www.calsec.or.kr
http://dx.doi.org/10.7838/jsebs.2013.18.1.097

유사도 측정 데이터 셋과 쓰레숄드

Practical Datasets for Similarity Measures and Their Threshold Values

양병주(Byoungju Yang)*, 심준호(Junho Shim)**

초 록

방대한 량의 전자상거래 데이터 객체를 다루는데 같거나 유사한 객체들을 찾는 유사도 측정은 중요하다. 객체간 유사도 측정은 객체 쌍의 유사도 측정값을 비교하므로 객체 량이 많아질수록 오랜 시간이 걸린다. 최근의 여러 유사도 측정 연구에선 이를 더 효율적으로 수행하는 기법을 제시하고 실제 데이터 셋에서 그 성능을 평가해왔다. 본 논문에서는 이들 연구에서 사용하는 데이터 셋의 특성과 실험에서 사용되는 쓰레숄드 값이 가지는 의미에 대해 분석해본다. 이러한 분석은 새로운 유사도 측정 기법의 성능 평가 실험의 참조 기준을 제시하는 역할을 한다.

ABSTRACT

In the e-business domain where data objects are quantitatively large, measuring similarity to find the same or similar objects is important. It basically requires comparing and computing the features of objects in pairs, and therefore takes longer time as the amount of data becomes bigger. Recent studies have shown various algorithms to efficiently perform it. Most of them show their performance superiority by empirical tests over some sets of data. In this paper, we introduce those data sets, present their characteristics and the meaningful threshold values that each of data sets contain in nature. The analysis on practical data sets with respect to their threshold values may serve as a referential baseline to the future experiments of newly developed algorithms.

키워드 : 유사도 측정, 벡터, 쓰레숄드
Similarity Measure, Vector, Threshold

본 연구는 숙명여자대학교 2012 교비연구로 수행되었음.
* Security Solution Division, Samsung Techwin Co.
** Corresponding Author, Division of Computer Science, Sookmyung Women's University
(E-mail : jshim@sookmyung.ac.kr)
2013년 01월 24일 접수, 2013년 02월 12일 심사완료 후 2013년 02월 15일 게재확정.

1. 서 론

인터넷을 활용한 전자 상거래 시스템에는 방대한 양의 사용자 및 제품 정보 객체 등이 저장 관리 되고, 이러한 정보 객체의 양은 전자 상거래 시스템과 인터넷의 발전과 함께 급속히 확대되고 있다. 구글(google)로 대표되는 웹검색 엔진 등은 이러한 방대한 정보들을 수집, 처리한 후 일종의 웹 데이터베이스(Web database)의 형태로 일반 사용자들로 하여금 원하는 인터넷 정보 객체를 검색할 수 있도록 해준다.

웹 데이타베이스에 보관 제공되는 방대한 양의 정보의 상당 혹은 일정 부분은 서로 비슷하거나 동일한 정보인 경우가 현실이다. 웹 사용자가 원하는 정보 객체를 찾고자 할 때 동일하거나 비슷한 정보를 검색 결과로 제공하는 것은 바람직하지 않을 수 있다[4, 13].

예를 들어 웹 사용자가 최신휴대폰을 구매하고자 하려는 의도에서 검색 창에 '최신휴대폰'을 입력하고 결과로서 사진 등의 이미지를 검색하고자 한다고 하자. 그러면 웹 데이타베이스는 최신 휴대폰 정보 이미지 객체로서 예를 들어 아이폰 5의 비슷한 이미지 여러 개보다는 갤럭시 S3와 기타 다른 제품들의 이미지나 동일한 제품의 이미지들 중에서도 사용자에게 다른 느낌을 주는 서로 다른 이미지를 제공해주는 것이 바람직하다.

컴퓨터 분야에서 자주 사용되는 웹을 통한 논문 검색 데이터베이스에서도 마찬가지이다. 예를 들어 사용자가 'Cosine Similarity'에 대한 ACM, IEEE 등의 학회에서 발표된 논문을 찾고자 했을 때 동일한 논문을 여러 번 탐색 결과로서 제공하기 보다는 서로 다른 논문을 탐색 결과로서 제공하는 것이 사용자의 목적에 부합될 수 있어 더 바람직하다.

