개인 맞춤형 의료진단 서비스 제공을 위한 효율적인 데이터마이닝 기법

권은희°, 이승철°, 이주창, 김응모
성균관대학교 전자전기컴퓨터공학과
e-mail : {nonimi, eddie, lordeath, umkim}@ece.skku.ac.kr

Efficient Mining for Personalized Medical treatment Diagnosis Service

Eun-Hee Kaun°, Seung-Cheol Lee°, Joochang Lee, UngMo Kim
Electrical and Computer Engineering, Sungkyunkwan University

요 약

최근 유비쿼터스 환경의 발달로 인해 사용자 중심의 유비쿼터스 기술이 활발히 연구되고 있다. 이에 따른 각종 응용 분야가 활발히 연구 중이며, 그 중에서 특히 U-Health기술이 주목 받고 있다. U-Health 기술은 질병의 치료라는 전통적인 관점의 의료 서비스에서 벗어나 건강한 상태의 지속적인 관리와 질병의 예방이라는 적극적이고 확장된 개념으로 발전해가고 있다. 건강상태를 관리하고 진단하기 위해서는 기존의 진단데이터를 효율적으로 관리하고, 그것을 토대로 하여 유용한 정보를 얻어 낼 수 있는 방법이 필요하다. 지금까지는 데이터를 처리하기 위하여 통계적인 수치나 전문가에 의한 전문지식을 토대로 하는 방법을 사용하고 있다. 그러나, 건강상태를 관리하고 진단을 목적으로 하는 시스템에서는 높은 정확성이 보장되어야 한다. 또한 유비쿼터스 환경의 특성상 적은 메모리의 사용과 빠른 마이닝 속도가 수반되어야 한다. 본 논문에서는 튜플기반의 진단데이터들을 마이닝하여 진단패턴을 뽑아내는 의료 진단마이닝 알고리즘을 제안한다. 본 알고리즘은 진단패턴정보의 정확성을 높일 수 있는 장점을 가지며, 튜플기반의 데이터들을 트리 구조로 구성함으로써 마이닝 속도를 향상시킨다. 더 나아가 트리 구조의 컴팩트한 데이터 구조로 메모리 적재가 용이하다. 이는 센서가 부착된 개별 사용자로부터 실시간으로 들어오는 건강상태와 진단패턴과의 비교, 분석을 가능하게 함으로써 보다 정확하고 빠른 진단결과를 내려줄 수 있는 의사결정시스템의 사용에 적합하다.

1. 서 론

유비쿼터스 환경이란 눈에 보이지 않는 컴퓨터가 인간의 생활 공간 도처에 설치되어, 생활을 하면서 필요시 쉽게 서비스를 제공받을 수 있는 조용한 컴퓨팅의 패러다임이다. IT환경 및 패러다임이 유비쿼터스로 변함에 따라 의료서비스 및 의료 기술의 패러다임도 함께 변화하고 있다[1].* 현재의 의료 서비스는 공간의 제약을 받고 특정 장소에 한정되어 있으며, 의료진을 중심으로 이루어지고 있다. 그러나, 유비쿼터스 기술의 각종 응용분야 중 U-Health기술의 발전과 맞물려 병원 중심의 치료 개념에서 환자의 생활 공간에서의 건강관리 개념으로 의료 서비스 패러다임이 변화하고 있다. 더 나아가 의료소비자들에게 더욱 질 높은 의료 서비스를 제공하기 위해 개인에게 특화된 개인맞춤형 의료 서비스를 제공하는 방향으로 연구 개발되고 있다. 즉, 언제, 어디서나 컴퓨터와 연결이 가능하며, 휴대 가능한 장치, 초소형 센서와 같은 휴대형 진단기기를 통하여 개인이 수시로 건강상태를 확인함으로써 조기 진단이 가능하다는 것이다.[2] 이러한 서비스는 기존의 방대한 의료데이터를 처리하는 것이 선행되어야 한다. 더 나아가 기존의 데이터뿐만 아니라 센서로부터 들어오는 실시간 건강상태정보를 처리하고 관리하는 방법도 연구해야 한다. 지금까지 사용되고 있는 데이터를 분석하고 처리하는 기법에는 Naïve Bayes 기법[3], 포함알고리즘[4], 사례기반학습[5]등 이 있다. 이러한 방법들은 정보량계산, 정보이득비계산 등과 같은 통계학적인 방법을 사용하고 있다. 이것은 진단의 목적으로 사용하기에 정확성을 보장할 수 없으며, 많은 질병이 새로 생기고 새로운 증상들이 많이 발생하는 현시점에서 새로운 데이터를 처리하는데 에는 역부족일수 밖에 없다. 이에 따라 데이터마이닝 기술을 도입하여 정확성 향상 및 효과적인 데이터 처리를 하기 위한 연구들이 진행되고 있다.

