KIPS Tr. Software and Data Eng.
Vol.5, No.3 pp.145~154 pISSN: 2287-5905



http://dx.doi.org/10.3745/KTSDE.2016.5.3.145

# TF-IDF Based Association Rule Analysis System for Medical Data

Hosik Park[†] · Minsu Lee[††] · Sungjin Hwang[†††] · Sangyoon Oh[††††]

## ABSTRACT

Because of the recent interest in the u-Health and development of IT technology, a need of utilizing a medical information data has been increased. Among previous studies that utilize various data mining algorithms for processing medical information data, there are studies of association rule analysis. In the studies, an association between the symptoms with specified diseases is the target to discover, however, infrequent terms which can be important information for a disease diagnosis are not considered in most cases. In this paper, we proposed a new association rule mining system considering the importance of each term using TF-IDF weight to consider infrequent but important items. In addition, the proposed system can predict candidate diagnoses from medical text records using term similarity analysis based on medical ontology.

Keywords : Association Rule, Medical Data, FP-Growth, TF-IDF

# 의료 정보 추출을 위한 TF-IDF 기반의 연관규칙 분석 시스템

박 호 식[†] · 이 민 수[††] · 황 성 진[†††] · 오 상 윤[††††]

## 요 약

u-Health에 대한 관심과 IT 기술의 발전에 따라 의료 정보를 적극적으로 활용하고자 하는 요구가 커지고 있으며, 이에 대해 텍스트 형태의 의료 정보 데이터에 연관규칙 기법을 적용하여 질병과 증상과의 관계를 추론하는 시스템에 대한 연구들이 이루어지고 있다. 그러나 일반적인 연관규칙 기법을 의료 정보 데이터에 그대로 적용할 경우, 이전에는 새로운 연관규칙들보다 일반적이며 의미없는 연관규칙들이 많이 생성되는 문제가 발생한다. 또한 필터링으로 인해 빈번하게 함께 발생하지는 않지만 의학적으로 의미있는 항목들의 연관 규칙을 발견할 수 없다는 한계점을 가지게 된다. 본 논문에서는 의료데이터 특성을 고려하여 빈번한 항목과 빈번하지 않지만 의학적으로 의미 있는 항목들을 대상으로 연관규칙을 구성하여 의료 전문가의 의사 결정에 도움을 주기 위한 시스템을 제안한다. 제안 시스템은 의료 기록 데이터에서 용어들을 TF-IDF기반으로 가중치를 부여하고 기존 FP-Growth 알고리즘을 확장하여 TF-IDF 가중치를 고려한 빈번하게 발생하거나 빈번하지 않지만 의미 있는 연관규칙을 구성한다. 특정 질의 데이터가 입력되면 해당 데이터에 나타난 연관 규칙들의 유사도를 의학분야 온톨로지를 이용하여 평가하여 해당 데이터의 내용과 관련된 후보 질병들을 추론한다. 추론된 후보 질병명은 의료 전문가에게 의사 결정의 참고 자료로 제공된다. 실제 임상 진료 및 처방 기록 데이터에 대해 제안 시스템을 적용해 본 결과, 본 제안 시스템을 통해 도출한 연관 규칙이 기존 FP-Growth 알고리즘을 적용했을 때 보다 더 구체적인 질병과 증상과의 관계들을 포함함을 확인할 수 있었다. 또한 본 제안 시스템은 자유형식의 의료 및 병리데이터를 마이닝하고 후보 질병들을 가중치 기반으로 보여주므로, 의료 기록 정보로부터 질병 관련 새로운 정보를 획득하고 의료진의 의사 결정에 도움을 주는 시스템으로 활용될 수 있다.

키워드 : 연관규칙, 의료 데이터, FP-Growth, TF-IDF

## 1. 서 론

최근 IT기술의 발전과 u-Health의 관심으로 인해 다양한 첨단 기기 이용하여 의료 정보 데이터를 수집하고 분석 및 활용할 수 있는 환경이 조성되고 있다. 다양한 스마트 밴드 및 웨어러블 디바이스(Wearable device)들을 이용하여 연속적으로 생체 정보들을 수집하고, 유전체 분석 기술과 질병 진단 및 예측 키트 등을 통해 개인화된 의료 정보들을 얻을 수 있다[1, 2]. 이렇게 수집, 분석된 의료 정보 데이터를 효과적으로 관리하고 활용할 수 있도록 많은 의료 정보들은 처리 및 분석이 용이한 형태로 전산화 되어 있으며, 이러한 정보들을 바탕으로 유용한 정보들을 추출하여 의료진의 의사결정에 도움을 주는 시스템들에 대한 관심이 높아지고 있다.

그러나 진료와 처방 내역의 기록, 병리검사 결과 또는 멀티미디어 데이터 판독 기록 등의 텍스트 데이터들은 증상과 질병, 예후 등에 대한 중요한 정보를 의학적 전문용어를 사

※ 본 논문은 2015년도 정부(미래창조과학부)재원으로 지원된 한국연구재단 신진연구자지원사업(2015R1C1A1A01054305), 정부(교육부)재원으로 지원된 한국연구재단 기초연구지원사업(NRF-2015R1D1A1A01059557), 그리고 미래창조과학부 및 정보통신기술진흥센터의 ICT/SW창의연구과정(SW중심대학)지원사업의 연구결과로 수행되었음(R2215-15-1002).
† 준 회 원 : 아주대학교 컴퓨터공학과 석사과정
†† 종신회원 : 이화여자대학교 컴퓨터공학과 연구교수
††† 비 회 원 : (주)휴민텍 의료영상사업부 팀장
†††† 정 회 원 : 아주대학교 소프트웨어학과 부교수
Manuscript Received : December 14, 2015
First Revision : February 1, 2016
Accepted : February 2, 2016
* Corresponding Author : Sangyoon Oh(syoh@ajou.ac.kr)

용한 자연어 형태로 표현하고 있어 의미있는 의료 정보 추출을 위해서는 의료분야의 특수성을 고려한 추가적인 정보처리 과정이 요구된다[3]. 또한 의료 분야는 사람의 생명을 다루는 분야이므로 매우 정확한 판단 및 적용이 요구되는 특수 분야이므로 데이터 분석 및 마이닝 결과를 바로 의사결정으로 적용하기보다 의료 전문가의 의사 결정에 참고사항으로 활용되어야 한다[4].

텍스트 형태의 의료 정보 데이터를 적극적으로 활용하여 유용한 정보를 얻기 위한 의료 정보 시스템들에 대한 연구들도 활발히 이루어지고 있다. 텍스트 마이닝 기법을 이용하여 의학 문헌 정보에서 유전자와 질병명의 동시 발생 빈도에 기반한 유전자-질병간의 관계를 규명하거나[5], 유전자의 상태 변화에 따른 암과의 연관성을 발견하는 연구[6] 등이 수행되었다. 더 나아가 텍스트 마이닝 기법을 의료 정보 검색에 활용하여 유전자와 암과의 관계를 검색하는 시스템들도 제안되었다[7, 8].

