pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2013.2.4.281

Machine Learning Process for the Prediction of the IT Asset Fault Recovery

Young-Joon Moon† · Sung-Yul Rhew†† · Il-Woo Choi†††

ABSTRACT

The IT asset is a core part that supports the management objective of an organization, and the fast settlement of the IT asset fault is very important. In this study, a fault recovery prediction technique is proposed, which uses the existing fault data to address the IT asset fault. The proposed fault recovery prediction technique is as follows. First, the existing fault recovery data were pre-processed and classified by fault recovery type; second, a rule was established for the keyword mapping of the classified fault recovery types and reported data; and third, a machine learning process that allows the prediction of the fault recovery method based on the established rule was presented. To verify the effectiveness of the proposed machine learning process, company A's 33,000 computer fault data for the duration of six months were tested. The hit rate for fault recovery prediction was approximately 72%, and it increased to 81% via continuous machine learning.

Keywords : IT Asset Management, Fault Management, Machine Learning, Prediction Process

IT자산 장애처리의 사전 예측을 위한 기계학습 프로세스

문 영 준† · 류 성 열†† · 최 일 우†††

요 약

IT자산은 조직의 경영목적을 지원해주는 핵심영역이며, IT자산의 장애 발생시 신속한 처리를 지원하는 것은 매우 중요하다. 본 연구에서는 IT자산의 장애가 발생할 경우, 장애해결을 위하여 기존의 장애 데이터를 기초로 장애처리 예측 기법을 제시한다. 제안한 장애처리 예측 기법은 첫째, 기존의 장애처리 데이터를 전처리하여 장애처리 유형별로 분류하고 둘째, 분류된 장애처리 유형과 장애 발생 후 접수된 내용을 키워드 매핑시키는 규칙을 제정하였으며 셋째, 제정된 규칙에 의하여 장애 발생 후 장애처리 방법이 사전에 예측 가능한 기계학습 프로세스를 제시하였다. 제시한 기계학습 프로세스의 유효성을 입증하기 위하여 A사에서 6개월 동안 접수된 33,000여건의 전산기기 장애 데이터를 실험한 결과 장애처리 예측의 적중률이 약 72%였으며, 지속적인 기계학습을 통하여 81%로 향상되었다.

키워드 : IT 자산관리, 장애관리, 기계학습, 예측프로세스

1. 서 론

IT자산은 조직의 경영 목표를 효율적으로 달성할 수 있도록 지원해주는 중요 자산이다. 특히 조직 내에 수백 대 이상의 동일한 기종이 배치되어 운영되는 경우라도 서로 다른 업무환경으로 인하여 다양한 장애신고가 접수된다. 일반적인 장애처리 프로세스는 장애신고 접수내용을 확인한 후, 장애처리담당자가 원격 또는 현장을 방문하여 장애를 처리한다. 이때 장애를 처리하는 다양한 방법들 중에서 장애처리담당자는 어떤 방법으로 해결해야할지 개인적인 경험에 의하여 판단하며 실수와 반복을 거듭하여 개인의 처리 능력이 향상된다. 하지만 개인의 경험보다는 여러 사람들의 경험을 바탕으로 시스템에 축적된 데이터가 신뢰성이 높을 것이라고 예측된다.

본 연구에서는 첫째, 장애처리 내용의 텍스트를 어휘 분석하여 유사 키워드를 표준화하는 전처리 작업을 수행한 후, 키워드를 분류하고 출현 빈도수를 적재하는 기계학습 규칙을 정하였다. 여기서 키워드란 용어의 정의는 문장을 공백이나 특수문자 등의 구분자로 분류하고 난 후의 단어나 문맥을 지칭한다. 둘째, 장애접수 내용의 텍스트를 전산기기 기종별, 장애처리 유형별 키워드를 분류하는 알고리즘을 적용하고 지도학습에 의하여 장애내용과 무관한 키워드를 예

※ 이 논문은 숭실대학교 교내 연구비 지원에 의하여 연구되었음.
† 종신회원 : 숭실대학교 컴퓨터학과 박사수료
†† 종신회원 : 숭실대학교 컴퓨터학부 교수
††† 정 회 원 : 강남대학교 교양학부 교수
논문접수 : 2012년 10월 15일
수 정 일 : 1차 2012년 11월 28일
심사완료 : 2012년 12월 3일
* Corresponding Author : Young-Joon Moon(yjmoon@yonsei.ac.kr)

외처리 하였다. 예외처리 후 단일키워드와 조합키워드로 분류하고 추출된 키워드 간의 유사성을 학습하여 같은 의미의 키워드들을 표준화된 하나의 키워드로 추출하는 규칙을 제정하였다. 셋째, 제정된 규칙에 의하여 장애 발생 후 장애처리를 위한 여러 개의 유형 중에서 해결 방법을 사전에 예측할 수 있는 기계학습 프로세스를 제시하였다.

본 연구에서는 고객의 장애접수 내용을 확인하고 가장 최적의 처리방법을 알려주므로 장애 해결을 위한 가장 근접한 판단을 먼저 접근하여 검토할 수 있는 프로세스를 제안하였다. 기계학습을 통하여 시스템이 알려준 처리방법이 실패한 경우, 신규 패턴을 학습시키고 기계학습을 강화시켜 더욱 향상된 장애처리 해결 방법의 예측을 확인할 수 있었다.

2. 관련 연구

IT자산은 조직의 업무를 지원하는 중요한 자산이며 신규 도입 후, 유지보수 프로세스의 효율화를 위한 측정 및 관리가 필요하다[1-4]. 또한 관리에 대한 표준화된 평가 프로세스를 통하여 양질의 IT서비스의 제공이 필요하다[5-6].

기계학습 알고리즘은 입력된 데이터의 학습을 통하여 유용한 정보를 제공해 준다[7-10]. 이를 조직의 업무에 응용할 경우 신속한 의사결정에 도움을 준다.