본 논문에서는 튜플기반의 진단데이터들을 마이닝 하여 진단패턴을 뽑아내는 의료진단 마이닝 알고리즘

* 본 연구는 한국전자통신연구원 지능형 상황정보 인식 및 관리기술 개발 (2007-0475-000) 지원으로 수행되었음.

(Medical Diagnosis Mining Algorithm: MD-mine)을 제안한다. 의료진단 마이닝 알고리즘은 진단 패턴을 찾기 위하여 의료진단트리(Medical Diagnosis Tree: MD-tree)를 생성한다. 트리 구조는 현대와 같이 새로운 병명들이 늘어나는 시점에서 업데이트가 용이하며, 방대한 데이터를 컴팩트하게 관리 할 수 있다. MD-tree는 진단기록 데이터베이스에 저장되어 있는 진단기록들을 바탕으로 구성되며, 트리의 구성을 위하여 진단기록 데이터베이스의 트랜잭션당 단 한번의 스캔을 필요로 한다. 이는 데이터베이스의 검색시간을 줄여줌으로써 마이닝 수행 속도를 향상시키는데 기여한다.

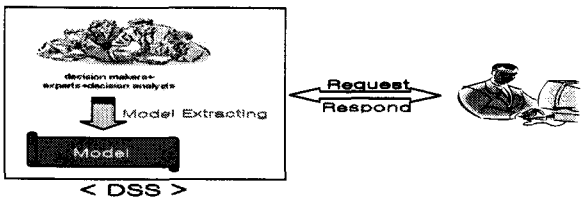
본 논문의 구성은 2장에서는 관련연구를 소개하며, 3장에서는 본 논문에서 가정하고 있는 시스템과 해당 시스템의 요구사항을 소개한다. 4장에서는 의료진단마이닝 알고리즘을 제안한다. 5장에서는 앞으로의 향후 발전 방향에 관하여 언급한다.

2. 관련연구

본 장에서는 의사결정시스템이 무엇인가에 대한 소개를 한다. 그리고 의사결정시스템에 데이터마이닝을 접목시킴으로써 데이터마이닝이 의사결정시스템에서 어떠한 역할을 수행하게 되고 어떠한 효율성이 있는지 간단하게 살펴본다. 마지막으로 연관규칙을 찾는 기본적인 개념정리를 위한 용어를 설명한다.

2.1 의사결정시스템

의사결정 시스템은 정보를 효율적으로 수집하고, 저장하고, 분배하기 위한 시스템이 아니라 사용자의 의사결정을 지원하여 의사결정의 효율성을 향상시키기 위한 정보시스템이다. 그러므로 의사결정지원시스템은 일반적인 정보시스템과는 다른 개발방식으로 개발된다. 즉, 일반적인 정보시스템에 있어서는 개발자가 사용자의 요구를 조사하고 시스템개발의 대상이 되는 업무를 분석하여 시스템을 설계한 다음 프로그래밍 과정을 통해 시스템이 개발되면 이를 사용자에게 제공하게 된다. 그러나 의사결정지원시스템의 개발방식은 [그림1]과 같이 사용자의 요구를 반영하여 시스템을 개발하고 이를 사용자가 직접 사용하고 평가하게 하여 새로운 요구와 문제점을 찾고 이들을 해결하면서 시스템을 개선해 나가는 방식을 거치게 된다. 이러한 개발방식을 프로토타이핑(prototyping)이라 하는데, 가능한 한 빠른 시간 안에 사용자의 기초적인 요구를 최대한 반영한 소규모 모형을 구축한 후 이를 점진적으로 개선해 나가는 시스템 개발방식을 말한다.