또 한편으로 의료데이터에 데이터 마이닝의 연관규칙 기법을 적용하여 질병과 증상의 관계를 추론하는 시스템에 대한 연구들이 이루어지고 있다. Alghamidi[9]는 FP-Growth 알고리즘을 이용하여 의료데이터의 연관규칙을 추출하였다. 이는 의료데이터의 각 속성(Attribute)의 관계에 대해 추출하고, 각 연관규칙에 대해 규칙 트리를 구성하여 의사결정에 도움이 되는 시스템을 구축하였다. Yang[10]은 방약합편이라는 책으로부터 질병의 증상과 약초와의 관계를 Apriori 알고리즘을 이용하여 추출하여 네트워크 분석을 하였다.

그러나 일반적인 연관규칙 기법을 활용하는 기존 시스템은 새로운 연관규칙들보다 대부분의 데이터에서 함께 빈번하게(Frequent) 발생하는 항목에 대한 의미 없는 연관규칙들이 많이 포함되게 된다. 또한, 후보 탐색 범위를 줄이기 위해 발생 빈도가 낮은 용어들은 사전에 필터링되어 빈번하게 함께 발생하지는 않지만(Infrequent) 의미있는 항목들의 연관 규칙을 발견할 수 없다는 한계점을 가지고 있다. 의료용어 및 분야의 특성상 빈번하지 않은 용어들이 특정 질병을 특징지을 수 있는 중요한 용어로 활용될 수 있으므로, 질병과 관련된 용어들에 대한 연관규칙을 구성할 때에는 단순한 발생 빈도와 함께 용어의 중요도를 고려하여야 중요한 질병과 증상 관련 정보들을 놓치지 않을 수 있을 것이다.

본 논문에서는 의료데이터 특성상 빈번한 항목과 빈번하지 않지만 의미를 가질 수 있는 항목들을 대상으로 연관규칙을 구성하여 의료 전문가의 의사 결정에 도움을 주기 위한 시스템을 제안한다. 제안 시스템은 의료 소견데이터에서 용어들을 TF-IDF 기반으로 가중치를 부여하고, 기존 FP-Growth 알고리즘을 확장하여 TF-IDF 가중치를 고려한 빈번하게 발생하거나 빈번하지 않지만 의미있는 연관규칙을 구성한다. 특정 질의 데이터가 입력되면 해당 데이터에 나타난 연관규칙들안의 병명들의 유사도를 의학분야 온톨로지를 기반으로 평가하여 해당 데이터의 내용과 관련된 후보 질병들을 추론한다. 추론된 후보 질병명은 의료 전문가에게 의사 결정의 참고 자료로 시각적으로 제공된다.

본 논문의 기여도는 다음과 같다. 1) 의학 분야의 특수성을 고려하여 빈번한 항목과 함께, 빈번하지 않지만 각 문서에서의 중요도가 높은 항목을 함께 고려한 연관규칙 분석 시스템을 제안하였다. 2) 기존의 PubMed의 초록과 같은 정형화된 문서 기반 의료정보 마이닝 시스템이 아니라 병원에서 의료진들이 기록 목적으로 비구조적(unstructured) 혹은 반구조적(semi-structured)으로 자유롭게 정리한 의료 기록물로부터 병리데이터를 마이닝 하는 시스템을 제안하였다.

본 논문의 구성은 다음과 같다. 2장에서는 전통적인 연관규칙 분석 방법과 의료 데이터를 이용한 기존 연관규칙 분석 방법에 대해 소개한다. 3장에서는 제안하는 시스템 구조와 연관규칙 추출 방법에 대해 설명한다. 4장에서는 제안 시스템을 활용하여 실제 병리 데이터를 이용하여 병명 추론 및 기존 연관규칙과 비교 분석하였으며, 5장에서는 결론 및 향후 연구 과제에 대해 제시하였다.

## 2. 관련 연구

### 2.1 전통적인 연관규칙 분석 방법과 알고리즘

연관규칙 분석은 주어진 트랜잭션 집합으로부터 어떤 아이템이 나타날지를 다른 아이템의 발생으로부터 예측하는 규칙을 찾는 작업을 말한다. 데이터베이스에 총 n개의 트랜잭션 데이터와 m개의 항목으로 구성된 집합을 I라고 할 때, 연관규칙 R은 “R:X⇒Y와 같이 표현된다. 연관규칙을 추출하는 것은 항목집합 X와 Y를 선택하는 문제인데, 지지도(support)와 신뢰도(confidence)를 사용한다. 지지도는 관심있을 정도로 빈발하게 나타나는 항목을 고려하는 값으로 X와 Y를 동시에 포함하는 트랜잭션 수의 비율을 말하며, Equation (1)과 같이 표현된다.

$$\text{Support(R)} = \frac{P(X \cup Y)}{T} \quad (1)$$

신뢰도는 규칙의 강도를 나타낸 것으로 X가 발생할 때 Y도 동시에 발생하는 조건부 확률을 의미한다. 트랜잭션 X의 항목들을 포함하는 경우 Y의 항목들도 동시에 포함할 확률을 나타내며 신뢰도가 높은 규칙일수록 의미가 크다고 할 수 있다. Equation (2)와 같이 표현된다.

$$\text{Confidence(R)} = P(Y|X) \quad (2)$$

지지도와 신뢰도를 통해 연관규칙을 생성한다. 대표적인 알고리즘으로 Apriori[11] 알고리즘이 있다. 이는 항목들의 지지도를 구해 최소 지지도를 넘는 빈발항목을 추출하여 이를 바탕으로 후보항목집합을 구성한다. 이들 중에서 다시 최소 지지도를 만족하는 항목들을 추출하는 과정을 반복하여 빈발항목집합을 구한다. Apriori는 연관성을 갖는 항목들을 발견하는데 초점을 두어 각 패스마다 전체 데이터 셋을 검색하는데, 후보가 될 가능성이 없는 집합들에 대해서도

고려하기 때문에 시간과 비용이 많이 든다.

후보항목집합을 생성하지 않고, FP-Growth[12]를 이용하여 각 항목에 대해 조건부 패턴트리를 생성하여 연관규칙을 찾는 FP-tree를 이용한 기법들이 최근 많이 사용되고 있다. Apriori처럼 반복적인 데이터베이스 접근을 필요로 하지 않으며, 트랜잭션 데이터 내에 존재하지 않는 항목집합 역시 생성하지 않아 Apriori보다 속도와 비용면에서 크게 향상되었다.

