혈연, 지역별로 분류한 개인 맞춤형 의료정보 서비스

황의철*, 이은주*, 김응모*
*성균관대학교 정보통신공학부
e-mail:zetenus@naver.com

Classified Medical Information Service by Blood and Local Group

Eui-Cheol Hwang*, Eun Ju Lee*, Ung-Mo Kim*
*School of Information and Communication Engineering,
Sungkyunkwan University

요 약

데이터베이스 시스템의 사용은 다양한 분야에서 필수적으로 사용되고 있다. 정보의 양이 증가함에 따라서 축적된 정보들의 연관성을 찾아내어 새로운 정보를 발굴하는 데이터 마이닝 기법이 지속적으로 연구 개발되고 있으며 데이터 마이닝 기법으로 얻게 된 새로운 정보들은 새로운 가치 창출을 위해 여러 분야에 적용되고 있다. 그 중에서도 의료 서비스 분야에서의 데이터베이스 시스템은 고령화 시대에 건강의 중요성이 강조됨에 따라 보다 적극적으로 활용되어지고 있다. 하지만 지금까지의 의료분야에서의 데이터베이스 시스템 활용은 개인의 의료정보를 데이터베이스 시스템에 저장하고 그것을 바탕으로 개인의 건강에 대한 검진이나 치료 등의 서비스를 제공하는데 국한되어왔다. 개인 맞춤 의료 서비스에도 불구하고 유전이나 지역특화적인 질병, 질환에 의한 때늦은 발견과 치료는 여전히 많은 이들을 고통 받게 하고 있다. 이에 본 논문에서는 각 의료기관에 등록된 환자의 각 질병의 발병 패턴과 치료 정보 등을 토대로 유전적요인과 환경적 특성을 고려한 집단으로 분류하고 환자가 속한 집단구성원에게 검진정보를 제공할 수 있는 의료검진정보 시스템을 제안한다.

1. 서 론

현대 의학이 발전하면서 이에 맞추어 사회보장제도의 의료복지 분야도 많은 발전을 이루어왔다. 의학의 발전은 많은 병들을 치료가능하게 만들었고 의료 복지제도는 국민 누구나 평등하게 이러한 혜택을 받을 수 있게 하고 있다. 의료 복지의 발전으로 더욱 많은 사람들이 각 종의 질병과 질환들의 위험으로부터 보호 될 수 있었으며 또한 질병, 질환을 겪는 사람들의 고통이 감소되어 질 수 있었다. 의료 복지 정책의 하나인 의료분야 정보 체계와 데이터베이스 구축을 통하여 얻게 되는 의료 관련 정보의 제공은 정보 통신기술이 복지국가의 국민들에게 제공할 수 있는 가장 비중이 높은 서비스 중의 하나이다. 또한 정보통신기술을 토대로 한 의료관련 정보의 접근 용이성은 진료의 최적화와 편의성 향상, 의료수준 및 의학 수준의 향상을 가져올 뿐만 아니라 국내 의학 관련사업의 경쟁력 증진에도 기여하고 있다[1]. 그러나 이러한 데이터베이스 분야의 의료분야로의 큰 기여에도 불구하고 현재의 국민 의료 관련 정보 서비스는 국민 개개인의 유전적, 환경적 요인과 같은 외부적 요인까지 고려한 정보를 제공하지 못하는 것이 현실이다.

따라서 본 논문에서는 축적된 국민 개인의 건강 정보를 데이터 마이닝 기법을 이용함으로써 개개인의 유전적, 환경적 요인에 따라 진행되는 각 종 질환, 질병 등에 대하여 패턴을 분석하고 각 요인들의 구분을 집단화하여 그 결과 정보를 해당 집단에 속한 국민에게 제공하는 것에 대하여 제안한다. 본 논문의 구성은 2장에서는 배경지식을 소개하며, 3장에서는 본 논문에서 제안하는 시스템에 대해서 설명한다. 4장에서는 제안한 시스템에 대한 보완점 및 향후 발전 방향에 대하여 언급한다.

2. 배경지식

본 장에서는 데이터 마이닝 기법에 대해 소개한다.

2.1 데이터 마이닝

데이터이닝(Date Mining)은 대용량의 데이터에서 숨겨진 유용한 패턴을 추출하는 방법론을 일컫고, 이것은 OLAP와 Data Warehousing을 구축할 때 중요한 도구로 사용하고 있다[2]. 즉, 데이터마이닝이라는 기술은 방대한 데이터에서 쉽게 드러나지 않는 유용한 정보를 발견하는 과정, 데이터간의 겉으로 드러나지 않거나 또는 기존의 통계학적 방법을 통해 뽑아내기에는 복잡한 관계를 찾아내고 이 관계를 바탕으로 미래의 상황을 예측하는 기술이다.