IT자산의 결함유형 분석 및 측정 지표를 제공하고, 프로그램별 오류 유형을 집계한 후, 측정된 결함의 유형 사례를 해당 업무부서에서 확인함으로써 지속적으로 결함을 개선할 수 있도록 지원하는 연구[11]나 결함 예방 모델의 연구[12] 사례는 있었으나 기계학습을 IT자산의 장애와 연계하여 장애처리 방법을 예측하는 연구는 없었다. 또한 기계학습 분야의 연구에서는 기계학습을 통하여 전문용어를 인식하는 시스템 연구[13]를 살펴보면 통계적 방법을 활용하여 문헌내의 후보용어들의 출현 빈도에 기반을 두어 전문용어를 추출하는 방식을 개선하는 연구가 진행되었으며 후보 용어들의 추출 패턴을 활용하여 추출된 각 후보 용어에 가중치를 할당하는 통계기반 및 기계학습 기반연구를 진행하였다. 이때 할당된 가중치를 기준으로 전문용어인지를 판단하는 과정으로 구성된다. 기계학습 기법을 이용한 전자게시판 질문 자동 분류 연구[14]에서는 문서로 이루어진 전자우편 또는 전자게시판의 고객 상담 내용을 기계학습의 분류기법을 활용하여 담당자를 자동 선정하고 고객의 요구사항에 신속하게 반응할 수 있는 방법을 제안하였다.

본 연구에서는 IT자산의 장애처리 방법을 예측하는 기계학습 프로세스를 제안하였으며, 실패하는 경우 기계학습 강도를 강화하도록 하였다. 장애처리 방법을 예측하기 위하여 장애 접수 건에 대한 문장을 분석하여 키워드를 추출하고 장애 내용과 무관한 키워드를 지도학습에 의하여 예외처리하였다. 키워드는 띄어쓰기나 구분을 나타내는 특수문자로 분류한 단일키워드와 장애 접수 건별로 동일 건에 대한 키워드 간의 연관성 의미를 부여하기 위하여 조합키워드로 구성하였다. 분류한 키워드는 정렬하여 같은 내용의 서로 다른 조사 표현 및 같은 의미의 서로 다른 키워드를 최다 빈도수 출현 키워드를 기준으로 표준화하였다.

본 연구의 예측 프로세스는 장애 접수 건의 문장을 제정된 규칙에 따라 분류하고 기계학습 된 과거 데이터를 활용하여 장애처리 방법을 예측한다. 기계학습에 의하여 추출된 예측이 실패할 경우 향후 성공률을 높이기 위하여 실패한 사례의 장애처리와 관련된 핵심키워드를 기계학습 시킴으로써 예측의 적중률을 향상 시킬 수 있었다.

3. 기계학습 프로세스

3.1 장애처리 유형 분류작업

IT자산의 장애접수 신고 후, 장애가 해결되면 어떤 방법으로 장애를 처리 했는지 시스템에 등록한다. 또한 장애처리 건에 대하여 사후 장애처리 유형별 통계를 내기 위하여 사전에 분류된 장애처리 유형 중 한 가지를 선택해야 한다. 이때 선택 대상 중에 알맞은 것이 없거나 비슷한 것이 여러 개 있으면 고민하게 되고 개인적인 판단에 의하여 임의로 한 가지를 선택하게 되어 실제 장애처리 유형과 장애처리 결과 내용이 상이한 경우를 많이 발견하게 된다.

본 연구의 첫 번째 단계에서는 이러한 문제점을 개선하기 위하여 과거에 등록된 장애처리 유형별로 장애를 처리한 내용을 텍스트마이닝하여 출현한 키워드별로 빈도수를 집계하였으며 추출방법은 Fig. 1과 같다.

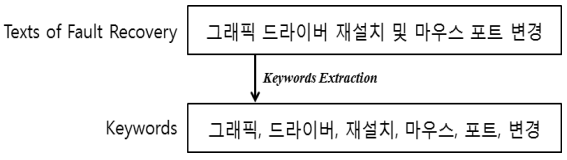

Fig. 1. Keywords Extraction from Texts of Fault Recovery

키워드별 빈도수는 향후 입력된 장애처리 내용이 어떤 유형에 해당되는지 예측할 수 있는 원천데이터로 활용되었다. 키워드별 빈도수는 새로운 장애가 발생하여 처리할 때마다 지속적으로 증가하므로 기계학습에 의한 예측은 변화되며 데이터가 많을수록 더 정확해 진다. 특히 “그래픽”과 “드라이버”, “마우스”와 “포트”는 상호 같이 출현하는 경향이 높다. 이런 경우 기계학습에서 연관성이 높은 키워드로 분류된다. 장애처리를 수행하는 담당자가 1명인 경우는 언어를 사용하는 패턴이 유사하기 때문에 예측의 적중률이 높아지지만 담당자가 많아질수록 같은 내용을 서로 다른 언어로 표현하기 때문에 키워드 분석에 의한 적중률이 낮아진다.

본 연구에서는 이러한 문제점을 해결하기 위하여 가장 많이 발생한 빈도수의 키워드로 표준화시키는 전처리 작업을 수행하였다. 본 연구의 데이터는 신뢰도를 높이기 위하여 5,000억 원 이상의 IT자산을 보유하고 25,000명이상의 계열사 직원들을 대상으로 서비스를 수행하는 IT전문기업인 A사의 전산기기 장애접수 및 처리내역을 대상으로 하였다.

장애유형별 장애처리 내용의 입력사례를 보면 서로 다른 장애처리담당자는 같은 장애처리 유형에 대하여 대부분 서로 다른 입력을 수행하였다. 장애처리담당자가 입력한 동일 장애유형에 대한 입력 데이터는 87건으로 집계되었으며 이 중에서 동일한 입력 내용을 제거하면 20개의 장애처리 유형으로 압축되었으며 Table 1과 같다.

Table 1. Fault Recovery Data on the Same Fault Recovery Type

Fault Recovery Type : Check Disk			
1	Check Disk	11	Check Disk & Terminal Optimization
2	Check Disk & Check Program	12	Check Disk Verification
3	Check Disk & Check Video Card	13	Check-Disk
4	Check Disk & Cleaning Internal	14	Chkdsk
5	Check Disk & Delete Program	15	Chkdsk; Delete Files not in Use
6	Check Disk & Disk Defrag	16	Chkdsk; File Recovery
7	Check Disk & Dust Remove	17	Chkdsk; Recovery Disk Error
8	Check Disk & Program Elimination	18	Run Check Disk
9	Check Disk & OS Reinstallation	19	Run Check-Disk
10	Check Disk & System Optimization	20	Run Chkdsk & File Recovery