전통적인 연관규칙의 경우 데이터베이스의 각 트랜잭션을 구성하는 항목들은 같은 특성의 항목으로 구성되어 있다고 가정한다. 하지만 다른 항목에 비해 임의의 항목이 더 중요하다면, 전통적인 연관규칙을 적용하기에 어려움이 있다. 따라서 각 항목에 대해 상대적인 가중치를 부여하여 연관규칙을 적용하는 연구가 많이 진행되고 있다. Yanbo[13]에서는 문서를 분류하는데 있어서 특정 관심 있는 항목에 대해 가중치를 부과하여 연관규칙을 추출하고 있다.

### 2.2 의료분야에서 연관규칙 분석을 활용한 기존 시스템

Lee[14]은 연관규칙 분석을 통해 심근경색을 일으키는 원인을 발견하는 방법에 대해 제안하였다. 특정 타겟 패턴을 미리 정의해 두고, CTP-Tree(Complete Target Pattern tree)라는 제안구조에 따라 병력(Medical history)의 관계에 대해 추론한다. 제안구조는 Fig. 1과 같다. Fig. 1의 과정은 빈발항목으로부터 패턴을 정의한다. 즉, {smoking, glucose} → {diabetes}, {smoking, total cholesterol} → {diabetes, hypertension}과 같은 패턴을 추출한다. 즉, 선행사건과 결과의 상관관계를 분석하여 정확도를 높였다.

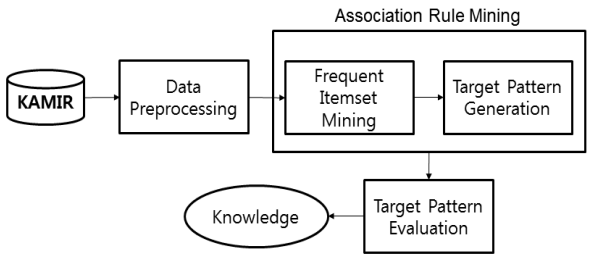


Fig. 1. Target Pattern reasoning through the association rule method[14]

또한, 빈발 항목패턴을 통해 심근경색을 일으키는 원인에 대해 추론가능하다. 하지만 이 연구는 특정 타겟을 정해 놓고 연관관계를 분석함으로 일반적인 상황에 적용하기 제한점이 있다. 또한 빈번하지 않은 항목에 대해 고려하지 않고 있기 때문에 특이한 증상은 연관규칙 생성에 포함되지 않는다.

Sajid[15]는 Apriori를 이용하여 빈번한 항목에 대해 Positive Association Rules(PARs), 빈번하지 않은 항목에 대해 Negative Association Rules(NARs)라 정의하여 증상과 병명간의 관계를 보였다. Sajid는 Positive와 Negative한 연관규칙들의 중요도가 같다고 보며, 지지도, 신뢰도, 향상도를 고려하여 Positive와 Negative의 기준을 나눈다. 또한, IDF 가중치를 이용하여 전체 문서 중에서 60%만 사용하였다. 이 연구에서는 Apriori를 이용하여 연관규칙을 생성하는데 후보집합을 생성하기 때문에 FP-Growth보다 많은 시간이 소모가 되는 단점이 있다. 또한, 질병간의 관계에 대해서 트레이닝 데이터에 의존하기 때문에 의미적으로 맞는지 검증이 필요하다.

## 3. 연관규칙 기반의 병리데이터 병명 추론 시스템

본 장에서는 의료소견 데이터를 이용한 제안하는 병명 추론 기법의 시스템 모델을 설명한다. 전처리 단계에서는 비정형 의료소견 데이터를 연관규칙 추출에 필요한 데이터를 추출하며, 연관규칙 추출 단계에서는 빈번하지 않은 키워드에 대해 가중치를 부가하여 연관규칙을 구성한다. 마지막으로 테스트 단계에서는 테스트할 의료 소견데이터가 들어오면, 정해진 연관규칙에 패턴 매칭하여 병명을 추론한다. 본 논문에 제안하는 시스템 구조는 Fig. 2와 같다.

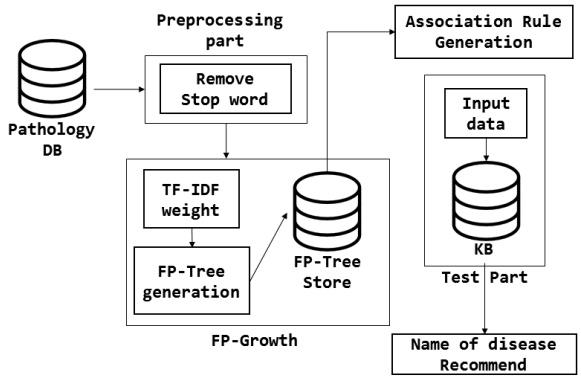


Fig. 2. Proposed structure for the name of disease reasoning

Table 1. MIMIC2 Example

| |
|---|
| DATE: [**2923-3-3**] 9:16 AM<br>BABYGRAM<br>Reason: check heart size, pulmonary markings<br><br>History of Present Illness:<br>Mr. [**Known patient lastname 3138**] is a 68-year-old male with worsening symptoms of<br>dyspnea on exertion and chest tightness, who has been followed with a known aortic stenosis.<br><br>Past Medical History:<br>Osteoarthritis<br>Basal Cell Cancer<br>Right hip replacement<br><br>Social History:<br>Lives with wife. Owns a retail store. Smoked for 25 years quitting 25 years ago.<br><br>Discharge Diagnosis:<br>[**3460-5-28**] - Status post Aortic Valve Replacement (#21 [**Last Name (un) 163**]<br>[**Doctor Last Name 164**] pericardial) |

### 3.1 전처리 단계

비정형 데이터인 의료소견 데이터를 정형화하기 위해, 전처리 단계를 거친다. 본 논문에서 사용된 의료소견 데이터는 MIMIC2[16]에서 제공받은 데이터를 이용한다. 이는 PhysioNet에서 연구를 목적으로 미국 국립 보건원의 후원하에 만들어진 임상 데이터베이스이다[17].

Table 1은 MIMIC2 데이터의 예시이다. 데이터 필드는 의료 전문가에 의한 진단과 병력, 환자의 상태 등이 있다. 본 전처리 단계에서는 영문에 경우 자주 발생하는 불필요한 am, are, is 등과 같은 불용어(stop word)를 제거한다. 이를 통해, 시스템이 봐야할 데이터를 대폭 감소할 수 있다.

### 3.2 연관규칙 추출 단계

기존 시스템에서는 FP-Growth를 이용한 연관규칙 분석을 적용하기 위해, 특정 질병을 정해두고 해당하는 증상들의 관계를 나타낸다. 본 연구에서는 질병과 증상 사이의 연관관계를 발견하는 것을 목적으로 하기 때문에, 증상끼리의 연관규칙을 배제한다. 또한, 기존 FP-Growth를 이용한 연관규칙 생성 기법은 단순히 빈번한 항목에 대하여 연관규칙을 생성하기 때문에 앞서 설명했듯이 의료데이터 특성에 맞지 않다. 즉, 증상들의 관계를 통해 질병을 추론하기 때문에 실생활에서 맞지 않다.