웹 데이타베이스에 보관된 정보 객체들 간의 중복 및 유사성을 파악하기 위하여 흔히 사용되는 기술적인 방법은 유사도 측정(similarity measures, 혹은 proximity, similarity coefficients라고 불림) 이다[4, 5]. 예를 들어 두 개의 정보 객체가 실제로 얼마나 유사한지 살펴보기 위해, 두 정보 객체를 표현하는 데이터 값을 수학적 함수인 유사도 측정 함수로 입력 값으로 받아 그 계산된 결과값이 바로 두 정보 객체간의 유사성을 가리키는 바로미터의 역할을 할 수 있다.

따라서 유사도 측정 기법은 정보 객체가 어떠한 방식으로 표현되며, 사용하는 유사도 측정함수가 무엇인지에 따라 그 효율성이 결정되는 것은 당연하다. 최근 들어 많은 논문에서 언급되고 있는 데이터 표현 방식 중 하나는 정보 객체를 벡터(vector)로서 표현하는 방식이고, 유사도 측정 함수로서 코사인 유사도 측정(cosine similarity measure)이라고 볼 수 있다[4, 5, 13]. 본 논문에서는 이 두 가지 기법에 집중한다.

정보 모델링과 유사도 측정 함수의 효율은 현실적으로 저장 관리하고 탐색하려는 데이터 응용(application)이나 도메인(domain)에 연관한다. 따라서 더 성능이 좋은 새로운 유사도 측정 함수를 개발하려고 하는 대부분의 연구 방법론은, 개발한 새로운 유사도 측정 함수를 해당 도메인에서 사용되는 데이터 셋(data set)에서 실험적으로 돌려보고 그 성능을 평가하는 실험 수행이 필수적이다. 이를 위해서는 실험에 사용할 적절한 실제 데이터 셋을 분석하고, 성능 평가에 필요한 유용한 객관적인

실험 환경을 구축하는 것은 필수적이다.

본 논문에서는 바로 이러한 점에 착안하여, 최근의 벡터 정보 객체 모델링 분야에서 개발된 유사도 측정 기법들이 사용하는 데이터 셋에 대한 유용한 정보를 분석 제공하고, 실제 성능 평가 환경 구축에서 사용되는 지표를 제공하는데 그 연구 목적을 둔다.

본 연구에서 한정하는 코사인 유사도 측정 함수 등의 개발에 있어서는 유사도 값이 특정 쓰레숄드(threshold) 값을 넘어가는 경우 유사하다고 판단하게 되므로, 실험하는 데이터 셋과 적정한 쓰레숄드의 설정은 매우 중요하다. 따라서 데이터 셋에 따른 의미 있는 쓰레숄드의 설정 등은 이전 연구들과 비교되게 되므로 각종 비교 연구들에서 사용한 데이터 셋과 의미 있는 쓰레숄드의 제시는 더욱 중요하다. 이러한 연구는 앞으로 개발될 새로운 유사도 측정 함수의 객관적인 실험 환경을 구성하는 정보로서 그 유용성이 존재한다.

본 논문의 구성은 다음과 같다. 먼저 제 2장에서는 최근의 유사도 측정 함수 개발에 대한 대표적 관련연구를 소개하고, 제 3장에서는 대표적 연구에서 사용된 데이터 셋에 대한 소개를, 제 4장에서는 데이터 셋에 대한 의미 있는 쓰레숄드 값의 설정을 하나의 데이터 셋을 예로 들어 설명한다. 마지막으로 제 5장에서는 결론과 연구 방향을 제시한다.

2. 관련 연구

2.1 유사도 측정

Lee[4], Lee[5]에서는 유사도 측정 함수의 종류, 적용되는 분야 등을 정보 객체의 표현과 함수들의 대수적 특성에 따라 비교 분석하였다. 특히 정보 객체의 표현을 벡터 정보 객체로 한정하고, 해당 객체의 특질(feature)에 대응하는 각 차원에 0이 아닌 값을 통해 표현한 웹 문서에 대한 벡터 유사도 함수를 다루고 있다.