Table 1의 사례는 Desktop PC의 “Check Disk”에 해당하는 장애처리 유형이다. 이중에서 “Check Disk”가 58회로 가장 많이 사용되었고 뒤를 이어 “Check”, “Disk”가 각각 12회, 13회 사용되었다. 이러한 경우에는 가장 많이 반복 출현한 “Check Disk”를 장애처리 유형으로 선정한다. 예를 들어 장애처리담당자가 장애처리 결과를 입력할 때 입력된 키워드가 “Check Disk”라는 장애처리 유형의 장애처리 결과 키워드와 가장 많이 일치하면 시스템은 자동으로 “Check Disk”라는 장애처리 유형으로 분류하라고 장애처리담당자에게 알려준다. 즉 장애처리 완료 후 장애처리 결과를 등록할 때, 전처리 작업 수행 후에 이미 기계학습 된 키워드와 일치하면 일치된 키워드 중에서 가장 많이 사용되었던 장애처리 유형을 장애처리담당자에게 알려주는 것이다. 장애처리담당자는 자동으로 제시한 장애처리 유형이 입력하려고 하는 장애처리 유형과 일치하지 않을 경우에만 이미 등록된 장애처리 유형 중 한 가지를 선택하게 하면 훨씬 효율적이다. 여기서 선택하려고 했던 장애처리 유형으로 적중하는 Hit Ratio는 다음 식과 같다.

$$\text{Hit Ratio} = \frac{\text{장애처리 유형 일치 수}}{\text{장애처리 결과 입력 수}}$$

입력한 장애처리 내용의 키워드는 과거 입력된 키워드와 가장 많이 일치하는 장애처리 유형을 알려준다. 장애처리 유형별 데이터가 50건 이상인 유형을 대상으로 측정한 결과 Hit Ratio는 표준화된 장애처리 유형으로 전처리 작업을 수행하지 않을 경우 약 52%이지만 전처리 작업을 수행하고 난 후 약 83%로 증가 되었다.

장애처리담당자가 장애처리 내용을 입력하면 자동화 시스템은 키워드를 분석하고 전처리과정 수행 후, 과거 기계학습 된 장애처리 키워드 빈도수를 파악하여 가장 높은 빈도수의 장애처리 유형을 장애처리담당자에게 알려준다. 알려준 장애처리 유형이 장애담당자가 입력하려고 하는 내용과 일치한다고 판단되면 시스템에 자동 등록되며, 기계학습에 의한 예측 결과가 틀릴 경우에만 장애처리담당자가 장애유형 중 한 가지를 선택하여 시스템에 등록한다. 이에 대한 흐름도는 Fig. 2와 같다.

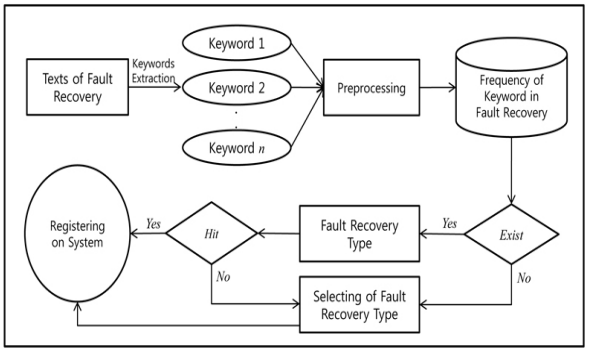

Fig. 2. Fault Recovery Classification Flow

만일 시간의 흐름에 따라 동일 건에 대한 장애처리 입력 키워드의 내용이 달라져서 기존의 키워드보다 새로운 키워드가 더 많이 발생하면 기계학습 결과에 의해 새로운 예측 결과를 제시한다. 이때 장애처리 유형은 새로운 결과를 반영하여 업데이트 된다. 장애처리 유형은 코드로 관리되므로 기존 내용이 신규 내용으로 배치작업에 의하여 자동 업데이트 된다. 이렇게 함으로써 항상 같은 결과를 주는 것이 아니라 변화하는 입력데이터에 맞추어서 기계학습을 수행하고 변화가 반영된 가장 알맞은 결과를 제시할 수 있다.

3.2 매핑과 처리 규칙

앞에서는 전처리 과정을 거쳐서 표준화된 장애처리 유형을 추출하였다. 그 다음 작업으로 장애처리 결과에 해당하는 장애접수 내용의 키워드를 추출하기 위하여 문장을 공백 및 특수문자 중 구분을 나타내는 “/”, “-”, “,”를 기준으로 키워드를 분류하였다. 이중에서 장애와 관련이 없는 특수문자, 숫자, 1개의 글자는 제외시켰으며 또한 장애 내용과 무관한 장소 및 대상을 가리키는 키워드도 사전에 예외처리로 구분하여 기계학습 대상에서 제외시켰다. 실제 사례를 보면 장애처리 유형이 “그래픽카드 교체”인 경우 장애 접수 키워드는 총 76건으로 집계되었다. 각각의 키워드는 기계학습에

Table 2. Result after Correction of Keyword in Texts of Fault

Classification	Keywords	Top 10
Before (76 Cases)	PC, 갑자기, 그래픽카드, 그래픽카드이상, 긴급, 꺼져, 꺼짐, 나오고, 노란색으로, 다시, 다운됨, 단말기, 동일증상, 들어가, 떨림, 멈춤, 모니터이상, 발급기, 발생, 버림, 변하면서, 변함, 본지, 부분이, 부팅불, 부팅시, 부팅안됨, 부팅중멈춤, 분홍색으로, 불가, 블루, 블루스크린, 블루스크린발생, 비디오카드교체했으나, 사용중, 사용중다운, 사용중다운됨, 사용중다운부팅불, 센지, 소음발생, 수차례반복됨, 스캔pc, 스크린, 안전모드, 연결, 오후동일시간멈춤 및, 움직임, 윈도우화면깨짐, 음영생김, 이상, 장애, 잦은장애, 재부팅시, 재부팅하여, 전원불, 전화요망, 전화요청, 정상부팅불, 정상부팅안됨, 조정불가, 조치바랍니다, 주변기기, 중간에, 카드, 파워세이브, 하얀, 해제하였으나, 화면, 화면깨짐, 화면노랗게, 화면번짐, 화면안나옴, 화면안나타남, 화면에, 화면에노란줄생김, 화면크게나옴	스캔pc, 전원불, 전화요청, 부팅불, 부팅안됨, 긴급
First Exception		전화요청, 긴급
Second Exception	갑자기, 나오고, 다시, 동일증상, 들어가, 발생, 버림, 본지, 부분이, 불가, 사용중, 센지, 수차례반복됨, 연결, 오후동일시간멈춤 및, 움직임, 이상, 장애, 잦은장애, 전화요망, 조정불가, 조치바랍니다, 주변기기, 중간에, 카드, 해제하였으나	
After (48 Cases)	PC, 그래픽카드, 그래픽카드이상, 꺼져, 꺼짐, 노란색으로, 다운됨, 단말기, 떨림, 멈춤, 모니터이상, 발급기, 변하면서, 변함, 부팅불, 부팅시, 부팅안됨, 부팅중멈춤, 분홍색으로, 블루, 블루스크린, 블루스크린발생, 비디오카드교체했으나, 사용중다운, 사용중다운됨, 다운부팅불, 소음발생, 스캔pc, 스크린, 안전모드, 윈도우화면깨짐, 음영생김, 재부팅시, 재부팅하여, 전원불, 정상부팅불, 정상부팅안됨, 파워세이브, 하얀, 화면, 화면깨짐, 화면노랗게, 화면번짐, 화면안나옴, 화면안나타남, 화면에, 화면에노란줄생김, 화면크게나옴	스캔pc, 전원불, 부팅불, 부팅안됨