본 논문에서는 의료데이터 특성상 해당 질병에 끼치는 증상의 가중치를 측정하여 연관규칙에 적용하는 방법을 제안한다. 이를 통해, 기존에 지지도와 신뢰도 값에 의해 무시되었던 빈번하지 않은 항목을 연관규칙 생성에 포함시킬 수 있다.

1) 가중치 측정

본 논문에서는 가중치 측정을 위해 TF-IDF(Term Frequency-Inverse Document Frequency) 방법을 사용한다. TF-IDF는 문서내에서 특정 단어의 빈도를 전체 문서군의 단어 출현 빈도로 나눈 값이다. TF-IDF 값을 통해 특정 단어가 문서 내에서 얼마나 중요한지를 나타낸 통계적 수치이다. Equation (3)과 같이 계산할 수 있다.

$$TF-IDF_{w,d} = f_{w,d} \cdot \log(\frac{|D|}{f_{w,D}}) \quad (3)$$

$w$는 특정 단어, $d$는 특정 문서, $D$는 전체 문서를 의미한다. $f_{w,D}$는 $w$가 등장한 문서의 횟수를 $|D|$ 는 전체 문서의 수를 의미 한다. DF는 해당 단어가 나타난 문서의 수를 의미하는데, 전체의 문서군에서 몇 개의 문서에서 나타났는지를 의미한다. 만약, DF 값이 높은 단어는 많은 문서에서 나타난 것이므로 중요한 단어가 아니다. 따라서 DF 값에 역수를 취해 IDF 값을 사용하여 해당 단어의 중요도를 나타낸다. IDF 값에 따라서 특정 단어가 다수의 문서에 등장하면 가중치가 감소하고, 소수의 문서에 등장하면 증가하게 된다. 따라서 IDF 값에 의해 빈번하지 않은 항목도 고려하여 연관규칙을 생성할 수 있다.

Table 2는 각 단어에 대해 TF, IDF, TF*IDF 값이 각각 표시되어 있다. 기존 방법을 사용하여 연관규칙을 추출한다면 TF 값으로만 구성되기 때문에 'Patient'와 같은 거의 모든 문서에서 나타나는 전체 문서에서 중요하지 않은 단어가 포함될 가능성이 높다. 기존 알고리즘은 최소 지지도 $\alpha$보다 빈번한 항목에 대해서 FP-Tree를 구성한 후 연관규칙 $\beta$를 생성한다. 빈발항목으로 정렬되어 있는 헤더테이블을 기준으로 최소 지지도 $\alpha$보다 큰 모든 항목에 대해 연관규칙을 만든다.

Table 2. Importance of symptoms corresponding to each weight

| Term | TF(Rank) | IDF | TF*IDF(Rank) |
|---|---|---|---|
| Patient | 4735 (1) | 1.128E-6 | 0.00534 (9332) |
| Pain | 3069 (2) | 1.324E-6 | 0.00406 (9666) |
| Year | 2528 (3) | 0.723E-6 | 0.00182 (12544) |
| Left | 2523 (4) | 1.625E-6 | 0.00176 (12550) |
| Day | 2413 (5) | 0.730E-6 | 0.00176 (12550) |
| Status | 2269 (6) | 1.624E-6 | 0.00369 (9717) |
| History | 2082 (7) | 1.391E-6 | 0.00289 (11939) |

본 논문에서 제안하는 가중치 측정 기법은 앞서 설명한 TF-IDF 가중치를 기반으로 FP-Tree를 구성한다. 즉, 알고리즘1과 같이 단순 빈번한 항목에 대해 구성하는 것이 아니라, 중요도에 따라 FP-Tree를 구성한다. TF는 기존 방법과 같은 빈도수를 나타내고, IDF는 해당 단어가 나타난 문서의 수의 역수를 취해 줌으로써 전체 문서군에서 중요도를 나타낸다. 본 논문에서 TF는 전체 데이터의 단어 수와 해당 단어의 빈도수로 나눠 정규화 한다. 정규화를 해주지 않게 되면, TF의 범위가 매우 넓어지기 때문에 문서의 전체 단어대비 단어의 빈도수로 Equation (4)와 같이 정규화를 한다. T는 전체 단어를 나타내고, $T_c$는 해당 단어의 빈도수를 나타내어 해당 단어의 빈도수를 전체 단어로 나누어 줌으로써 정규화를 한다.

$$Normalize(t_i) = \frac{T_c}{T} \quad (4)$$

사용되는 데이터는 진단서로써 증상, 수술명, 진단명, 병력 등이 포함되어 있다. 3만 여개의 데이터로 구성되어 있다. 이 데이터를 FP-Tree로 구성하는데 트레이닝 데이터로 사용한다. 제안된 TF-IDF 방법으로 가중치를 줌으로써 기존 방법에 의해 무시되던 항목들 즉, 임계값 아래의 항목들이 본 논문에서 제안하는 방법을 통해 연관규칙 생성에 포함된다.

본 논문에서는 Table 3과 같이 전체 트렌젝션 데이터에서 나타나는 단어들을 TF-IDF를 계산하여 높은 순으로 정렬한다. 다음으로 트렌젝션 데이터를 TF-IDF가 높은 단어의 순으로 정렬하여 FP-Tree를 구성한다.

Table 3. Importance of symptoms of the TF-IDF weight

| Symptom | TF(Rank) | Symptom | TF-IDF(Rank) |
|---|---|---|---|
| Patient | 4735(1) | Hyperlipemia | 11.658(1) |
| Pain | 3069(2) | Abciximab | 10.564(2) |
| Year | 2528(3) | Nonketotic Hyperglycinemia | 10.323(3) |
| Left | 2523(4) | Hypomagnesia | 9.824(4) |
| Day | 2413(5) | Ileocecostomy | 9.771(5) |

알고리즘 1은 본 논문에서 제안하는 방법이다. 먼저, 트렌젝션 데이터에서 TF-IDF 값을 구한다. 이를 통해, 트렌젝션 데이터를 TF-IDF가 높은 순서로 정렬한다. 다음은 이 TF-IDF 값을 FP-Growth의 헤더테이블로 사용해서 임계값 $\gamma$보다 작은 데이터는 제거한다. FP-Tree는 트렌젝션 데이터를 읽으면서 최소 지지도 $\alpha$보다 큰 항목들을 순차적으로 트리에 저장한다.

