논문 2014-2-1

빅 데이터 소프트웨어 비용산정을 위한 FP 응용

권기태*

An Application of FP for Big Data Software Cost Estimation

Ki-Tae Kwon*

요 약

데이터가 기하급수적으로 증가하는 대용량의 데이터의 시대가 소셜 서비스의 이용과 개인당 스마트 기기 보유량이 늘어남에 따라 도래했다. 빅 데이터에 관심을 가지는 이유는 과거에는 규모가 너무 커서 분석이 불가능하였던 데이터의 분석이 이제는 가능해졌고 이것을 통해 새로운 가치(Value)를 찾아내고 있기 때문이다. 빅 데이터 프로젝트 관리에 있어서 초기 개발 단계에서 정확한 소프트웨어 개발 비용 및 규모 산정의 중요성에 불구하고 빅 데이터 환경을 반영하는 비용산정 연구는 거의 이루어지지 않았다. 본 연구는 이러한 한계를 극복하기 위한 시도로서 퍼지 논리를 적용한 FP 기법을 제안하였다. 본 연구에서 제안된 기법은 FP 구성요소의 실제 규모를 반영하여 적절한 규모산정이 가능하다. 복잡도 경곗값에서 증가(감소)에 따르는 수준만큼 FP값이 증가(감소)하므로, 개발 초기 단계에서 FP 산정이 모호하더라도 파급 효과를 최소화할 수 있다. 본 연구에서 제안하는 퍼지 FP 기법은 빅 데이터 환경에서 대규모 소프트웨어에 적절한 비용산정이 가능하다.

Abstract

Era of large volumes of data that data increases exponentially has arrived. This is because the holdings of smart devices per use and one of the social services has increased. Regardless of the development costs and the importance of scale calculation of precise software in the early stages of development in big data project management, cost estimation study reflects the big data environment was not nearly done. In this study, for an attempt to overcome these limitations, we have proposed a FP method of applying fuzzy logic. The proposed method in this paper is to reflect the actual size of the components of the FP, it is possible to correct cost estimation. Only the level FP value associated with an increase(decrease) in the boundary value increases(decreases), it is possible to minimize the ambiguity even ripple effect FP during the early stages of development. Fuzzy FP method proposed in this study, it is possible to calculate the appropriate cost for large-scale software in big data environment.

한글키워드 : 빅 데이터, 비용산정, 퍼지 FP 기법

Ⅰ. 서 론

최근데이터가 기하급수적으로 증가하는 대용량의 데이터의 시대가 소셜 서비스의 이용과 개인당 스마트 기기 보유량이 늘어남에 따라 도래했다. 빅 데이터의 핵심은 단순한 스토리지 서비스나 데이터 분석만을 의미하는 것이 아니라 대

* 강릉원주대학교 컴퓨터공학과 교수
(email : ktkwon@gwnu.ac.kr)
접수일자: 2014.12.14 수정완료: 2014.12.23

량의 다양한 데이터를 빠르게 검색하고 분석해 경제적인 가치를 이끌어 나는데 있다.

무한한 빅 데이터 활용 가능성은 기존 금융, 통신 사업을 비롯해 의료, 농업, 국방, 교통 정보 등 전 세계 다양한 분야에서 증가하고 있다. 또한 전염병과 자연재해의 이동경로를 파악하는 등 데이터 자원의 무궁무진하게 분석, 예측이 빅 데이터를 이용한 분석과 해석으로 가능하게 되었다. 또한 기업 환경에서는 의사결정에 필요한 의미 있는 정보의 발견, 분석 능력이 방대한 정보 속에서 기업 비즈니스의 핵심 경쟁력으로 부상하고 있고 분석 도구와 비즈니스 인텔리전스 플랫폼 서비스에 대한 수요 증가가 이어지고 있다. 위키피디아의 빅 데이터는 다음과 같이 정의되고 있다[3]. “빅데이터란 기존 데이터베이스 관리도구의 데이터 수집, 저장, 관리, 분석의 역량을 넘어서는 대량의 정형 또는 비정형 데이터 세트 및 데이터로부터 가치를 추출하고 결과를 분석하는 기술을 의미한다.” 즉 기존에 다루던 수준을 뛰어 넘는 대규모의 자료를 뜻하며, 이와 같은 빅 데이터의 분석을 위한 IT의 발전은 최근 들어 급격하게 진행되어 빅 데이터 분석이 현실화되어 가고 있다[1].