코사인(cosine) 벡터 유사 측정 함수는 Tanimoto, Dice와 같이 측정 함수의 대수적 계산식에 다음과 같이 두 벡터의 곱(dot product)을 가지고 있다[4].

$$S_{\cos}(x,\ y) = \frac{\sum_{i=1}^{m} x_i y_i}{\sqrt{\sum_{i=1}^{m}} \sqrt{\sum_{i=1}^{m} y_i^2}}$$

여기서, x, y는 주어진 두 벡터를 나타내는데, 각 벡터는 m개의 차원을 가지고 있고, 이들 구 벡터의 코사인 유사도는 $S_{\cos}(x,\ y)$으로 나타내고 있다.

코사인 유사도 측정은 추천시스템에서의 사용자 성향, 특성(features)으로 모델된 이미지 검색, SNS 소셜 데이터에서의 사용자 클러스터링 등 도메인에서 데이터가 벡터로 모델링에서, 데이터 간 유사도 측정의 주안점으로 데이터 특질 및 그 경향(tendency) 혹은 방향성이 중요시 되는 분야에서 폭넓게 사용된다.

2.2 벡터 유사도 측정 기법

Yang[12], Yang[13]에서는 최근 대표적으로 사용되는 유사도 측정에 대한 작동 원리와 성능 향상에 대해 소개하고 있다. 벡터 유

사도 측정은 기본적으로 두 벡터간 유사도 값을 계산해야 하므로, 계산하려는 벡터 쌍의 수가 늘어날수록, 유사도 측정 함수가 복잡할수록 더 많은 시간이 걸리게 된다. 웹에서 데이터 크기가 늘어나고, 동시에 데이터 간의 유사도 비교가 중요해짐에 따라, 벡터 유사도 검색을 빨리 하려는 연구가 최근 각광받고 있다.

벡터 유사도 검색을 빨리 하는 데에는 다양한 방법을 취할 수 있지만, 최근 제일 주목받는 두 가지 기법은 별로 비슷하지 않을 벡터 쌍을 가능한 많이 미리 제거하여 벡터 유사도 측정 시간을 줄이려는 필터링 기술(filtering techniques)과 맵리듀스(MapReduce)[2]로 대표되는 대용량 분산 처리시스템 등을 통하여 주어진 일을 다수의 컴퓨터에서 동시에 수행하여 그 수행 시간을 단축시키려는 방법이라 볼 수 있다.

이러한 관점에서 보았을 때[1, 5, 6, 11, 12, 13] 등은 최근의 대표적 벡터 유사도 측정 알고리즘을 소개한 연구로 볼 수 있고, 이들 연구들에서는 공히 제시된 유사도 측정 알고리즘의 성능을 여러 실제 데이터를 사용한 실험을 통해 제시하고 있다. 하지만 이들 대부분의 연구에서 실제 사용된 데이터와 실험 파라미터(parameter)로 사용된 쓰레숄드 값 등에 대해서는 사용 근거 등이 충분히 제시되지 않았다.

3. 데이터 셋

<Table 2>는 앞 절의 연구들의 성능 실험에서 나타나는 데이터 셋의 일부이다. 이들 연구들에서는 제시하는 유사도 측정 알고리즘의 성능 평가 실험에서 이들 데이터 셋의 일부 혹은 몇 개를 뽑아 사용하고 있다.

UKBench는 이미지 추출 및 비교 검색 벤치마킹을 위해 University of Kentucky에서 개발한 데이터 셋이다[8]. 각 이미지는 특질(feature) 벡터 값으로 표현되는데 데이터는 벡터로 표현되고 데이터 셋 안에는 동일한 이미지가 여러 가지 다른 모양으로 저장되어 이미지 검색 및 추출 검사를 위해 사용되도록 되어 있다.

MoviesLens는 사용자에게 영화(movies)를 추천해 주기 위하여 영화에 대한 기본 정보와 사용자 별 영화에 대한 선호도를 벡터 값으로 표현하는 데이터베이스이며, university of Minnesota에서 Collaborative 필터링 기법을 사용해 사용자가 원하는 영화를 추천해주는 시스템을 개발하기 구축한 데이터이다[7].

LiveJournal은 LiveJournal.com이란 온라인상에서 Facebook과 같이 소셜 네트워크 서비스를 제공하는 회사이다. Stanford 대학교의 SNAP(Stanford Large Network Dataset Collection) 프로젝트[9]에서는 LiveJournal에서 제공한 사용자와 친구들 간의 관계 정보를 노드(node)와 에지(edge)로 구성된 그래프로서 표현 제공하는데 해당 정보는 다시 벡터로서 재구성 되어질 수 있다.