**Algorithm 1. Adjust weight for infreqeunt item**

INPUT : Pathology data

```
1   First Scan transaction database. collect the set of items F
2   Calculate TF*IDF weight. and Sort F by TF*IDF weight
3   Create initial header table by TF*IDF weight
4   For each instance in instances length
5       Split it into items
6       Construct header table (Item, Count)
7       Delete items less than γ
8
9   Second Scan transaction database.
10          create initial FP-Tree and link table
11
12  IF FP-Tree contains a single path P then
13      FOR each combination do generate pattern β, α with
14              support = minimum support of nodes in β
15  ELSE
16      FOR each header ai in the header of Tree
17      DO
18          Generate pattern β = ai α with support = ai.support
19          Construct β.s conditional pattern base
20  and β.s conditional FP-tree Tree β
```

OUTPUT : Complete set of frequent pattern

2) 연관규칙 생성

본 논문에서 제안하는 연관규칙 생성은 알고리즘2를 기반으로 FP-Growth를 변형하여 FP-Tree를 생성한 후 연관규칙을 추출한다. FP-Growth는 Apriori 알고리즘의 단점인 후보 집합을 생성하지 않는 연관규칙 추출기법이다. 또한, 후보 집합을 생성하지 않고 연관규칙을 추출하기 때문에 반복적으로 데이터베이스 접근을 하지 않으며, 트랜잭션 데이터 내에 존재하지 않는 항목집합 역시 생성하지 않는다. 본 논문에서는 FP-Tree를 생성 후 이로부터 연관규칙을 추출한다.

Table 4. Transaction DB example

| TID | Symptom |
|---|---|
| 1 | {f, c, a, m, p} |
| 2 | {c, b, p} |
| 3 | {f, c, a, b, m} |
| 4 | {f, b} |

Table 4를 이용하여 Fig. 3과 같이 FP-Tree를 구성하게 된다. 본 논문에서 설명의 편의를 위해 알파벳으로 표현한다. 완성된 FP-Tree로부터 연관규칙을 추출하는 방법은 다음과 같다. 항목 b와 관련된 패턴은 Header Table의 b의 링크를 따라 각각의 부모 노드를 살펴보면 된다. 그 결과, {(b, a, c, f : 1), (b, f : 1), (b, c : 1)}가 연관규칙으로 추출되는 것으로 볼 수 있다. ":1"은 지지도를 나타낸다.

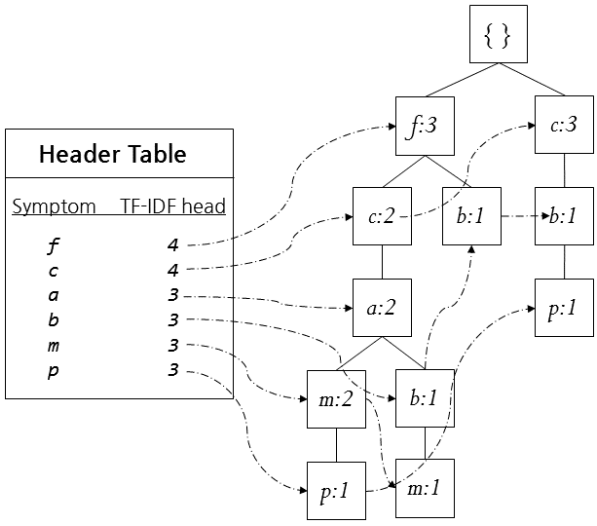


Fig. 3. FP-Tree

3.3 테스트 단계

Fig. 3과 같이 FP-Tree가 구성되어 있다면, a, b, m이 증상으로 테스트될 때, 연관관계는 $r_i$ = {(a, c, f), (b, a, c, f), (b, f), (b, c), (m, a, c, f), (m, b, a, c, f)}가 된다.

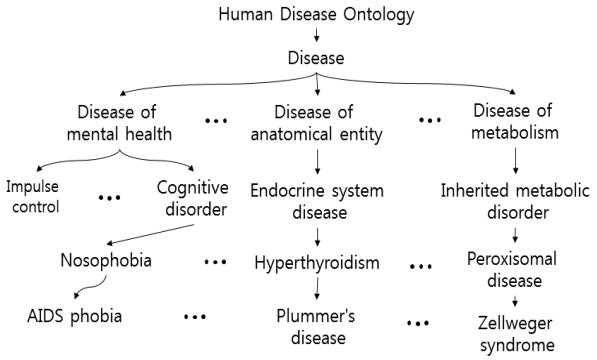


Fig. 4. OBO Foundry Ontology

c가 Acute rhinitis(급성 비염), f가 Acute pharyngitis(인두염)이라 할 때, a, b, m의 증상은 연관규칙에 의해 급성 비염 또는 인두염 질병을 가질 수 있다. 이는 연관규칙에 질병명이 포함되어 있는 연관규칙을 후보로 한다. 본 연구에서는 질병과 증상 사이의 연관관계를 발견하여, 테스트 데이터의 질병명을 추론하는 것을 목적으로 하기 때문에, 증상들 끼리 연관규칙이 추출되면, 의미가 없는 연관규칙이기 때문에 사전에 질병명 Dictionary를 정의해 질병명을 포함하는 연관규칙만 추출한다. 또한, 환자의 증상을 특정 질병으로 확진하는 시스템이 아니라, 의료 전문가에 도움을 주는 시스템이기 때문에 확진을 하게 되면, 위험성이 따른다. 따라서 본 논문에서는 OBO Foundry[18]에서 제공하는 의료 온톨로지를 이용하여, 위 과정을 통해 추출된 후보 질병들과의 관계를 보임으로써 의료 전문가 의사결정에 도움을 주는 표현 방법에 대해 제안한다.

Fig. 4와 같은 구조의 OBO Foundry Ontology는 의료 관련 온톨로지를 다루고 있는데, 본 논문에서는 사람 질병에 관한 온톨로지를 사용한다. Fig. 4는 OBO Foundry Ontology의 계층도를 나타내며, 각 질병에 대해 계층 구조에 따라 각 질병을 분류한다.

온톨로지의 Root 를 정하기 위하여, 각 후보들의 지지도(Support), 신뢰도(Confidence), 향상도(Lift) 값을 계산하여 결정한다. 지지도는 앞서 설명한 항목 A와 항목 B가 동시에 일어나는 확률을 의미한다. 신뢰도는 A가 발생한 경우 중 B가 발생하는 경우의 조건부확률을 의미한다. 향상도는 Equation (5)와 같이 신뢰확률을 기대 신뢰확률로 나눈 값으로 1에 가까우면 항목간에는 상관관계가 거의 없음을 뜻하며, 1보다 크면 양의 상관관계, 1보다 작으면 음의 상관관계를 뜻한다.

$$\text{Lift(R)} = \frac{P(Y|X)}{P(Y)} \tag{5}$$

따라서 본 논문에서는 이 값들이 특정 임계점 미만이면 후보 병명으로 채택될 수 없다. 각 후보들의 Root를 정하기 위하여 순위를 정해야 하는데, 본 논문에서는 향상도, 신뢰도, 지지도 순으로 순위를 정한다. 왜냐하면, 지지도의 경우 각 항목은 각 항목의 관심변수 모두의 비중이 크고, 연관성도 큰 경우에 유용하게 사용되지만, 관심변수 전체에 대한 비중이 낮은 경우에는 연관성을 판단하기 어려움이 있다. 따라서 지지도의 단점을 보완하는 것이 신뢰도인데, 신뢰도는 지지도와 달리 대칭적이지 않다. 하지만 신뢰도가 높은 연관규칙 중에는 우연하게 연관성이 높게 보이는 것들이 나타나기 때문에 이를 보완하기 위해 향상도가 사용되었다. 따라서 본 논문에서는 향상도가 가장 높은 후보를 Root로 사용한다.