그림 1과 같이 물리적 하드웨어로부터 시작해 인프라 소프트웨어에서 서비스 소프트웨어 부문으로 빅 데이터 시장이 확장되고 있다. 특히 빅 데이터 의미 파악 및 이해 능력, 분석을 위한 총체적이고 직관적인 시각화 연구가 확대되고 있다. 이를 위해 다양한 사용자와 IT 기기를 통해 생성되고 수집된 대규모 정형 데이터와 비정형 데이터를 전략적으로 활용하기 위한 통찰력을 발견하고 예측력을 제공하는 업무수행 과정 및 지원 플랫폼을 필요로 하고 있고 나아가 분석기법에 대한 부분들도 새롭게 해석이 되어져야 하는 과제를 가지고 있다.

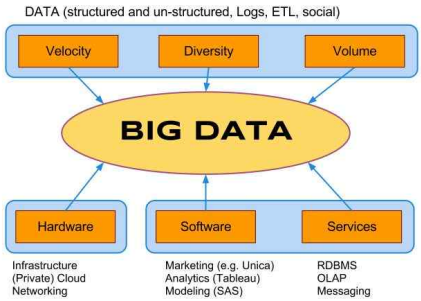

그림 1. 빅 데이터의 배경

오늘날 우리가 빅 데이터에 관심을 가지는 이유는 과거에는 규모가 너무 커서 분석이 불가능하였던 데이터의 분석이 이제는 가능해졌고 이것을 통해 새로운 가치(Value)를 찾아내고 있기 때문이다. 데이터는 과거에 비해서 더 빨리 늘어나고(Volume), 더 다양한 형태(Variety)를 가지고, 실 시간에 가까운 속도(Velocity)로 생성되고 있고, 이러한 데이터를 이용한 빅 데이터 분석 기술들이 현실화 되고 있다[2].

본 논문은 서론에 이어, 2장에서 빅 데이터 소프트웨어 비용산정의 필요성, 3장에서 FP의 문제점, 4장에서 조정된 FP를 제시하고, 5장에서는 연구의 결론과 향후 연구 방향으로 구성되었다.

2. 빅 데이터 소프트웨어 비용산정

2.1 빅 데이터의 개념

1990년대 초반부터 빅 데이터 연구의 기반이 형성되었으나, 최근 들어 하드웨어와 소프트웨어, 네트워크를 비롯한 정보기술의 현저한 발전으로 인해 빅 데이터를 제대로 제어할 수 있는 능력을 가지게 되었다. Miller는 빅 데이터를 “수년 동안 존재해 왔으나 지금에 와서야 더 빨리, 더 큰 규모로 적용되어 더 많은 이용자가 접근 가능한 분석적 기술”이라고 했다[4]. Loukides는 빅 데이터

분석이 가장 효과적일 때는 데이터의 사이즈 자체가 연구 문제의 일부일 경우일 것이라고 주장했다[5]. 이와 같이 빅 데이터의 개념은 데이터의 크기를 이야기하는 것이 가장 기본적인 접근법이다. 그러나 단지 크기의 관점으로만 접근하는 것은 아니다. 방대한 양의 데이터를 처리하는 실제적인 분석이 가능해졌기 때문에 최근 몇 년간 재조명되고 활발한 논의와 함께 집중적인 관심을 받게 되었다. 이런 의미에서 빅 데이터는 그냥 커다란 데이터라는 단순한 개념이 아니라 다학제적인 연구가 융합된 통합적인 이해가 필요하다. 어쨌든 빅 데이터의 가장 단순한 개념은 직접적이며 단순한 형태로 다루고 처리하기 어려운 방대한 양의 데이터를 의미한다. 데이터의 규모가 너무 방대해 기존의 일반적인 방법이나 도구로 수집, 저장, 검색, 분석, 시각화 등을 하기 어려운 정형 또는 비정형 데이터 집합을 의미하는 것이다. 현재 빅 데이터의 개념은 외부 데이터, 비정형, 실시간 데이터 및 서로 상이한 정보의 결합으로 인한 새로운 지식 창출을 포함하는 영역으로 그 의미가 확대되었다.