DBLP는 컴퓨터과학 분야의 각종 저널과 학술대회의 논문 및 저자 정보 데이터베이스로서 University of Trier와 DBLP Team에 의해서 운용되고 있다[10]. 데이타베이스 분야로부터 출발하여 현재는 2M 이상의 컴퓨터과학 분야 각종 논문에 대한 저자, 제목, 출판정보 등을 XML 등의 형식으로 제공해준다.

Last.FM은 각종 음악에 대해 음원 뿐 아니라 가수, 작곡가 등의 음악 정보를 사용자에게 제공하는 웹 사이트로서, 사용자가 청취한 음악에 대한 정보를 데이터베이스화하여 각 사용자에게 맞춤형 음악을 추천하기 위해 데이터베이스와 해당 데이터베이스데 액세스하기 위한 API를 제공한다[3].

TREC은 미국의 National Institute of Standards and Technology(NIST)에서 최신 문서 검색 기법을 개발 공유하기 위한 프로젝트로서, 여러 분야의 문서 데이터베이스를 제공하고 있다. 예를 들어 TREC-9 프로젝트 분과에서는 MEDLINE 온라인 의료 문서 데이타베이스에서 추출된 의료 분야 논문을 데이터베이스로 제공하고 있다.

4. 쓰레숄드 분석

<Table 2>는 앞 절에 소개된 데이터 셋들에 해당 데이터 셋을 사용한 최근의 유사도 측정 알고리즘 연구들을 일목요연하게 정리해서 보여주고 있다.

유사도 측정 연구를 수행하기 위해선 데이터 별로 의미 있는 쓰레숄드 값을 설정해 그 성능을 분석해야 하는데, 알고리즘의 주된 목적이 검색된 데이터 쌍간의 실제 유사도를 정성적으로 높이기 위한 연구와는 달리, 유사도 조인과 같이 더 빠르게 연산을 수행하기 위해서인 경우라면, 다른 연구에서 사용되는 쓰레숄드를 참조하는 것이 객관적인 성능 비교를 위해 바람직하다. <Table 3>에서는 해당 쓰레숄드를 제시하고 있다.

물론 유사도 측정 알고리즘의 성능은 실제 응용에 있어서는 그 정성적인 측면이 가미되어야 한다. 다시 말하면 실제 응용에서 의미가 없거나 사용되지 못할 쓰레숄드를 설정해놓고 실험을 수행하는 것은 의미가 크지 않다. 따라서 사용하는 데이터 셋의 특성을 알고 그에 걸맞은 쓰레숄드를 설정하는 것이 바람직하다.

본 논문에서는 그 예로 이미지 유사도 실험이라면 UKBench 이미지 데이터베이스와 MovieLens 데이터 셋을 사용해, 사용하는 쓰레숄드의 값으로서 0.9, 0.7, 0.5, 0.3 등의 값이 실제 어떠한 데이터를 서로 유사하다고 그 결과로 제공하는지를 분석하고 그 결과를 제시한다.

<Figure 1>은 쓰레숄드 값이 0.9로 설정한 경우 유사하다고 판정되는 이미지로서 그 실제 코사인 유사도 값은 0.9999이다. <Figure 2>는 쓰레숄드 값이 0.7로 설정한 경우 유사하다고 판정되는 이미지로서 그 실제 코사인 유사도 값은 0.84이다.

본 예에서는 유사하다고 검색된 이미지 쌍을 보면 0.99인 유사도에서는 같거나 눈으로 보기에 거의 같은 이미지를 찾아준다는 것을 알 수 있고, 0.7 이상으로 검색된 예에서는 이미지의 상당부분이 유사한 것을 알 수 있다.

<Figure 4>는 쓰레숄드 값이 0.5로 설정한 경우 유사하다고 판정되는 이미지로서 그 실제 코사인 유사도 값은 0.57이다. <Figure 5>는 쓰레숄드 값이 0.3으로 설정한 경우 유사하다고 판정되는 이미지로서 그 실제 코사인 유사도 값은 0.30이다. 그림의 예를 통해 보건데 유사도 0.5나 0.3인 경우에는 실제 이미지의 유사성을 눈으로 보기에 0.7이나 0.9처럼 뚜렷이 찾아내기 어려운 것을 알 수 있다.