각 후보들 간의 시각적으로 표현하기 위해 Root노드와 후보 노드들 간의 온톨로지 거리를 구하여 질병간의 얼마나 유사도가 있는지 표현한다. 거리는 Equation (6)과 같이 구하며, Equation (6)은 Resnik[19]에 의해 제안된 노드 간 유사도를 계산하는 방법이다. 각 $c_i$와 $c_i$에 대해 거리를 구하는데, p(c)는 클래스 집합 안에 c가 발생하는 확률빈도이며, -log(p(c))는 c의 정보량으로 c의 계층적 정보량을 나타내는 값이다. S($c_i$, $c_i$)는 상위노드 c를 의미하며, 두 노드가 공유하고 있는 상위노드가 많을수록 두 노드는 유사하다고 할 수 있다. 본 논문에서는 Equation (6)에 따라 각 후보 노드의 거리를 구하여 표현한다.

$$\text{sim}(c_i, c_j) = \max_{c \in S(c_i, c_j)}[-\log(p(c))] \tag{6}$$

Fig. 5는 테스트증상에 top-4를 나타낸 것이며, 특정 증상에 의해 추출된 병명들을 온톨로지로 표현한 것이다. 신뢰도에의해 독감(Influenza)이 Root 노드로 선택되었고, 각 질병들의 온톨로지 거리를 구하였다. 이 표현에 따라 의료 전문가는 간암(Liver cancer)보다 독감(Influenza), 급성 인두염(Acute rhinitis), 급성 비염(Acute pharyngitis)으로 의사결정 할 수 있을 것이라 기대한다. 즉, 정점의 크기가 크고 간선이 가깝고 모여 있을수록 테스트 증상과 더 가까운 질병이라고 말할 수 있다. 후보로 채택된 질병에 대해 온톨로지 간의 거리를 구하여 의료 전문가 의사결정에 도움을 주기위해 시각적으로 표현한다.

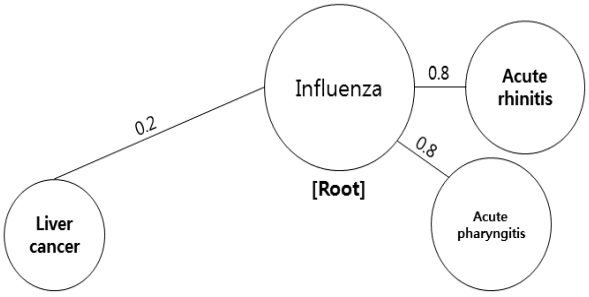


Fig. 5. Ontology representation

## 4. 성능 평가

본 논문이 제시한 TF-IDF 기반의 FP-Growth 알고리즘의 성능을 평가하기 위해, 실제 임상 기록 데이터셋을 사용하였다. 이는 각 환자 기준으로 의료 전문가에 의해 작성된 진료 및 처방 이력에 대한 자연어로된 텍스트 데이터이다. 본 장에서는 성능 평가를 위한 실험 환경에 대해 설명하고 제안 시스템의 성능을 기존 FP-Growth 알고리즘을 사용한 시스템과 비교하여 정성적으로 평가한 결과를 제시한다. 또한, 본 제안 시스템에서 제공하는 의료진의 병명 진단을 돕기 위한 입력 임상 병리 테스트 데이터에 대한 후보 병명 추론 및 온톨로지 표현 결과도 제시한다.

### 4.1. 실험 환경

1) 실험 데이터 셋

본 논문에서 사용된 의료소견 데이터는 MIMIC2 데이터는 PhysioNet에서 연구를 목적으로 미국 국립 보건원의 후원하에 제공되는 임상 데이터베이스이다[17]. 데이터 필드는 의료 전문가에 의한 진단과 병력, 환자의 상태, 처방 내역 등이 있다.

2) 성능 평가 환경

본 논문은 Table 5와 같은 환경에서 실험을 진행하였으며, 제안 시스템은 Java로 구현되었다.

Table 5. Experimental environment

| | 환 경 |
|---|---|
| CPU | Intel i7-5930K |
| | 3.50GHz |
| RAM | 16GB |
| OS | Windows7 |
| Data | MIMIC2[15] |

본 논문에서 $\gamma$값인 TF-IDF의 임계값은 10으로 하였으며, $\alpha$인 최소지지도의 임계값은 3으로 하였다. 문서의 수는 3만개이며, Table 6과 같은 테스트 데이터를 이용하여 실험을 진행하였다.

3) 실험 비교 대상

본 논문에서 제안한 TF-IDF 기반 FP-Growth 알고리즘을 이용한 병명 추론 시스템의 성능을 평가하기 위해 FP-Growth를 사용한 기존 시스템[9, 12]과 성능을 비교하였다. 성능 평가 기준이 정량적인 성능 평가가 아닌 정성적인 평가이므로, 성능 평가를 위해 제안 시스템과 기존 시스템으로부터 도출된 연관규칙들의 내용이 의학적 진단을 위한 구체적이고 유의미한 결과물인지 여부를 비교하였다.

### 4.2 실험 결과

Table 6은 원시 데이터에서 불용어(Stop word)를 제거한 후 시스템에 적용된 테스트 데이터로 기존 FP-Growth 기반 연관규칙 발견 시스템과 제안 시스템에 적용되었다. 각 시스템을 이용하여 3만개의 학습 데이터를 이용하여 연관규칙을 발견하고, 입력된 테스트 데이터와 관련된 연관규칙들을 찾아 그 결과를 보여준다. Table 7은 기존 연관규칙 분석을 사용하여 나온 결과물이고, Table 8은 본 논문에서 제안하는 시스템을 통해 나온 결과물이다.