2011년 McKinsey 보고서에서는 빅 데이터를 “보통의 데이터 베이스 소프트웨어 도구로 수집, 저장, 관리, 분석하기엔 그 능력의 한계를 넘어서는 방대한 사이즈의 데이터 집합”이라고 했는데[6], 이러한 정의는 “비용 효과적이며 혁신적인 형태의 정보처리를 요구하는 방대한 양, 속도와 다양한 정보로 통찰력을 증진시켜주고 의사결정을 증진시킨다”[7]는 가트너 연구소의 정의와 더불어 연구자들에게 가장 많이 사용되고 있다.

빅 데이터는 그냥 방대한 양인 것인 것만이 아니라 다양한 데이터 타입과 스트리밍 데이터들이 존재한다. 빅데이터 분석의 가장 기본적인 목적은 기술적, 사회적, 경제적 환경에서 존재하는 방대한 양의 데이터에서 의미를 파악하고 그 안에 숨겨진 패턴을 찾아내는 것이다[8].

2.2 빅 데이터 비용산정의 필요성

현재까지 정의되고 있는 빅 데이터의 특징은 그림 2과 같이 Volume(대용량), Variety(다양성), Velocity(실시간성)의 3V로 정의된다. 여기서 가트너 연구소는 한 가지 특성 Complexity(복잡성)를 덧붙여 설명하고 있다. 대부분의 연구자들이 빅 데이터의 특징을 앞서 말한 3V 또는 3V+1C로 설명하고 있으며 대용량, 다양성, 실시간성이 복합적으로 이루어지는 복잡성을 가졌다는 측면에서 4가지 측면 관점으로 설명하고 있다. 즉 규모, 다양성, 복잡성, 속도의 증가 특성을 중심으로 각각의 범주에서 원하는 가치를 얻을 수 있는 정도의 상대적 해석이 이루어지고 있다.

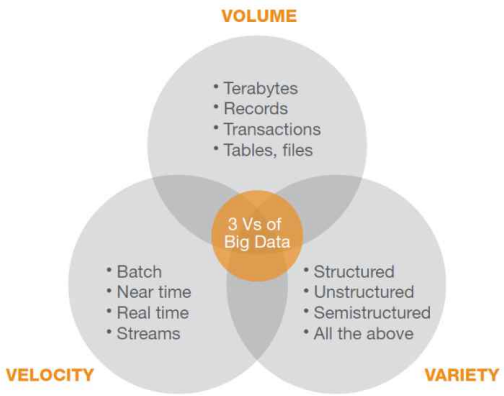

그림 2. 빅 데이터와 3V

빅 데이터의 가장 큰 특징은 취급하는 텍스트와 이미지가 정형적이지 않다는 것이다. 또한 데이터 자체의 양이 방대할 뿐 아니라 빠르게 전파되기 때문에 전통적인 방식으로 비용산정을 하기가 쉽지 않다. 다른 한편으로는 유용한 정보의 증가만큼이나 불필요한 정보도 급증하고 있어 방대한 데이터 안에서 의미가 있는 것을 찾아내고 분석하는 것이 아주 중요하며, 이를 반영하는 비용산정이 이루어져야 한다. 빅 데이터 비용산정에도 알고리즘 비용산정 기법을 적영할 수 있으

므로, 기본적으로 규모산정값에 대해 패러매트릭 모델을 적용하게 된다. 그러나 기존의 대표적인 FP의 경우, 빅 데이터 비용산정 시 본질적인 문제점을 내포하고 있으므로 빅 데이터 규모산정을 위한 FP의 조정이 필요하다.