<Figure 3>은 쓰레숄드 값이 0.7 이상으로 설정된 또 다른 유사 이미지인데 실제 코사인 유사도 값은 0.76인 경우이다. <Figure 2>에서와 같이 두 그림이 비슷하나 방향성이 틀린 경우로서 코사인 유사도 쓰레숄드의 경우 0.7 이상이면 눈으로 식별 인지하기에 상당히 비슷한 이미지를 찾는 것을 알 수 있다.

쓰레숄드 값이 0.9와 0.7인 경우 코사인유사도 측정 알고리즘이 상당히 비슷한 데이터를 찾는다는 것은 텍스트와 숫자를 벡터의 dimension 값으로 가지는 MovieLens 데이터 셋에서도 발견된다.

<Table 1>에서는 두 사용자 User_2와 User_11362가 각기 본 영화와 각 사용자가 추천한 rating 값이 표기되어 있다. 예를 들어 12 Monkeys(1995) 영화를 두 사용자 모두 관람하였고 그 rating 값이 각기 4를 준 경우인데, 실제 듀 사용자 간의 코사인 유사도 값은 표와 같은 경우 0.97로 되어 두 사용자가 상당히 유사하다고 결과가 나오게 된다.

〈Figure 1〉 Images with threshold 〉 0.9 (in UKBench)

〈Figure 2〉 Images with threshold 〉 0.7 (in UKBench)

〈Figure 3〉 Another images with threshold 〉 0.7 (in UKBench)

〈Figure 4〉 Images with threshold 〉 0.5 (in UKBench)

〈Figure 5〉 Images with threshold 〉 0.3 (in UKBench)

〈Table 1〉 Similar users with threshold 〉 0.9 (in MovieLens)

Movie	User_2	User_11362
12 Monkeys(1995)	4	4
Die Hard with a Vengeance(1995)	3	3
Heat(1995)	-	1
Natural Born Killers(1994)	1	-
Star Trek : Generations (1994)	4	4
Water World(1995)	4	3

〈Table 2〉 Datasets for Similarity Search

datasets	data type	applications	organization	source (as of Feb. 2013)
UKBench	Images	Image recognition	Univ of Kentucky	http://www.vis.uky.edu/~stewe/ukbench/
MoviesLens	Movie ratings	recom-mendation system	Univ of Minnesota	http://movielens.umn.edu
LiveJournal	Social network	Social network service	LiveJournal.com, Stanford Univ	http://snap.stanford.edu/da-ta/soc-LiveJournal1.html
DBLP	Author and paper database	Paper and author searching	DBLP Team, University of Trier	http://www.informatik.uni-trier.de/~ley/db/
Last.FM	Music	Music recommendation	Last.fm	http://www.last.fm/api
TREC	Paper	Information retrieval	National Institute of Standards and Technology	http://trec.nist.gov/data/t9_filtering.html

따라서 미래에 User_2가 어떤 영화를 보게 되는 경우 User_11362에게도 해당 영화를 추천하는 것이 이른바 추천(recommendation) 시스템에서의 유사도 활용이 되겠다.

〈Table 3〉 Datasets and Threshold Values

datasets	algorithms	used threshold(T) values
UKBench	[12]	$0.3 \leq T \leq 0.9$
MoviesLens	[12]	$0.3 \leq T \leq 0.9$
LiveJournal	[11, 12]	$0.1 < T \leq 0.9$
DBLP	[7, 5, 11]	0.8 = T(Jaccard) [11] $0.7 \leq T \leq 0.95$ [7] $0.5 \leq T \leq 0.95$ [5]
Last.FM	[7]	$0.7 \leq T \leq 0.95$
TREC	[7]	$0.7 \leq T \leq 0.95$
Synthetic(IPs)	[9]	0.5

5. 맺음말

모든 벡터 데이터에 대한 벡터 유사도 검색은 그 데이터 량이 많아질수록 데이터 량의 제곱에 비례하여 그 소요 시간이 증가된다. 전자상거래와 같은 웹 기반 데이터베이스의 데이터 량은 빅테이터(big data)의 대표적 분야에 걸맞게 급속히 증가하고 있다.