의하여 사용 빈도수가 산출되며 이 값을 가중치로 결정하였다. 이중에서 가중치가 높은 상위 10%이내의 자주 사용하는 키워드 중에서 장애내용과 무의미한 키워드는 Table 2와 같이 지도학습에 의하여 1차적으로 예외처리 하였다. 비지도학습보다 지도학습을 선정한 이유는 어떤 데이터가 올바른지 아닌지 전문가가 개입되어 알려주므로 더욱 효율적이고 정밀하게 학습할 수 있다. 이때 많은 양의 지도학습은 많은 공수를 필요로 하므로 최대한 공수를 줄이기 위하여 지도학습의 노력이 적게 발생하는 상위 10%이내의 키워드는 1차적으로 예외처리 하였고 그 밖의 키워드는 2차적으로 예외처리 하였다. 그러므로 조직의 상황에 따라 전문가에 의한 지도학습으로 1차 예외처리 주기를 2차 예외처리 주기보다 자주 수행할 수 있다. 예외처리는 전문가가 개입되어 지도학습 이후, 새로 발생되는 키워드에 대해서만 예외처리 하므로 최초 지도학습 이후 점차적으로 지도학습의 양이 줄어든다. 전산기기 중 Desktop PC의 장애처리 내용이 “그래픽카드 교체”인 경우, 장애접수 내용 키워드의 보정 후 결과는 Table 2와 같으며 보정 전보다 약 37%가 축소되었다.

Table 2의 “그래픽카드 교체”라는 장애처리 유형의 예외처리 대상 키워드는 장애와 관련이 없는 키워드이므로 공통테이블에 등록하여 모든 장애접수 내용에서 예외처리 되도록 대상에 포함시킨다.

장애접수 키워드의 보정처리 후, 장애접수 내용 중 선정된 키워드를 장애처리 결과에 매핑시키는 작업을 수행하였다. 각각의 장애처리 결과 항목에 매핑된 장애접수 내용의 키워드와 빈도수가 테이블에 함께 저장된다. 또한 연관도 분석을 위하여 동일 장애접수 건에 대하여 보정후에 추출된 키워드가 2개 이상일 경우에는 키워드를 결합하여 장애처리 결과에 매핑시켰다. 2개 이상이 결합된 조합키워드는 전체의 37.5% 였다. 조합키워드는 데이터마이닝의 연관규칙을 활용하여 키워드 간의 신뢰도를 산출하고 상대평가를 수행하였다. 신뢰도가 높은 키워드는 기계학습에 의하여 관리되므로 만일 고객의 장애 접수 내용의 단어가 적어 장애처리 담당자가 판단하기 어려울 때 시스템에 요청하면, 시스템은 출현한 키워드와 신뢰도가 높은 키워드를 자동으로 추출하여 문맥의 이해 형성에 도움을 줄 수 있다. 단일키워드와 조합키워드의 분포 사례는 Table 3과 같다.

Table 3. Frequency of Keyword in Texts of Fault

Classification	Keywords	Frequency
Single Keywords	부팅불	9
	부팅안됨	9
	스캔pc	5
	전원불	5
	소음발생	4
	정상부팅불	4
	블루스크린	3
	Omitted	
	화면에노란줄생김	1
	화면크게나옴	1
Combination Keywords	블루스크린+발생	5
	다운+블루스크린	4
	단말기+멈춤	4
	Omitted	
	스캔pc+부팅시+화면안나타남	1
	스캔pc+부팅안됨	1
	스캔PC+정상부팅불	1
	화면+노란색으로	1
	화면+음영생김	1
	화면노랗게+떨림	1

키워드를 분석하면 특정 키워드가 반복되어 출현하는 현상이 많으므로 장애처리 유형 추출을 위한 키워드의 분석 및 보정 작업 후에, 유사한 키워드를 최다 출현한 키워드로 표준화 작업을 수행하였다. Table 3에서 “부팅불” 9건, “부팅안됨” 9건, “정상부팅불” 4건은 같은 의미이므로 합산되어 22건으로 시스템에 기록한다. 이렇게 하면 기계학습에 의하여 같은 의미의 키워드를 서로 다르게 카운트하여 예측을 잘못하는 경우를 예방할 수 있다.

장애처리 예측의 기계학습 생성규칙은 Table 4와 같으며 장애접수 문장내용 S는 여러 개의 단어 조합으로 구성되어 있으며 키워드인 K로 분류되고, K는 기계학습 처리함수인 ML에 의하여 관련이 있는 유형끼리 군집화하여 하나이상의 장애처리 예측결과를 생성한다.

Table 4. Production Rule of Machine Learning in Fault Recovery Prediction

$G = (\{S, K\}, \{r_p\}, P, S)$
P : A set of Production Rule
S : Texts of Fault reported data
K : A set of Keywords Extraction from Texts of Fault reported data
n : A count of Keywords Extraction from Texts of Fault reported data
ε : Empty
i : Keyword
ML : Machine Learning processing function
r_p : Result of Fault Recovery Prediction
M_t : Type of Matrix

$P : S \rightarrow K_1 \mid K_2 \mid \bullet \bullet \mid K_n \mid \varepsilon$
$\sum_{i=1}^{n} K_i \xrightarrow{ML} r_{p\,i} \mid \varepsilon$
$\sum_{i=1}^{n} r_{p\,i} \rightarrow M_t + r_{p\,i}$
$M_t \xrightarrow{ML} r_p$

기계학습 처리함수 ML의 장애처리 예측결과 추출은 Fig 3과 같으며 장애접수 문장의 키워드뿐만 아니라 신뢰도가 높게 형성된 키워드도 연관규칙에 의해 관련성이 높으므로 테이블에 별도 저장한다. 만일 예측이 실패하더라도 실패의 원인이 된 핵심키워드를 추가로 입력하여 상호 연관성 학습을 수행한다. 만일 Fig. 3의 B그룹으로 장애처리 방법이 결정된 사례가 자주 발생하면 A그룹으로 옮겨서 장애처리 예측결과에 포함시켜 예측의 적중률을 향상 시킬 수 있다.