3. 전통적인 FP의 문제점

3.1 경계 설정의 비합리성

표 1의 ILF의 가중치 값을 살펴 보면, RET와 DET의 수가 각각 (1, 1) (1, 50) (5, 19)인 경우 모두 7FP로 동일한 규모로 정의된다. 그러나 실제로 DET의 수가 50배 이상 혹은 RET의 수거 5배 이상되는 규모를 동일한 규모로 정의하는 모순이 발생한다. 마찬가지로 (2, 51) (2, 100) (8, 200)인 경우에도 15FP로 모두 동일한 규모로 산정되지만, 실제 규모를 고려하여 볼 때 이들이 모두 동일한 규모라는 것은 매우 비합리적이다.

표 1. ILF의 복잡도에 따른 가중치

RET의 개수	DET의 개수		
	1~19	20~50	51이상
1	7	7	10
2~5	7	10	15
6이상	10	15	15

또한 표 1의 복잡도 분류 기준의 경계선 상에서 DET의 수와 RET의 수가 1 증가하는 경우에 3FP 혹은 5FP 증가하는 것을 알 수 있다. 1FP에 해당하는 C언어의 라인 수가 128 정도라는 것을 고려할 때 단 한 개의 DET나 RET 수 증가로 인해 384라인 혹은 640라인 정도의 규모가 증가하는 것으로 산정되는 FP의 전통적인 계산 규칙은 매우 비합리적이라는 것을 알 수 있다.

표 2. EIF의 복잡도에 따른 가중치

RET의 개수	DET의 개수		
	1~19	20~50	51이상
1	5	5	7
2~5	5	7	10
6이상	7	10	10

이러한 문제점은 표 2의 EIF와 표 3부터 표 5의 EI, EO, EQ의 경우에도 동일하게 존재한다.

표 3. EI의 복잡도에 따른 가중치

FTRT의 개수	DET의 개수		
	1~4	5~15	16이상
0~1	3	3	4
2	3	4	6
3이상	4	6	6

표 4. EO의 복잡도에 따른 가중치

FTRT의 개수	DET의 개수		
	1~5	6~19	20이상
0~1	4	4	5
2~3	4	5	7
4이상	5	7	7

표 5. EQ의 복잡도에 따른 가중치

FTRT의 개수	DET의 개수		
	1~5	6~19	20이상
0~1	3	3	4
2~3	3	4	6
4이상	4	6	6

3.2 경곗값 증가 문제

표 6은 전통적인 FP 방법을 사용하여 산정된 실제 규모산정 사례이다.

표 6. 실제 규모산정 사례

구 분		낮음			보통			높음			집계	
		개수	가중치	계	개수	가중치	계	개수	가중치	계	기능수	계
데이터	ILF	24	7	168	4	10	40	2	15	30	30	238
	EIF	8	5	40	0	7	0	0	10	0	8	40
트랜잭션	EI	8	3	24	12	4	48	4	6	24	24	96
	EO	12	4	48	20	5	100	8	7	56	40	204
	EQ	25	3	75	16	4	64	8	6	48	49	187
											113	765

표 7. 경곗값의 증가 결과

구 분		낮음->보통			보통->높음			높음			집계	
		개수	가중치	계	개수	가중치	계	개수	가중치	계	기능수	계
데이터	ILF	24	10	240	4	15	60	2	15	30	30	330
	EIF	8	7	56	0	10	0	0	10	0	8	56
트랜잭션	EI	8	4	32	12	6	72	4	6	24	24	128
	EO	12	5	60	20	7	140	8	7	56	40	256
	EQ	25	4	100	16	6	96	8	6	48	49	244
											113	1014

만약 표 5의 각 FP 각 구성요소별로 DET의 값이 경곗값에서 단지 1씩 증가하는 경우 표 7과 같이 총 765FP에서 총 1,014FP로 증가하게 되며, C 언어로 구현하는 경우 31,972라인이 증가하게 되는 문제점이 발생한다. 개발 초기 단계에서 전통적인 FP 산정의 모호성으로 인해 발생하는 문제는 심각할 수 밖에 없다.