Table 6. Experimental data

| Pathologic Data |
|---|
| small term female omphalocele rule sepsis ampicillin gentamicin hypoglycemia treated baby girl week female omphalocele born year old prima gravida history early sabs prenatal screens antibody negative rubella immune rpr nonreactive gc negative hepatitis surface antigen negative gbs unknown conceived clomid hcg pregnancy complicated lupus anticoagulant diagnosed work sabs treated heparin aspirin aspirin discontinued chronic hypertension requiring medications pregnancy aldomet started weeks switched labetalol weeks gestational diabetes treated |

기존 FP-Growth 기반 결과로 추출된 연관규칙을 살펴보면(Table 7), Table 6의 실험 데이터를 통해 교모세포종(glioblastoma), 대동맥판 협착증(aortic stenosis), 낭창(lupus)이 후보 병명으로 추론되었을 알 수 있다. 연관규칙에서는 year, old, female, pain 등과 같은 병명과 관련 없는 의료 소견 데이터에서 매우 빈번하게 발생하는 용어들이 연관규칙으로 추출된 것을 볼 수 있는데, 이는 기존 연관규칙이 빈번한 항목을 기반으로 생성되었기 때문이다. 이렇게 기존 FP-Growth 기반 방법을 의료 분야 텍스트 데이터에 그대로 적용하게 되면 의학적으로 특이적인 용어가 아닌 대부분의 데이터에서 빈번하게 나타나는 용어들에 의해 조건식이 구성된 연관규칙들이 높은 점수로 생성되게 된다. 따라서, 연관 규칙의 조건부에 나타난 용어와 결과부에 나타난 병명 사이의 특이적 관련성이 매우 떨어지게 되어 예측된 후보 병명이 실제 테스트 데이터의 특이적인 내용과는 관련이 적을 가능성이 커지게 된다.

제안 시스템을 사용했을 때에는 3만개의 학습 데이터로부터 12550개의 FP-Tree의 노드가 생성되었으며, Table 6의

Table 7. Association Rules in traditional method

| Term | Association Rule |
|---|---|
| glioblastoma | {year old female days pain} → {glioblastoma} (Support: 0.982, Confidence: 0.922, Lift: 2.17) |
| aortic stenosis | {old man medical history early white gentleman} → {aortic stenosis} (Support: 0.923, Confidence: 0.897, Lift: 0.945) |
| lupus | {aspirin sixth cycle complicated twin diabetes left right sided girl sepsis} → {lupus} (Support: 0.971, Confidence: 0.932, Lift: 1.56) |

Table 8. Association Rules in our System

| Term | Association Rule |
|---|---|
| omphalocele | {sabs gc clomid aldomet intestines hypospadias omphalocele} → {omphalocele} (Support: 0.988, Confidence: 0.952, Lift: 1.17) |
| clomid | {aldomet intestines omphalocele clomid} → {clomid} (Support: 0.973, Confidence: 0.942, Lift: 1.03) |
| hypertension | {hypercholesterolemia diabetic popliteal neuropathy hypertension} → {hypertension} (Support: 0.936, Confidence: 0.911, Lift: 0.991) |

실험 데이터를 테스트 데이터로 입력하였을 때 Table 8의 관련 연관규칙들이 추출되었다.

제안 시스템을 통해 얻은 연관규칙 결과에는 테스트 데이터 중 'old, female'과 같은 문서 집합에서 빈번한 단어에 대해서는 TF-IDF 값에 의해 중요도가 낮기 때문에 연관규칙에 포함이 되어있지 않다(Table 8). 이로 인해 Table 7과 비교하면 높은 점수로 발견된 연관규칙의 조건부가 매우 구체적인 의학용어들 기반으로 구성되어 있음을 확인할 수 있다. 따라서 발견된 연관규칙들이 특정 의학적 상황들과 병명의 연관성을 구체적으로 표현하게 된다.

또한 제안 시스템은 기존 시스템에서 더 나아가 발견된 연관 규칙들로부터 해당 입력 데이터와 관련된 후보 질병명을 온톨로지 형태로 보여준다. 발견된 연관규칙들 중 각 단어들이 병명이 될 후보가 되며, OBO Foundry[18]에서 각 단어들을 검색한 후 병명이 포함되는 단어들이 후보가 된다. 후보 병명에 대한 온톨로지를 구성하기 위한 Root노드는 연관규칙 결과의 대푯값으로 사용될 수 있는 향상도를 이용하여 결정한다. 따라서 테스트 데이터와 관련된 연관규칙들(Table 8)로부터 얻은 후보 병명들(Table 9) 중 향상도 1.17을 갖는 'omphalocele(배꼽탈장)'이 Root노드가 되어 온톨로지를 구성한다.

Table 9. Candidate

| No | Candidate name of disease |
|---|---|
| 1 | Omphalocele |
| 2 | Hypospadias |
| 3 | Clomid |
| 4 | Hypercholesterolemia |
| 5 | Diabetic |
| 6 | Neuropathy |

Root노드가 Omphalecele이고 각 후보 병명들간의 온톨로지 간의 거리를 구하여 Fig. 5와 같은 결과를 얻을 수 있다. Fig. 5의 직관적인 후보 병명 온톨로지를 통해, 의료 전문가는 Table 6의 임상병리 기록 데이터를 보고 환자의 병명을 결정할 때, 배꼽탈장(Omphalocele), 요도하열(Hypospadia)과의 관련성이 높고, 불임(Clomid), 신경통(Neuropathy), 고콜레스테롤혈증(Hypercholesterolemia), 당뇨병(Diabetic) 등과도 관련이 있음을 참고할 수 있다.

본 제안 시스템에 Table 6과 같은 테스트 데이터를 입력하면, 관련 연관규칙 리스트(Table 8)와 후보 병명들(Table 9), 그리고 후보 병명들의 관련성에 대한 온톨로지(Fig. 5)가 의료 전문가에게 제공되어 의료진의 진단을 위한 의사결정 참고자료로 활용될 수 있다.

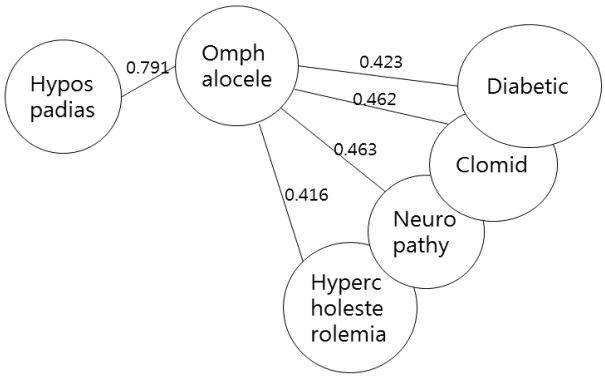


Fig. 6. Ontology result

## 5. 결 론

기존 의료 데이터 마이닝 연구들에서 연관규칙 기반의 의료데이터를 활용한 연구들은 주로 특정 질병을 정해두고 그와 관련된 증상들의 연관규칙을 찾아 발견되는 연관규칙의 범위가 제한되어 있었다. 또한, 기존 연구들은 연관규칙을 생성하는데 있어서 용어의 빈발도를 기준으로 임계값 아래의 항목들은 연관규칙 생성에 포함되지 않아 의학적으로 중요하지만 빈번하지 않은 용어들은 연관규칙에서 제외된다는 문제점이 있었다.