3.3 장애처리의 사전 예측

앞에서 장애처리 유형별로 매핑된 테이블의 키워드별 빈도수는 장애 접수건의 증가와 함께 지속적으로 변화하며 기계학습에 반영된다. 기계학습은 불필요한 장애접수 키워드를 제외되고 유용한 데이터만 지속적인 학습이 이루어지도록 하였다. 신규 장애접수 건이 발생하면 장애접수 내용의

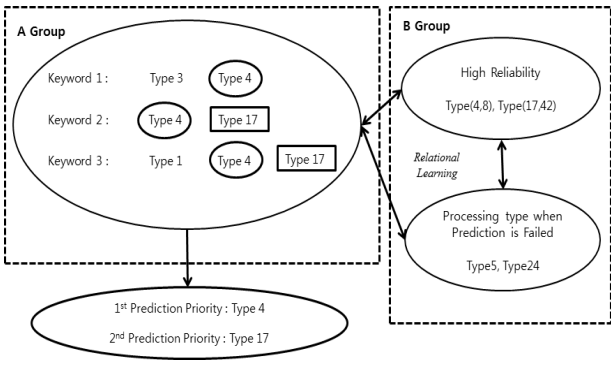

Fig. 3. Fault Recovery Result Extraction Technique

어휘를 분석하여 사전 등록된 테이블의 키워드와 일치하는지 확인하고 빈도수가 가장 높은 장애처리 방법을 찾아내어 예측하도록 한다. 예를 들어 장애접수 내용의 문장을 분석한 결과 키워드가 3개라면 각 키워드별로 빈도수가 가장 높은 단일키워드 장애처리 결과와 조합키워드 장애처리 결과를 모두 반영한 이행적 폐쇄행렬의 결과 함수 값 M_t을 추출하여 장애처리 유형을 예측할 수 있도록 하였으며 Fig. 4와 같다.

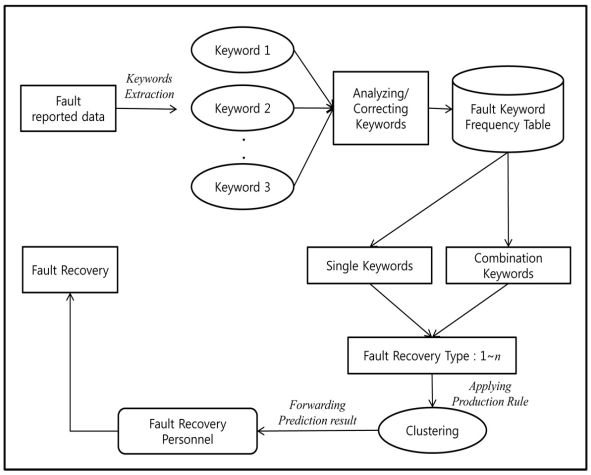

Fig. 4. Clustering Process of Fault Recovery Type Prediction

유사한 장애처리 유형이 많을수록 예측 결과가 한 개의 유형으로 일축되기 어렵다. 이런 경우 유사한 장애처리 유형끼리 군집화를 형성하여 장애처리담당자에게 제시하도록 하였다. 이렇게 장애처리담당자에게 기계학습에 의하여 군집화 된 몇 가지의 장애처리 예측 유형을 알려주면 장애처리의 판단이 빨라진다.

장애처리 예측결과는 장애처리담당자가 인지하고 판단하며, 만일 시스템이 알려준 예측결과가 단일키 우선순위와 조합키 우선순위 중 일치하는 쪽에 인센티브를 부여하여 기계학습을 강화시킨다. 만일 일치하지 않더라도 장애처리 학습테이블에 실패사례를 저장하고 향후 반영되도록 학습 시킨다. 이렇게 단일키워드와 조합키워드의 두 가지 측면을 모두 활용하여 예측결과의 실패 시, 학습을 강화하는 것을

본 연구에서는 제안하였다. 즉, 단일키워드는 최빈수를 구하는 통계적 학습을 수행하고 조합키워드는 키워드 상호간의 연관성을 학습한다. 또한 예측의 결과가 틀렸을 경우 별도의 연관규칙을 설정하여 강화학습을 수행하도록 하였다. 이러한 관계는 Fig. 5와 같이 이행적 폐쇄행렬로 구현하여 나타내었다. 행렬의 주대각선은 행과 열이 같으므로 단일키워드에 대한 빈도수를 나타내며, 주대각선이 아닌 경우는 행과 열이 서로 다른 키워드 간의 연관성을 의미하며 조합키워드로 분류할 수 있도록 하였다. Fig. 5의 (a)를 보면 이행적 폐쇄행렬은 대칭행렬로 구성되므로 같은 데이터가 저장되는 장소를 절약하기 위하여 (b)와 같이 상삼각행렬로 변경하였다. 또한 예측결과가 틀렸을 경우 학습을 강화하기 위하여 (c)와 같이 별도의 이행적 폐쇄행렬로 나타내었다. 원소 값이 1인 경우 행과 열로 표현된 키워드 간의 가중치를 부여하는 학습이 이루어진다. 만일 주어진 키워드의 예측 결과가 반복하여 틀릴 경우 학습의 강도를 강화하기 위하여 가중치를 반영하여 준다. 가중치는 예측결과의 실패로 인한 학습의 횟수를 n으로 보았을 때 log로 계산하며 별도 항목으로 관리한다. 예를 들어 예측의 실패로 인한 학습의 횟수가 8번이면 log8 이므로 3배의 가중치를 부여하고 9번부터 16번까지는 log16 으로 간주하여 4배의 가중치를 부여한다. 가중치를 포함하여 장애처리 유형 중에서 상삼각 행렬의 각 원소의 합이 가장 높은 경우를 우선순위로 예측한다.