벡터 유사도 검색 시간을 단축하려는 알고리즘의 개발에는 실험을 통한 검증이 필수적으로 요구되며, 실험에는 적절한 데이터의 사용 및 올바른 실험 환경의 설정이 필요하다. 필터링 기법을 사용하는 코사인 벡터 유사도 기반 알고리즘에서는 쓰레숄드 값이 그 예라 할 수 있다.

본 논문에서는 살펴본 최근 관련 연구에서의 데이터 셋과 쓰레숄드 값들에 대한 분석

은 단순히 기존에 발표된 유사도 알고리즘의 성능 분석에 국한하지 않고, 새로운 유사도 알고리즘의 제안에 사용될 수 있다. 빅데이타가 사회적으로 화두가 되고 있기에 빠르고 스케일 가능한 유사도 알고리즘은 당분간 많은 연구가 더욱 필요할 것으로 보이며 본 연구는 그 성능 실험 환경 설정에 참조자료로서 유용하게 사용될 수 있다.

References

[1] Bayardo, R. J., Ma, Y., and Srikant, R., "Scaling up all pairs similarity search," In Proceedings of the 16th international conference on World Wide Web, WWW '07, USA, 2007.

[2] Dean, J. and Ghemawat, S., "Mapreduce: simplified data processing on large clusters," Communications of ACM, Vol. 51, No. 1, pp. 107-113, 2008.

[3] Last.fm Web Services, http://www.last.fm/api, 2012.

[4] Lee, D. and Shim, J., "Survey on Vector Similarity Measures : Focusing on Algebraic Characteristic," The Journal of Society for e-Business Studies, Vol. 17, No. 4, pp. 209-219, 2012.

[5] Lee, D., Park, J., Shim, J., and Lee, S. G., "An efficient similarity join algorithm with cosine similarity predicate," In Proceedings of the DEXA (2), 2010.

[6] Metwally, A. and Faloutsos, C., "V-smart join: a scalable mapreduce framework for all-pair similarity joins of multisets and vectors," Proc. VLDB Endow, Vol. 5, No. 8, pp. 704-715, 2012.

[7] Movielens data sets, grouplens research. http://www.grouplens.org/node/73, 2011.

[8] Nister, D. and Stewenius, H., "Scalable recognition with a vocabulary tree," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, pp. 2161-2168, 2006.

[9] Stanford Large Network Dataset Collection, Stanford University, http://snap.stanford.edu/data/, 2012.

[10] The DBLP Computer Science Bibliography, http://www.informatik.uni-trier.de/~ley/db/, 2012.

[11] Vernica, R., Carey, M. J., and Li, C., "Efficient parallel set-similarity joins using mapreduce," In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 2010.

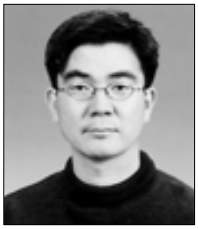
[12] Yang, B., Kim, H., Shim, J., Lee, D., and Lee, S. G., "A MapReduce-based Filtering Framework for Vector Similarity Joins," Technical Report, Seoul National Univ, 2013.

[13] Yang, B., Myung, J., Lee, S. G. and Lee, D., "A mapreduce-based filtering algorithm for vector similarity join," In Proceedings of the ICUIMC(IMCOM) '13, 2013.

[14] Yeon, J., Lee, D., Shim, J., and Lee, S. G., "Product Review Data and Sentiment Analytical Processing Modeling," The Journal of Society for e-Business Studies, Vol. 16, No. 4, pp. 125-137, 2011.

저 자 소 개

양병주 (E-mail : matthew.yang@samsung.com)
2009년 서울대학교 전기공학부 졸업 (학사)
2013년 서울대학교 전기컴퓨터공학부 컴퓨터공학전공 (석사)
2009년 삼성전자 영상보안솔루션사업팀 연구원
2010년~현재 삼성테크윈 시큐리티솔루션사업부 선임연구원
관심분야 데이터베이스, 데이터마이닝, 추천시스템, 맵리듀스, 유사도측



심준호 (E-mail : jshim@sookmyung.ac.kr)
1990년 서울대학교 계산통계학과 졸업 (학사)
1994년 서울대학교 계산통계학과 전산과학전공 (석사)
1998년 Northwestern University, Electrical & Computer Engineering (박사)
2001년~현재 숙명여자대학교 컴퓨터과학부 교수
관심분야 데이타베이스, 전자상거래, 데이타웨어하우스, 빅데이타