[그림 1] 의사결정시스템의 기본적인 구조

2.2 데이터마이닝과 의사결정시스템

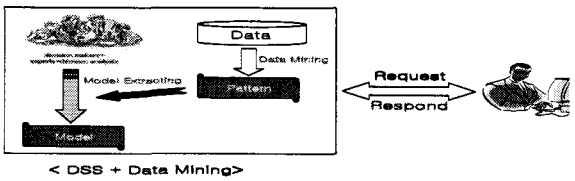
2.2.1 데이터마이닝의 개요

현시대에는 대용량에 저장된 수많은 데이터와 빠르게 증가하는 엄청난 양의 데이터로 인해"데이터의 무덤"현상이 발생되고 있다. 따라서 데이터를 빠르게 처리하고 유용한 정보를 얻어내기 위한 강력한 도구가 필요하게 되었다. 데이터 마이닝은 수많은 데이터로부터 유용한 패턴 및 경향을 찾는 효율적인 기법이다.

2.2.2 데이터마이닝과 의사결정시스템의 연계 필요성

지금까지의 의사 결정자는 방대한 데이터에 묻혀 있는 가치 있는 지식들을 추출해주는 도구들이 없었기 때문에, 데이터베이스에 저장된 풍부한 정보를 바탕으로 의사결정을 하지 못하였다. 때문에, 자신의 경험에 의하여 결정을 거나, 전문가들의 지식, 통계적인 수치 등에 근거한 의사결정시스템에 의존하여 중요한 의사 결정을 하곤 하였다. 그러나, 현시대는 빠르게 변화하고 있으며, 전문가들이 예상치 못하는 일들과 통계상으로 발생할 수 없는 이들이 발생되고 있다. 또한 통계적인 수치로 얻어지는 정보이기 때문에 정확성이 다소 떨어지게 된다. 따라서, 중요한 의사결정을 위한 시스템을 위하여 사용되기 에는 역부족이다.

이러한 문제점을 해결하기 위해 [그림2]와 같이 데이터마이닝 기술을 의사결정시스템과 연계시키려는 연구들이 진행되고 있다. 데이터 마이닝 기법은 기존에 저장된 방대한 데이터를 분석하며, 변화되는 데이터들을 주기적으로 수용함으로써 중요한 데이터 패턴을 찾아내는 역할을 한다. 이와 같이 발견된 패턴 및 지식들은 수많은 데이터를 기반으로 하여 추출되었기 때문에 신뢰성을 지닌 정보가 될 수 있다.



[그림 2] 의사결정시스템 + Data Mining

2.3 연관규칙

연관규칙을 찾는 기법에는 대표적으로 Agrawal et al.

에 의해 소개된 Apriori알고리즘[6]과 Han et al.에 의해 소개된 FP-growth method[7]가 있다. 연관규칙은 X=>Y 와 같이 표현되며, X와 Y는 데이터베이스를 이루는 트랜잭션을 구성하는 아이템집합이다. 이때, X와 Y의 교집합은 존재하지 않는다. 연관규칙은 지지도와 신뢰도 이 두 가지의 측정값으로 유도된다. 지지도(Support)는 전체 데이터베이스의 트랜잭션 중에서 "X U Y"를 포함하는 트랜잭션의 비율을 나타내며, 신뢰도(Confidence)는 데이터베이스에서 X를 만족하는 트랜잭션 중에서 "X U Y"를 포함하는 비율을 나타낸다. 지지도와 신뢰도는 다음과 같이 계산된다.

지지도(X=>Y) = freq(X U Y, D)
신뢰도(X=>Y) = freq(X U Y, D) / freq(X, D)

유용한 연관규칙이란 사용자에 의해 주어진 최소지지도(minimum support)와 최소신뢰도(minimum confidence)를 임계 값으로 정한 후 두 가지의 임계 값 이상이 되는 지지도와 신뢰도를 갖는 특정 규칙을 말한다.

3. 시스템 요구사항 및 동작

앞으로 제안할 의료진단마이닝(MD-mine)알고리즘은 개인 맞춤형 의료진단을 제공을 위한 의사결정시스템에서 동작하는 마이닝 알고리즘으로써 정확성이 높은 의료진단 패턴정보를 추출한다. 본 논문에서는 다음과 같은 시스템을 가정한다. 일정시간 간격으로 사용자의 센서로부터 건강상태정보가 무선 네트워크을 통하여 DSS 시스템에 전달된다. 전달된 데이터와 마이닝된 패턴을 비교 분석하여 특정병명이 의심되는 경우에 병명과 해당 조치를 사용자에게 전달한다. 동시에 사용자가 등록한 담당의사에게도 전달된다. 만일 기존에 존재하던 병명과 일치하는 병명이 발견되지 않은 경우에는 사용자와 담당의사에게 경고메시지를 보내며, 사용자가 원하는 경우 병원시간을 예약해줌으로써 빠른 진료가 가능하도록 한다. 사용자의 상태가 정상이라면 별도로 알리지 않는다. 이와 같이 일정시간 단위로 사용자의 상태를 일정시간 간격으로 관리 함으로써 조기 진단과 예방이 가능하다. 이와 같은 시스템이 정상적으로 작동되기 위한 요구사항은 [표 2]와 같다.