결과적으로 장애처리 학습테이블에는 중요 정보가 축적되며 적중할수록 선택의 가중치는 높아진다. 장애처리 학습별로 가중치가 10이 넘으면 온라인 FAQ게시판에 입력하여 사용자가 스스로 장애처리 하도록 안내할 수 있으며, 향후에 지식관리시스템(KMS)과 연계하여 신속하게 일을 처리할 수 있는 기반을 마련할 수 있는 계기가 될 수 있다. 전체적인

$$A^{+}=\begin{pmatrix} 36 & 7 & 3 & 8 \\ 7 & 14 & 11 & 6 \\ 3 & 11 & 24 & 10 \\ 8 & 6 & 10 & 50 \end{pmatrix} \quad A'^{+}=\begin{pmatrix} 36 & 7 & 3 & 8 \\ 0 & 14 & 11 & 6 \\ 0 & 0 & 24 & 10 \\ 0 & 0 & 0 & 50 \end{pmatrix}$$

(a) Transitive Closure Matrix (b) Upper Triangular Matrix

$$M\!L^{+}=\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} * \log n$$

(c) Reinforcement Learning

Fig. 5. Matrix for Prediction

세부 프로세스는 Fig. 6과 같으며 본 연구에서 제안하는 장애처리 예측 기법의 세부 처리내용은 다음과 같다.

① 여러 명의 장애처리담당자가 장애 조치 후 처리내용을 입력한 텍스트 문장을 표준화된 장애처리 유형별로 분류하기 위하여 전처리 작업을 준비한다.

② 전처리 작업 후 장애처리담당자가 입력한 비표준화된 텍스트 문장은 표준화된 장애처리 결과 문장으로 변경된다.

③ 장애처리 결과는 장애처리 유형으로 귀속되며 빈도수가 함께 저장되어 장애발생 시 어떤 방법으로 조치하였는지 집계가 되고 기계학습에 활용할 수 있는 데이터로 축적된다.

④ 분류된 장애처리 유형은 장애내용별 키워드 빈도수와 매핑하기 위하여 테이블의 장애처리결과 속성 값에 반영된다.

⑤ 고객이 장애접수 게시판을 통하여 입력한 데이터 또는 고객이 전화로 요청하여 전화 접수 담당자가 입력한 장애내용의 텍스트 데이터를 키워드별로 분류한다.

⑥ 사전에 지도학습 된 제거될 데이터가 있는지 분석하고 필요한 키워드만 추출한 후 분석 및 보정작업을 걸쳐 유사 키워드를 표준화 하고 DB에 저장한다.

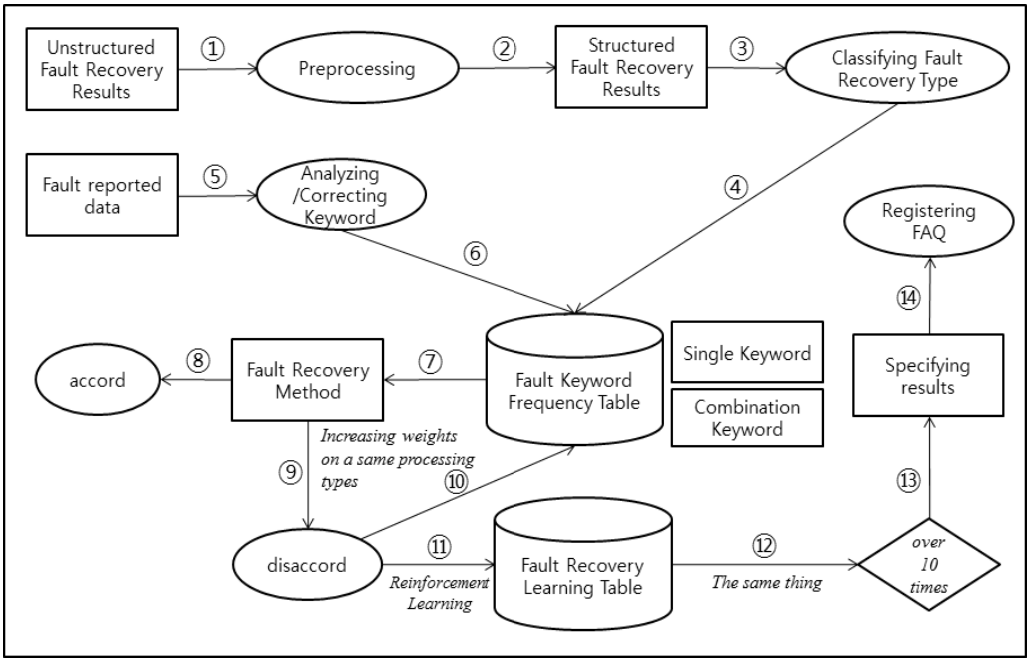

Fig. 6. Flow of Fault Recovery Prediction Technique

Table 5. Fault Recovery Type and its Frequency

Fault Recovery Result	Frequency	Fault Recovery Result	Frequency
Printer Configuration Change	3095	Operating-System Reinstallation	445
Operating-System Configuation Change	2678	Scanner Configuration Change	396
Printer Cleaning	1504	Barcode Reader Configuration Change	360
Device Driver Reinstallation	1386	Malignant Code Elimination	348
HDD Replacement	1174	Network Configurating	328
Ddrivers Reinstallation	1105	Overall System Reinstallation	311
Printer Reinstallation	970	Display Reinstallation	302
Connecting Cable Reconnection	965	Power Replug-in	301
Printer Unit Replacement	856	Scanner Reconfiguration	286
Card Issuing Device Reinstallation	837	Cooling-Fan Cleaning	277
Printer-Sharing Reconfiguration	723	Printer Paper Feeder Replacement	253
Web-Browser Configuring	720	Display Connecting Cable Replacement	252
Installed Program Elimination	652	External Device Reconfiguration	237
Connecting Cable Replacement	538	Temporary Files Elimination	236
Power Supply Unit Replacement	513	Barcode Reader Unit Replacement	223

⑦ 장애접수 내용에 규칙을 적용한 기계학습을 수행한 후, 장애처리 방법을 예측하여 알려준다.
⑧ 확인결과 일치되었으면 Success를 등록한다.
⑨ 확인결과 불일치되었으면 Fail을 등록한다.
⑩ 성공한 장애접수 키워드의 빈도수를 증가시켜 추가적인 기계학습을 수행한다.
⑪ 기계학습 결과와 불일치되었을 때 불일치한 원인을 시스템이 분석하고 Success 유형의 추가 학습 가중치를 높여서 다음에 선택할 확률이 유리하도록 기계학습 시킨다.
⑫ 장애처리 기계학습을 통하여 축적된 성공사례가 10회가 넘으면 Best Practices로 간주한다.
⑬ 축적된 성공사례가 10회가 넘으면 처리결과를 전문가가 시스템에 상세히 입력하여 지식을 향상시킨다.
⑭ 장애접수 유형별로 장애처리 결과를 상세화 한 후, 시스템 내에 FAQ 메뉴에 등록하여 고객이 동일 장애 및 처리 사례가 있는지 FAQ를 먼저 검색하고 스스로 조치할 수 있도록 하여 조직의 업무처리를 효율화 시킨다.