본 논문에서는 이러한 문제점들을 개선하고, 진료 소견데이터로부터 학습된 질병들의 증상을 이용해 질병명을 추론하는 시스템을 제안하였다. 의료 소견 데이터 문서군에서 빈번하지 않은 항목이 연관규칙 생성에 포함되지 않은 문제점을 해결하기 위한 TF-IDF 가중치 기반의 FP-Growth 기법을 제안하여 빈번하지 않은 항목도 연관규칙 생성에 포함되도록 개선하였다. 또한, 제안 시스템은 특정 진료 소견 데이터에 나타난 연관규칙들에 나타난 용어들의 의학 온톨로지상 거리

를 분석하여 해당 진료 소견 데이터에 대한 후보 병명을 추론하고 이를 시각화하여 제공함으로써 의료 전문가의 진단에 참고 자료로 활용될 수 있도록 하였다. 실제 임상 텍스트 데이터를 활용한 실험을 통해 제안 시스템이 기존 FP-Growth 알고리즘기반 시스템에 비해 의학적으로 더 의미있고 구체적인 연관규칙을 발견함을 확인할 수 있었다.

향후 연구로는 연관규칙 생성에 FP-Growth 알고리즘을 병렬화 및 분산화하여 성능 향상을 하는 연구가 필요하다. 또한, 의료 온톨로지가 현재 병명 위주로 구성되어 있는데, 질병과 증상간의 관계에 관한 온톨로지 구조가 확립되어 활용된다면 더욱 의학적으로 의미있고 정확한 연관규칙을 생성하여 의료 전문가의 의사결정에 도움을 주는 시스템이 될 것이라 기대한다.

## References


[1] S. H. Kim, "Health IT Technology Trends," *Electronics and Telecommunication Trens*, Vol.25, No.6, pp.37-46, 2011.

[2] Ottes, Leo, "Health 2.0 - It's up to You.," *Medicine 2.0 Conference*, JMIR Publication, 2010.

[3] Jorge C. G. Ramirez, Lon A. Smith, and Lynn L. Peterson, "Medical Information Systems: Characterization and Challenges," *ACM SIGMOD*, Vol.23, No.3, pp.44-53, 1994.

[4] Moon Koo Kim, Jong Hyun Park, and Young Hwan Joe, "A Study on the Key Success Factors of Big Data for Health Car," *KSII*, pp.239-240, 2013.

[5] Hisham Al-Mubaid and Rajit K Singh, "A new text mining approach for finding protein-to-disease association," *American Journal of Biochemistry and Biotechnology*, Vol.1, No.3, pp.145-151, 2005.

[6] J. Bjorne, Filip Ginter Heimonen, and Antti Airola, "Extracting complex biological events with rich graph-based feature sets," *Proceedings of the Workshop on Current Rends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics*, pp.10-18, 2009.

[7] Kim Jung-jae, Piotr Pezik and Dietrich Rebholz-Schuhmann., "MedEvi: retrieving textual evidence of relations between biomedical concepts from Medline," *Bioinformatics*, Vol.24. No.11, pp.1410-1412, 2008.

[8] Jeongkyun Kim and Jung-jae Kim, "DigSee: disease gene search engine with evidence sentences(version cancer)," *Nucleic Acids Research,* 41(Web Server issue), pp.510-517, 2013.

[9] Abdullah Saad Almalaise Alghamdi, "Efficient Implementation of FP-Growth Algorithm-Data Mining on Medical Data," *International Journal of Computer Science and Network Security*, Vol.11, No.12, pp.7-16, 2011.

[10] Dong Hoon Yang, Ji Hoon Kang, and Seoung Bum Kim, "Association Rule Mining and Network Analysis in Oriental Medicine," *PLOS one*, Vol.8, No.3, 2013.

[11] Rakesh Agrawal and R. Srikant, "Fast algorithms for mining association rules," *VLDB*, Vol.1215, pp.287-499, 1994.

[12] J. Han, J. Pei, and Y. Yun, "Mining frequent patterns without candidate generation," *ACM SIGMOD Int. Conf. Manag. Data*, Vol.29, No.2, pp.1-12, 2000.

[13] Yanbo J. Wang, Q. Xin, and F. Coenen, "A Novel Rule Weighting Approach in Classification Association Rule Mining," *Seventh IEEE International Conference on. IEEE*, pp.271-276, 2007.

[14] Dong Gyu Lee, Kwang Sun Ryu, Mohamed Bashir, Jang Whan Bae, and Keun Ho Ryu, "Discovering Medical Knowledge using Association Rule Mining in Young Adults with Acute Myocardial Infraction," *Journal of Medical System*, Vol.37, No.2, pp.1-10, 2013.

[15] Sajid Mahmood, Muhammad Shahbaz, and Aziz Guergachi, "Negative and Positive Association Rules Mining from Text Using Frequent and Infrequent Itemsets," *The Scientific World Journal*, 2014.

[16] MIMIC2 [Internet], https://physionet.org/.

[17] Goldberger, Ary, Jeffrey M. Hausdorff, Joseph E. Mietus, and H. Eugene Stanley, "PhysioBank physiotoolkit, and physionet components of a new research resource for complex physiologic signals," *Circulation*, Vol.101, No.23, pp.215-220, 2000.

[18] OBO Foundry [Internet], http://www.obofoundry.org.

[19] Philip. Resnik, "Using information content to evaluate semantic similarity in a taxonomu," arXiv preprint cmp-lg/9511007, 1995.



**박 호 식**

e-mail : pamiers@ajou.ac.kr

2014년 아주대학교 컴퓨터공학과(학사)

2014년~현 재 아주대학교 컴퓨터공학과 석사과정

관심분야 : 분산/병렬 컴퓨팅, 데이터 마이닝



**이 민 수**

e-mail : michelle.lee@ewha.ac.kr

2001년 이화여자대학교 수학과(학사)

2003년 이화여자대학교 컴퓨터공학과(석사)

2007년 이화여자대학교 컴퓨터공학과(박사)

2007년~2011년 서울대학교 컴퓨터공학과 박사후연구원

2013년~2014년 (미)인디애나대학교 정보컴퓨팅학부 방문연구원

2014년~현 재 이화여자대학교 컴퓨터공학과 연구교수

관심분야 : Active learning, Adaptive stream data mining, Bioinformatics, Machine learning

황 성 진

e-mail : seongjins@humintec.com
2015년 아주대학교 산업공학과(석사)
2007년~현 재 (주)휴민텍 의료영상사업부 팀장
관심분야 : 의료정보, 웹 프로그래밍

오 상 윤

e-mail : syoh@ajou.ac.kr
2006년 (미)인디애나대학교 컴퓨터공학과 (박사)
2006년~2007년 SK텔레콤
2007년~현 재 아주대학교 소프트웨어학과 부교수
관심분야 : 분산/병렬 시스템, 고성능컴퓨팅, Large Scale Software System, Semantic Web