3.3 빅 데이터 미반영 문제

앞에서 빅 데이터 분석이 가장 효과적일 때는 데이터의 사이즈 자체가 연구 문제의 일부일 경우이며, 이와 같이 빅 데이터의 개념은 데이터의 크기를 이야기하는 것이 가장 기본적인 접근법이라고 주장하였다. 그러나 전통적인 FP 산정 방법론은 일정 규모 이상은 모두 동일한 규모로 간주하고 있다.

표 1의 ILF 복잡도 가중치를 살펴 보면 DET의 수가 51개 이상인 경우 모든 동일한 규모로 간주한다. 즉, DET의 수가 51개인 경우와 100개인 경우가 동일한 규모로 간주되며, DET의 수가 200개 이상인 극단적인 경우에도 51개와 동일한 규모로 간주된다. RET의 경우에도 마찬가지이다. 나머지 표 2부터 표 5를 살펴 보아도 마찬가지 문제점이 존재한다.

이와 같이 데이터의 규모가 크고, 다양한 3V 혹은 게다가 복잡한 3V+1C 특성을 가지는 빅 데이터의 경우, 실제 규모를 반영하지 못하는 전통적인 FP 산정 방법에는 비합리적인 규모 산정 방법을 내포하고 있다.

4. FP의 조정과 응용

4.1 맘다니형 퍼지 추론

퍼지 추론 시스템은 출력값에서 분명한 해 하나를 얻기 위해 모든 출력 퍼지 집합을 단일 출력퍼지 집합으로 통합하는 방법을 의미한다.

일반적으로 퍼지 추론 방법은 맘다니형 추론과 스게노형 추론으로 구분된다.

맘다니형 추론은 가장 흔히 쓰이는 퍼지 추론 기법으로 1975년에 런던 대학교 교수 맘다니가 보일러가 결합된 증기 기관을 제어하기 위해 최초의 퍼지 시스템을 만들기 위해 경험 있는 기술자가 제공한 퍼지규칙을 적용했다.

맘다니형 퍼지 추론 과정은 다음의 네 단계로 진행된다[9].

- 1단계 : 입력 변수의 퍼지화
- 2단계 : 규칙 평가
- 3단계 : 출력으로 나온 규칙의 통합
- 4단계 : 역퍼지화

맘다니형 추론 규칙을 설명하기 위해 다음과 같이 가정하자.

- x, y, z: 언어 변수
- A1, A2, A3: 논의 영역 X 상의 퍼지 집합에서 정해지는 언어 값
- B1, B2: 논의 영역 Y 상의 퍼지 집합에서 정해지는 언어 값
- C1, C2, C3: 논의 영역 Z 상의 퍼지 집합에서 정해지는 언어 값