4. 검증 및 실험

검증 및 실험을 위하여 5,000억원 이상의 IT자산을 보유하고 있으며 25,000명 이상의 계열사 직원들에게 IT자산을 제공하고 유지보수를 주관하는 A사의 전산기기 장애접수 내용을 대상으로 하였다. A사는 10개 계열사를 대상으로 IT 서비스를 수행하며, 그중에서 가장 규모가 큰 계열사를 대상으로 실험을 수행하였다. 22종의 전산기기 중에서 장애접수 내용이 적어 기계학습이 불충분한 7종의 전산기기를 제외하고, 총 15종의 전산기기를 대상으로 6개월간 접수된 약 33,000건의 장애 접수 및 처리 데이터를 대상으로 기계학습을 수행하였다. 각 전산기기별로 비표준화된 장애처리 텍스트 데이터를 텍스트마이닝 기법을 이용하여 키워드의 출현 빈도수를 반영하였으며 전처리 작업 후 표준화된 장애처리 유형으로 분류하였다. 15종의 전산기기를 대상으로 기계학습 실험결과 장애처리의 유형은 101건으로 분류되었으며 발생 빈도가 높은 상위 30건은 Table 5와 같다.

이렇게 비정형적인 장애처리 유형을 기계학습에 의한 전처리 작업을 수행하여 최다 출현한 키워드를 기준으로 정형적인 장애처리 유형을 추출하였다. 고객이 요청한 장애접수 내용은 기계학습을 위하여 키워드별로 분류하였으며 알고리즘은 Table 6과 같다.

Table 6. Keyword Classification Algorithm

```
void main()
   {
    FILE *fp1;
    FILE *fp2;

    char buff[500];
    char result[500] = " ";
    int i,j,k,count;

    fp1=fopen("input.txt", "r");
    fp2=fopen("output.txt", "w");

    if(fp1 != NULL){
      while(!feof(fp1)){
            fgets(buff,500,fp1);
            count=strlen(buff);
            j=0;
            for(i=4; i<count; i++){
```

```
                if(buff[i]!=' ' && buff[i]!='\n' &&
                buff[i]!=',' &&
                buff[i]!='/' && buff[i]!='-'){
                result[j] = buff[i];
                j=j+1;
                }
            else{
            for(k=0; k<4; k++) fputc(buff[k], fp2);
            fputs("\t", fp2);
            for(k=0; k<j; k++) fputc(result[k], fp2);
            fputs("\n", fp2);
            j=0;
            }
            }
    }
  }
  fclose(fp1);
  fclose(fp2);
 }
```

키워드 분류 알고리즘은 C언어로 작성하였으며 입력 및 출력 결과의 예시는 Table 7과 같다.

Table 7. I/O results of Keyword Classification Algorithm

```
< Input Data >

A05     PDA 포트 및 연결상태 확인하라고 나옴.
A11     PDA마스터수신은 되는데 송수신안됨

중략

O02     핀패드인식안됨/IC카드인식안됨
O03     팬패드가 누르면 기계에는 전송완료라고 나옴

< Output Data >

A05     PDA
A05     포트
A05     및
A05     연결상태
A05     확인하라고
A05     나옴
A11     PDA마스터수신은
A11     되는데
A11     송수신안됨

Omitted

O02     핀패드인식안됨
O02     IC카드인식안됨
O03     팬패드가
O03     누르면
O03     기계에는
O03     전송완료라고
O03     나옴
```

Table 8. The number of Fault Recovery Types in IT Asset

IT Asset	Code	Number of Fault Recovery Type
Barcode Scanner	A	17
BPR Scanner	B	10
LCD Monitor	C	6
PC Attachments	D	13
PDA	E	5
Finance IC Issuing Device	F	8
Notebook PC	G	44
Desktop PC	H	51
Laser Printer	I	15
Inkjet Printer	J	11
Operator Reader Unit	K	8
Fingerprint Identification Devices	L	9
Card Issuing Device	M	15
Bankbook Printer	N	15
PIN Pad	O	9

총 15종의 전산기기 코드는 A부터 O까지 분류하였고 장애처리 유형을 구분하기 위한 2자리의 숫자코드를 부여하였으며 Table 8과 같다.

고객에 의하여 접수된 장애처리 요청건의 키워드별 발생빈도수를 6개월 간 축적하며 기계학습을 수행한 이후, 신규로 접수된 장애 건에 대하여 키워드를 분석하였다. 장애처리담당자가 장애접수 내용을 보고 장애처리 유형을 3가지 한도 내에서 예측한 경우는 성공률이 평균 56%로 조사되었다. 본 연구에서는 신규로 발생한 장애접수 건을 랜덤하게 100건을 추출하고 3회로 나누어 실험하였다. 장애에 대한 기계학습 예측 결과 평균 72건이 성공하였다. 실패한 경우 핵심키워드를 추가적으로 입력하여 키워드별 가중치를 조정하고 2차 기계학습을 수행하고 다시 실험한 결과 성공률이 평균 81건으로 향상되었으며 실험결과는 Fig. 7과 같다.