	요구사항
1	의료진단서비스에 정당하게 등록된 사용자이여야 함
2	사용자는 담당의사나 담당병원을 지정하여야 함. (지정하지 않았을 경우, 임의의 병원을 안내함.
3	센서의 상태및 무선네트웍 상태가 안정적이여야 함.
4	병원 시스템과의 연결이 안정적이여야 함.
5	반드시 센서는 "ON"이 된 상태여야 함.

[표 2] 시스템 요구사항

4. 의료진단트리를 이용한 마이닝 기법

4.1 필요성

방대한량의 데이터를 처리하여 의사결정을 하기 위한 기법은 많이 소개 된 바 있다.[3,4,5] 이들은 정보량 계산, 정보이득계산 등의 통계학적인 계산과 인공지능적인 요소로부터 정보를 습득하는 형식이다. 이러한 과정들은 계산이 복잡하며, 많은 시간을 소비하게 된다. 또한 새로운 병명과 증상들이 나타나고 있는 현시점에서 이러한 기법들은 새로운 정보들을 반영하기에는 역부족이다. 따라서 정확한 패턴의 추출과 빠른 수행속도를 가지는 마이닝 기법이 필요하며, 기존에 존재하는 병들과 증상들을 비롯한 새로 발생하는 건강상태정보들을 효율적으로 관리하며 처리할 수 있는 방법이 필요하다.

4.2 의료진단마이닝 알고리즘(MD-mine Algorithm)

사용자 별 건강상태를 진단하기 위하여 진단 데이터의 처리를 간단하고 효율적으로 빠르게 수행하며, 정확하고 정밀한 패턴을 습득 할 수 있도록 해주는 MD-mine 알고리즘을 제안한다. MD-mine 알고리즘은 MD-tree를 생성하여 마이닝을 수행한다.

4.2.1 의료진단트리(MD-tree)생성

MD-mine 알고리즘에서 트리 생성시 진단기록데이터베이스의 트랜잭션당 단 한번의 스캔이 필요하다. 이것은 대규모 데이터베이스의 탐색시간을 최소화함으로써 마이닝하는 시간을 줄여주는 요소가 된다. MD-tree의 마지막 노드는 항상 병명으로 구성 되며, 마이닝은 병명을 Key로 하여 진행되게 된다. MD-tree의 생성과정을 예를 들어 살펴보자.

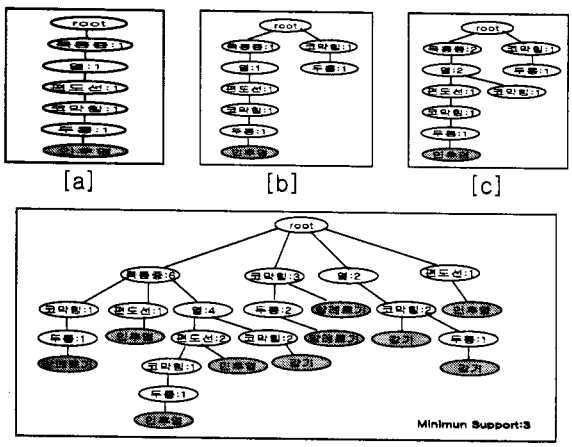
PID	목통증	열	편도선 부종	코막힘	두통	병명
1	Yes	Yes	Yes	yes	Yes	인후염
1	No	No	No	Yes	Yes	알레르기
2	Yes	Yes	No	Yes	No	감기
1	Yes	No	Yes	No	No	인후염
2	No	Yes	No	Yes	No	감기
3	No	No	No	Yes	No	알레르기
3	No	No	Yes	No	No	인후염
2	Yes	No	No	Yes	Yes	알레르기
1	No	Yes	No	Yes	Yes	감기
3	Yes	Yes	No	Yes	No	감기
2	Yes	Yes	Yes	No	No	인후염
3	No	No	No	Yes	Yes	알레르기