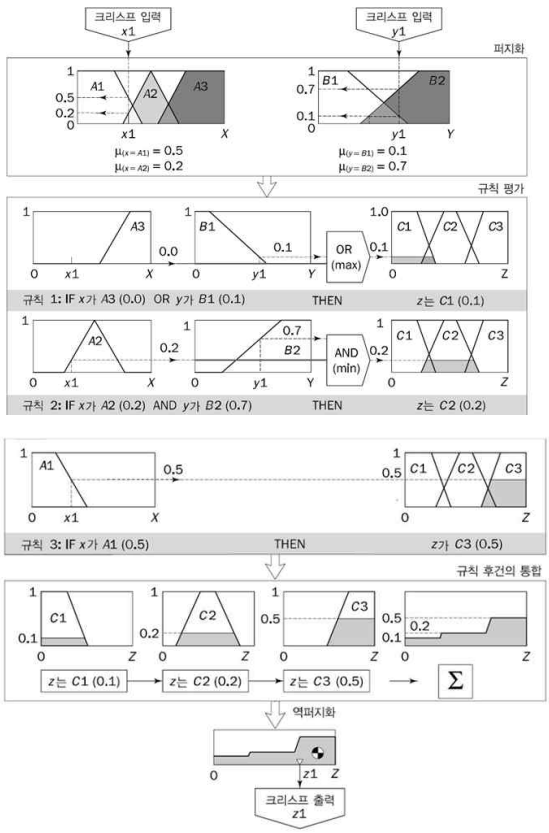

그림 3. 퍼지 추론 시스템

위의 가정을 이용하면 맘다니형 퍼지 추론 시스템은 그림 3과 같이 정의된다.

4.2 퍼지 추론 시스템을 적용한 FP 조정

Matlab Toolbox를 이용하여 개발한 퍼지 FP 산정 시스템의 소속도 함수는 그림 4와 같다.

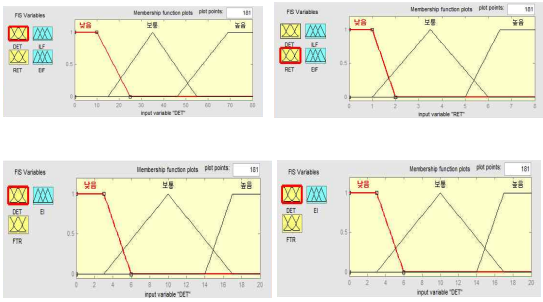

그림 4. 퍼지 FP 시스템의 소속도 함수

퍼지 FP 시스템에서 DET, RET의 개수가 각각 (1, 1) (50, 1), (19, 5)일 때 산정된 FP값은 그림 5와 같다.

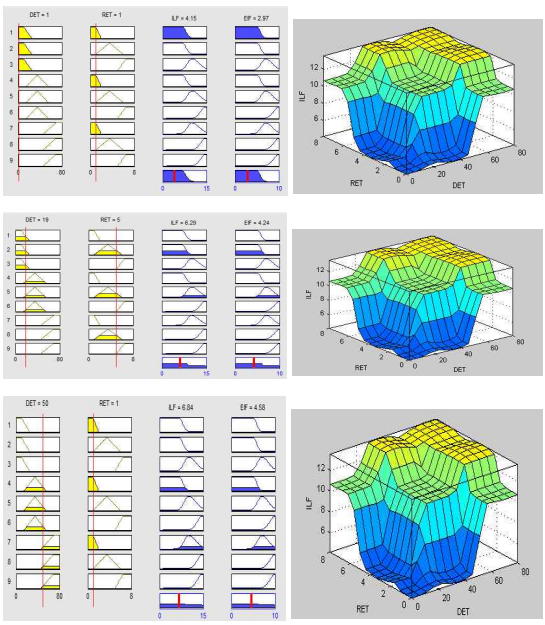

그림 5. 퍼지 FP 시스템의 산정 결과

퍼지 FP 시스템의 데이터 기능요소별 복잡도 가중치와 기존 FP를 비교한 결과는 표 8과 같다.

표 8. 기존 FP와 퍼지 FP의 가중치 비교

기능점수	DET	RET	복잡도	ILF	EIF
기존 FP	1	1	낮음	7	5
	50	1	낮음	7	5
	19	5	낮음	7	5
Fuzzy FP	1	1	낮음	4.15	2.97
	50	1	낮음	6.84	4.58
	19	5	낮음	6.29	4.24