2.2 데이터 마이닝 기법

데이터 마이닝의 각 기법들의 간단한 개념은 다음과 같다.

2.2.1 연관 규칙(Association rules)

상품 혹은 서비스 간의 관계를 살펴보고 이로부터 유용한 규칙을 찾아내고자 할 때 이용될 수 있는 기법 이다. 상품이나 서비스의 거래기록 데이터로부터 상품 간의 연관성정도를 측정하여 연관성이 많은 상품들을 그룹화 하는 클러스터링의 일종이며, 동시에 구매 될 가능성이 큰 상품들을 찾아냄으로써 시장바구니분석에서 다루는 문제들에 적용될 수 있다. 연관규칙을 찾는 대표적인 기법에는 Apriori알고리즘[3]과 FP-growth-method[4]가 있다.

2.2.2 연속규칙(Sequence)

일정 시간 동안 레코드를 분석하여 순서 패턴(Sequential Pattern)[5]을 찾아낸다. 시간이 흐름에 따라 발생하는 패턴을 발견해 내는 작업이다. 즉 개인별 트랜잭션 이력 데이터를 시계열적으로 분석하여 트랜잭션의 향후 발생 가능성을 예측하는 작업이다.

2.2.3 분류탐사(Classification)

분류탐사(classification)[6]란 이미 알려진 그룹의 특징을 부여하여 어떤 범주에 근거하여 사전에 정의된 분류를 구분하는데 사용된다. 예를 들어 소득액이 일정액 이상이고 거주지에 따라 고객등급을 부여하는 방법 등이 이에 속한다.

2.2.4 의료진단마이닝 알고리즘(md-mine)

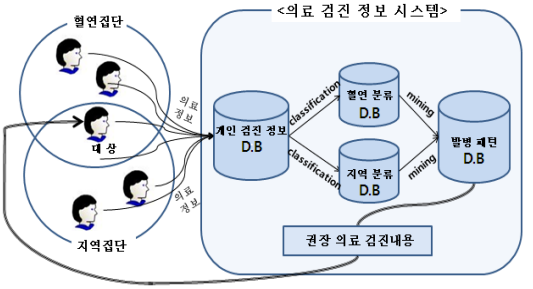
사용자 별 건강상태를 진단하기 위하여 진단 데이터의 처리를 간단하고 효율적으로 빠르게 수행하며, 정확하고 정밀한 패턴을 습득 할 수 있도록 해주는 알고리즘이다. MD-mine 알고리즘[7]은 MD-tree를 생성하여 마이닝을 수행한다.

3. 혈연, 지역집단으로 분류한 의료검진정보

본 장에서는 축적된 국민 건강 정보를 혈연, 지역별로 분류하고 패턴을 분석하여 집단별 개인 맞춤 건강 검진 정보 서비스를 제공하는 방법에 대해 설명한다. 오늘날 대부분의 사람들이 암을 비롯해 유전자가 직접 혹은 간접적으로 관여하는 질병으로 사망하고 있으며 전염병 같은 경우는 지역이나 생활 여건 등의 환경적 요인이 크게 작용한다[8]. 때문에 유전적 질병, 질환이나 지역에 국부적으로 빈번하게 발병되는 병들의 정보를 데이터 마이닝 기법을 이용하여 집단별로 분석하여 패턴을 찾아낸다면 개개인에게 특화된 더욱 나은 의료 검진 정보 서비스를 제공 할 수 있다.

3.1 의료검진 정보시스템의 구조

본 시스템의 정보 제공과정은 저장된 정보를 바탕으로 분류탐사기법(classification)으로 집단을 분류하고 집단내의 의료정보를 md-mine 알고리즘과 연속규칙(Sequence pattern)기법을 적용하여 해당 개인에 맞는 건강검진정보를 제공하도록 한다. (그림 1)은 의료검진정보시스템이 개략적인 정보제공 과정을 설명한다.