[표1] 진단 데이터베이스

표[1]은 예제 데이터로써, 3명의 환자의 진단 데이터

트랜잭션으로 이루어진 진단데이터베이스라고 가정한다. [표1]은 임의의 3명의 환자에 대한 진단데이터베이스를 보여주고 있다. 본 진단 데이터베이스에는 인후염, 알레르기, 감기의 세 가지의 병명이 있으며, 각각의 증상은 목통증, 열등 5가지로 분류하여 yes, no로 표시하였다. 이때, 가장 우측의 속성은 항상 병명으로 정렬한다.

이제 트리를 구성해보자. 해당증상이 있는 경우(yes경우)는 노드를 생성하고 없는 경우(No경우)는 노드를 생성하지 않는다. 트리가 구성되는 과정을 설명하기 위하여 첫번째에서 세번째 트랜잭션까지 한 트랜잭션씩 스캔하며 트리가 구성되어가는 과정을 설명하도록 한다. 첫번째 트랜잭션에서 목통증, 열, 편도선부음, 코막힘, 두통의 증상이 다 나타났으며 병명은 인후염인 경우의 트리는 [그림3-a]와 같이 생성된다. 두번째 트랜잭션에서 코막힘과 두통증상이 있었을 경우 알레르기라는 병명을 가졌고 [그림3-b]와 같이 생성된다. 세번째 트랜잭션에서 목통증, 열, 코막힘 증상이 있었을 경우 감기라는 병명을 가졌다. 세번째 트랜잭션을 스캔 한 후에는 첫번째 트랜잭션에서 의 증상과 중복되는 증상이 발생하였다. 목통증과 열의 증상이 동일하게 나타난다. 따라서 겹치는 증상의 경우 count를 1씩 증가시킨다. 코막힘도 중복되는 증상이지만 편도선이 중복되는 증상이 아니므로 코막힘 증상의 노드는 새롭게 생성된다[그림 3-c]. 이와 같은 동일한 방법을 통하여 최종으로 만들어진 MD-tree는 [그림 3-d]와 같다.



[a] [b] [c]

[d]

[그림 3] 의료진단트리 생성

트리가 생성 되어질 때 동시에 헤더테이블이 구성이 된다. 각각의 노드들은 헤더테이블에 등록이 되고 같은 동일한 이름을 가진 노드들은 서로 링크로 연결 되어진다. 헤더테이블은 노드명, 링크, Flag 필드로 구성되며, Flag 필드는 가장 마지막 노드, 즉 병명에 해당하는 노드에 Flag값을 1로 지정함으로써 병명에 해당하는 노드임을 표시한다. Flag필드를 구성하는 이유는 마이닝시에 병명별로 진행되기 때문에 마지막 노드에 해당하는 Flag필드 값을 1로 표시 하게 되면 Flag가 1로 표시된 노드에 대해서만 마이닝을 진행하도록 할 수 있다. 이때 마이닝하는 순서는 관계없다.

4.2.2 의료진단마이닝 알고리즘(MD-mine Algorithm)

진단패턴을 얻기 위한 작업 수행은 병명 별로 진행한다. Flag가 1인 노드 중에서 선택된 노드의 속성값을 동일하게 갖는 노드들의 단일 경로를 모두 뽑아내어 count의 값이 minimum Support보다 작은 값의 노드는 삭제하고 minimum Support와 같거나 큰 값의 노드는 삭제하지 않는다.

예를 들어 감기에 대한 마이닝이 진행된다고 가정하자. 감기에 해당하는 단일 경로는 [목통증:2, 열:2, 코막힘:2], [열:1, 코막힘:1], 그리고 [열:1, 코막힘:1, 두통:1] 이렇게 세개의 단일경로가 존재한다. 증상 별로 count값을 종합해보면, [목통증:2, 열:4, 코막힘:4, 두통:1] 과 같다. [그림3-d]에 표시되어 있듯이 Minimum Support값은 3이며, 3보다 작은 목통증과 두통의 증상에 해당하는 노드는 삭제 된다. 따라서 감기에 걸리면 열과 코막힘 증상이 수반된다는 진단패턴을 얻을 수 있다. 다음의 [그림 4]는 MD-tee를 구성하여 마이닝을 수행하는 MD-mine 알고리즘이다.