5. 결 론

최근에 대두된 무한한 빅 데이터 활용 가능성은 기존 금융, 통신 사업을 비롯해 의료, 농업, 국방, 교통 정보 등 전 세계 다양한 분야에서 증가하고 있다. 또한 전염병과 자연재해의 이동경로를 파악하는 등 데이터 자원의 무궁무진하게 분석, 예측이 빅 데이터를 이용한 분석과 해석으로 가능하게 되었다. 또한 기업 환경에서는 의사결정에 필요한 의미 있는 정보의 발견, 분석 능력이 방대한 정보 속에서 기업 비즈니스의 핵심 경쟁력으로 부상하고 있고 분석 도구와 비즈니스 인텔리전스 플랫폼 서비스에 대한 수요 증가가 이어지고 있다. 그러나 빅 데이터 프로젝트 관리에 있어서 초기 개발 단계에서 정확한 소프트웨어 개발 비용 및 규모 산정의 중요성에 불구하고 빅 데이터 환경을 반영하는 비용산정 연구는 거의 이루어지지 않았다.

빅 데이터 환경에서 전통적인 FP 기법은 빅 데이터의 특성을 반영할 수 없는 한계를 가지고 있다. 본 연구는 이러한 한계를 극복하기 위한 시도로서 Fuzzy Logic을 적용한 FP 기법을 제안하였다. 본 연구에서 제안된 기법은 ILF의 실제 DET가 50배, RET가 5배로 증가하더라도 실제 규모를 반영할 수 있다. 또한 ILF의 DET가 100개 이상인 경우에도 실제 규모를 반영하여 적절한 규모산정이 가능하다. 복잡도 경계값에서 DET가 1개 증가(감소)하면 증가 (감소)하는 수준만큼 FP값 증가(감소)하고, 개발 초기 단계에서 FP 산정이 모호하더라도 파급 효과를 최소화할 수 있다. 본 연구에서 제안하는 퍼지 FP 기법은 빅 데이터 환경에서 다량의 DET와 FTR을 가지는 소프트웨어에 적절한 규모산정이 가능하다. 추후 연구과제는 실제 빅 데이터 소프트웨어 비용산정에 적용한 후 세부적인 가중치를 조정하고 완전한 퍼지 추론 FP 산정 시스템을 개발하는 것이다.

참 고 문 헌

[1] 이병엽 외, “빅 데이터를 이용한 소셜 미디어 분석 기법의 활용”, 한국콘텐츠학회논문지 '13 Vol. 13 No. 2, 2013. 2.

[2] Gryman, G, “Tapping into power of Big Data”, Technology Forecase, pp.4-13, 2010(3).

[3] 이정미, “빅데이터의 이해와 도서관 정보서비스에의 활용”, 한국비블리아학회지, 24(4): 53-73, 2013. 12.

[4] Miller, H. E. “Big-data in Cloud Computing: A Taxonomy of Risks.” Information Research, 18(1), 2013.

[5] Loukides, M. “What is Data Science?”, Sebastopol, CA: O'Reilly Media, 2012.

[6] Manyika, J. et al., “Big Data: The Next Frontier for Innovation, Competition, and Productivity”, McKinsey Global Institute, 2011.

[7] Lehong, H. and D. Laney. “Toolkit: Board Ready Slides on Big Data Trends and Opportunities”, Stamford, CT: Gartner, 2013.

[8] Park, H. W. and L. Leydesdorff, “Decomposing Social and Semantic Networks in Emerging Big Data

Research." Journal of Informetrics, 7: 756-765. 2013."
[9] 김용혁 역, "인공지능 개론", 한빛미디어, 2013.
[10] 권기태, "빅 데이터 SW 산정을 위한 Fuzzy Logic과 FP의 결합", 2014년 한국소프트웨어감정평가학회 춘계학술발표대회 논문집, 2014, 5.

저 자 소 개

권 기 태

1986년 서울대학교 계산통계학과 졸업
1988년 서울대학교 계산통계학과 석사 졸업
1993년 서울대학교 계산통계학과 박사 졸업
1996년 미국 Univ. of Southern California, 전산학과 Post-Doc.
현재 강릉원주대학교 컴퓨터공학과 교수

<주 관심분야 : 소프트웨어공학, 데이터마이닝, 지능시스템>