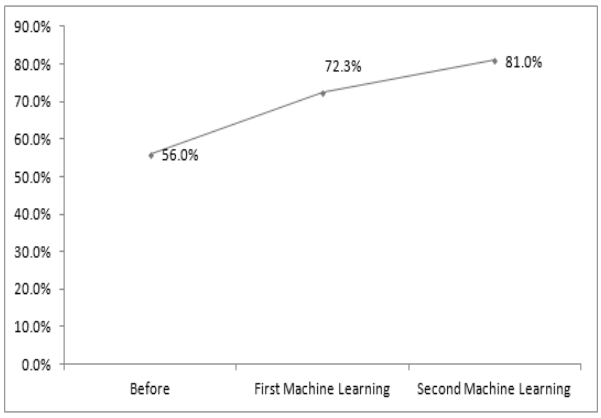

Fig. 7. Test Results

5. 결론 및 향후 연구

IT자산은 조직의 경영목적을 지원해주는 핵심영역이며, IT자산의 장애 발생 시 신속한 처리를 지원하는 것은 매우 중요하다. 본 연구에서는 6개월간 접수된 약 33,000건의 장애 접수 데이터를 대상으로 하였으며 전산기기 장애접수 및 처리현황 데이터를 활용하였다.

장애접수 된 문장을 키워드로 분류하고 같은 의미를 다르게 표현한 키워드를 표준화하였으며 키워드별 빈도수를 산출하였다. 키워드별 빈도수는 기계학습 생성규칙에 의하여 한 문장의 여러 개의 키워드별로 예측된 장애처리 유형을 생성하였다. 생성된 유형간의 최대 발생 건을 기준으로 예측의 우선순위를 제공하였으며, 예측이 실패하더라도 추가적인 학습 및 유형간의 신뢰도를 바탕으로 기계학습을 강화시켰다. 제안한 기계학습 프로세스의 유용성을 검증하기 위하여 신규로 접수된 장애접수 데이터를 랜덤하게 100건을 추출하고 3회로 나누어 실험하였다. 키워드 분석에 의한 매핑으로 장애처리 유형이 예측 가능하였다. 실험한 결과 장애처리 예측의 적중률은 약 72.3%였으며, 지속적인 기계학습을 통하여 81%로 향상되었다. 본 연구의 한계점은 모든 IT자산을 대상으로 하기에는 장애발생 건수가 기계학습 프로세스를 적용하기에는 부족하였다. 그리고 제안하는 기계학습 방법은 전체를 자동화하여 시스템으로 완성하기에는 시간, 인력, 예산 등의 자원을 조달하기에는 한계가 있어서 각 단계별로 저자가 제안하는 기계학습과 관련된 부분만 알고리즘으로 작성하고 프로그램을 부분적으로 구동하였으며, 논문에서 제안하는 것을 입증할 수 있었다.

향후 연구과제로 IT자산의 장애뿐만 아니라 기계학습과 연관성 규칙을 더욱 확대하여 IT자산의 적정한 도입 수량 및 교체주기 예측, 장애와 연관된 BMT(Bench Marking Test) 항목 도출 등 연구범위의 확대가 필요하다. 또한 기계학습을 지원하는 시스템을 위한 자동화 연구가 지속되어야 하며, 나아가 조직의 KMS(Knowledge Management System)와 연동하여 장애가 나면 신속히 복구하기 위한 방법을 알려주는 자동화 도구 설계가 필요하다.

참 고 문 헌

[1] Penny Grubb, Armstrong A Takang, “Software Maintenance Concepts and Practice”, World Scientific, 2003.

[2] Daryl Mather, “The Maintenance Scorecard”, Industrial Press, 2005.

[3] Matthew B. Doar, “Practical Development Environments”, O'Reilly & Associates, 2005.

[4] Roger S. Pressman, “Software Engineering A Practitioner's Approach”, McGraw-Hill, 2005.

[5] United Kingdoms Office Of Government Commerce, Information Technology Infrastructure Library Ver3 - Service Transition, The Stationery Office, 2008.

[6] IEEE, IEEE Std 1220-2008 - Systems and software engineering — Software life cycle processes, IEEE Computer Society, 2008.

[7] Y. K. Cho, Y. S. Hong, J. S. Lee, Algorithm, Ehan IT Books, 2006.

[8] K. Y. Lee, J. W. Kim, Algorithm, KNOU PRESS, 2012.

[9] K. H. Lee, B. R. Lee, Artificial Intelligence, KNOU PRESS, 2011.

[10] S. J. Kim, “The Machine Learning of making and learning”, Hanbit Media, 2012.

[11] Young-Joon Moon, Sung-Yul Rhew, “A Study on Software Fault Analysis and Management Method using Defect Tracking System”, KIPS Transactions on Part D, pp.321-326 1598-2866, 2008.

[12] Hyo-Young Kim, Hyuk-Soo Han, “A Defect Prevention Model based on SW-FMEA”, Journal of KISS, pp.605-614 1738-6322, 2006.

[13] Yun-Soo Choi, Sa-Kwang Song, Hong-Woo Chun, Chang-Hoo Jeong, Sung-Pil Choi, “Terminology Recognition System based on Machine Learning for Scientific Document Analysis”, KIPS Transactions on Part D, pp.321-326 1598-2866, 2011.

[14] Hyung-Rim Choi, Kwang-Ryel Ryu, Jae-Ho Kang, Jong-Il Shin, Chang-Sup Lee, “An Automatic Question Routing System using Machine Learning”, Spring Conference of KIISS, 2003.

문 영 준

e-mail : yjmoon@yonsei.ac.kr

2004년 한국방송통신대학교 컴퓨터과학과 (학사)

2006년 연세대학교 컴퓨터공학과(석사)

2008년 숭실대학교 컴퓨터학과(박사수료)

2006년~현 재 한국방송통신대학교 Tutor 및 출석수업 교수

관심분야 : 유지보수, IT자산관리, 기계학습, 형상관리, 정보보호

류 성 열

e-mail : syrhew@ssu.ac.kr

1976년 숭실대학교 컴퓨터학과(학사)

1980년 연세대학교 컴퓨터공학과(석사)

1997년 아주대학교 컴퓨터공학과(박사)

1981년~현 재 숭실대학교 컴퓨터학부 교수

1982년~1995년 숭실대학교 전자계산원장 및 중앙전자계산소장

1998년~2001년 숭실대학교 정보과학대학원장

2006년~2010년 공정거래위원회, 기획재정부, 보건복지부 정보화 위원

관심분야 : 요구공학, 유지보수, 오픈소스, IT정책경영학

최 일 우

e-mail : iwchoi@kangnam.ac.kr

1995년 숭실대학교 컴퓨터학과(학사)

1997년 숭실대학교 컴퓨터학과(석사)

2004년 숭실대학교 컴퓨터학과(박사)

2007년~현 재 강남대학교 교양학부 교수

관심분야 : 개발 프로세스, 레거시 재사용, USN, 모바일컴퓨팅, 임베디드 시스템