Input: 진단기록 데이터(트랜잭션단위)
Output: 병명 별 패턴

Step1. 진단기록 데이터베이스에서 진단기록을 트랜잭션단위로 가져옴
Step2. 트랜잭션들의 속성 중에 병명에 해당하는 것은 반드시 가장 오른쪽으로 정렬
Step3. 트랜잭션의 왼쪽에서부터 오른쪽으로 스캔 하면서 새로 발생하는 증상들은 노드로 생성, 겹치는 증상이 있을 경우 count를 1씩 증가하면서 트리 구성
(이때, 트리를 구성하면서 헤더테이블 또한 함께 생성, 같은 노드들은 링크로 연결됨. 헤더테이블은 노드명, 링크, Flag필드로 구성되며, Flag필드는 0으로 초기화되어 있음. 가장 마지막 노드에 해당하는 필드는 1로 표시)
Step4. Flag가 1로 표시된 노드를 하나 선택 후, 선택된 속성을 포함하는 경로를 탐색하고, 경로에 포함되어 있는 노드들의 count값을 확인하여, 주어진 Minimum Support값보다 작은 값을 가지는 노드들은 삭제. (남은 노드는 선택된 병이라고 판명되었을 때 수반될 확률이 높은 증상.)
Step5. Flag가 1인 노드에 대하여 Step4를 반복수행.

[그림 4] MD-mine Algorithm

5. 결론 및 향후 연구방향

지금까지 우리는 유비쿼터스 환경하에서 센서로 들어오는 사용자의 상태로부터 사용자의 상태를 진단하는 의사결정시스템에 적합한 MD-mine Algorithm에 대하여 살펴보았다. 의사결정을 위하여 데이터마이닝 기술을 도입함으로써 패턴의 정확성을 높일수 있었으며, 특히 MD-mine 알고리즘은 트리구조로 데이터를 처리함으로써 메모리의 사용량을 줄이고 마이닝의 속도향상을 도모할 수 있었다. 또한 트리를 생성할 때 데이터베이스의 스캔을 단 한번으로 간소화 함으로써 데이터베이스의 탐색시간을 줄임으로써 속도를 향상시키는 요소를 제공한다. 현 시대에는 많은 새로운 질병들과 증상들이 발생하고 있다. MD-mine알고리즘은 이와 같은 경우에서도 트리 구조로 데이터를 처리하기 때문에 유연성 있게 데이터를 처리할 수 있다. 건강에 대한 관심이 높아짐에 따라 개인 맞춤형 의료진단을 원하는 사람들이 증가하고 있다. 이러한 사회적인 흐름에 따라 앞으로도 더 많은 연구가 진행되어야 할 것이다. 또한 더욱 정확한 진단을 위해서 구체적인 의학적인 자료들이 동반되어야 할 것이다.

6. 감사의 말

본 연구는 21세기 프론티어 연구개발사업의 일환으로 추진되고 있는 정보통신부의 유비쿼터스 컴퓨팅 및 네트워크 원천기반 기술개발사업의 지원에 의한 것임.

* 참고문헌 *

[1]http://www.roboticslab.co.kr/ , HealthCare

[2]MindBranch Asia Pacific Co.Ltd, "U-Health 시장 현황 및 전망", 2005년 10월

[3]S.T.Dumais, J.Platt, D.Heckerman, and M.Sahmi, "Inductive learing algorithms and representations for text categorization," In CIKM, 1998

[4]geoffrey I. Webb, "A heuristic covering algorithm has higher predictive accuracy than learing all rules", Published in Preceedings of Information, Statistics and Induction in Science, Melbourne, World Scientific, 1996, pp.20-30

[5]Robert C. Holte and Stan Matwin (1998), "Machine Learning for the Detection of Oil Spills in Satellite Radar Images Miroslav Kubat", *Machine Learning*, volume 30, pp. 195-215. 2.8 Megabytes

[6]Dunja Mladenic, Nada Lavrac, Marko Bohanec, and Steve Moyle, "Data Mining and Decision Support Integration and Collaboration", Kluwer Academic Publishers Boston/Dordrecht/London

[7]R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rule," Proc. Int'l Conf. Very Large Data Bases, pp. 487-499, Sept. 1994

[8]JIAWEI HAN, JIAN PEI, YIWEN YIN, RUNYING MAO, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery, 8, 53-87,2004