는 x_{10}=2 로 상태전이 되어 최하위 모듈의 결과값이 최상위 모듈로 취합됨을 볼 수 있다. 사용자는 최상위 모듈의 x_{10}의 상태를 보고 $x_{10}\rightarrow x_{20}\rightarrow x_{300}$으로 탐색해 들어갈 수 있다. 최상위 모듈에서 성분값이 정해진 값 이상이 도출되면 바로 하위탐색 하여 발생한 지역과 그룹을 선별하고 이벤트를 해결 할 수 있다.

역순10진 트리 탐색은 깊이가 10인 경우 10^9개의 검색대상을 선정할 수 있고 선정된 대상들에 대한 실시간 탐색시간은 최하위 모듈까지 $\log 10^9$로 도출되므로 많은 대상에 대한 실시간 정보검색과 대응이 가능하다.

여러 연구자들의 관심과 좀더 진전된 모습으로 발전된다면 Big data 응용분야 뿐만 아니라 의료, 기상, 보안, 바이러스 확산, 등등 그 응용범위는 광범위 하다 할 수 있다.

REFERENCES

[1] S. J. Shin, “SNS using Big Data utilization research,” *Journal of the Institute of Internet, Broadcasting and Communication*, vol. 12, no. 6, pp. 268-272, Dec. 2012.

[2] J. H. Kim, S. H. Lim, I. K. Kim, H. S. Cho, B. k. No “Technical trends of cyber security with Big data”, *Journal of the Electronics and telecommunications trends*, vol. 28, no. 3, pp.19-29, Jun. 2013.

임광철(Kwang-cheol Rim)

2000년 조선대학교 대학원 이학석사
2006년 조선대학교 대학원 이학박사
현재 조선대학교 수학과 외래교수
※관심분야 : 응용수학, 정보보안, 양자암호, 암호학

설정자(Jung-Za Seol)

2000년 조선대학교 교육대학원 교육학 석사
2007년 조선대학교 컴퓨터공학 박사수료
현재 조선대학교 컴퓨터공학부 외래교수
※관심분야 : 통신 네트워크 및 정보보안, 유비쿼터스