(그림 1) 의료검진정보시스템의 구조

(그림 1)에서 검진을 받는 대상은 의료검진정보시스템에 자신의 진단기록을 제공한다. 의료검진정보 시스템의 축적된 개개인의 검진정보는 혈연과 지역에 따라 분류되어 저장된다. 집단별로 저장된 정보들은 질병, 질환의 발병패턴으로 마이닝 된다. 마이닝 정보에 의해 검진 대상에 대한 권장 의료 검진 내용을 판단하고 해당되는 정보를 제공한다.

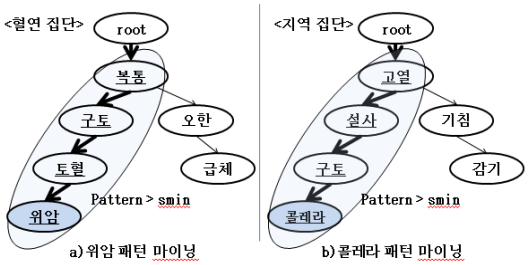
3.2 정보의 분류

본 논문에서 제안하는 서비스를 제공하기 위해서는 분류(classification)과정이 필요하다. 의료정보 마이닝의 대상이 되는 집단을 유전을 바탕으로 한 혈연집단과 환경과 지리적 특성을 고려한 지역집단으로 분류하는 것이 첫 번째 과정이며 이는 대상을 집단에 특화 된 분류기준에 맞추어 독립적으로 진행 한다. 특히 혈연집단 분류에서는 대상의 가족과 직계 조상뿐 만 아니라 유전자의 활용범위가 유효한 친계 대상까지 고려해야 한다. 또한 지역집단을 나누는 과정도 지역특화 질병, 질환이 미치는 환경적, 지리적 범위에 대한 선행 연구가 필요하다. 이는 질병, 질환에 대한 심층적인 의학 연구가 필요하므로 본 논문에서는 혈연집단과 지역집단으로 분류한다. 혈연집단은 직계 가족에 의한 분류로 하며 지역집단은 도시, 농촌, 어촌 등의 환경적 요인과 지명을 기준으로 분류한다.

3.4 집단별 발병패턴 마이닝

미리 결정된 최소지지도 smin 이상의 지지도를 갖는 모든 병의 증상들에 대한 빈발 항목집합들(large itemsets)을 찾은 후 증상 항목집합 L에 대한 부분집합 A를 고려한다. 미리 결정된 최소신뢰도 cmin에 대하여 $supp(L)/supp(A) \geq cmin$ 이면, R: $A \Rightarrow (L-A)$ 형태의 규칙을 출력한다. 즉, 이 규칙의 지지도는 supp(R)=supp(L)이며, 신뢰도는 conf(R)=supp(L)/supp(A) 가 된다[9]. 이를 바탕으로 기존의 연관규칙을 적용하면, 빈발도가

낮은 마지막 단계의 질환이 최소 지지도를 만족하지 못할 때, 정확한 서비스를 제공하지 못하게 된다. 이를 보안하기 위해서 본 논문에서는 MD-mine tree 기법을 사용한다.



(그림 2) 집단별 발병패턴 분석

(그림 2)에서 root는 집단의 구성원에 해당한다. 분류된 혈연집단에서의 위암 발병패턴과 지역집단에서의 콜레라 발병 패턴이 각 집단에서 발생할 경우 count가 1씩 증가하며 집단의 구성원 비례하여 적용된 smin보다 많을 경우 해당 패턴은 주요 발병 패턴으로서 마이닝 된다. 아래 그림은 패턴이 모아지는 과정을 나타낸 것이다.

Input: 개인 진단기록 데이터
Output: 집단별 의료검진 정보시스템

Step1. 진단기록 데이터베이스에서 진단기록을 트랜잭션 단위로 가져옴
Step2. 대상자의 진단기록을 md-mine 알고리즘을 이용하여 진단 패턴을 찾아냄
Step3. 대상이 속한 혈연집단과 지역집단의 주요 발병 패턴을 가져옴
Step4. 대상자의 진단 패턴과 각 집단의 발병 패턴을 비교하여 동일증상이 smin을 만족하고 smin을 만족하는 패턴들이 연속적일 경우 집단 발병 패턴의 최종 단계인 병을 '발병 위험성 있음' 으로 판단함
Step5. 대상자의 진단 패턴이 최종단계까지 진행 되었을 경우에는 대상자의 패턴을 의료검진정보시스템에 집단별로 저장함

(그림 3) 집단별 발병 패턴 마이닝과정

(그림 3)의 과정을 통해서 모아진 정보는 대상자가 발생할 경우 아래 표와 같이 smin을 충족하는 순서패턴을 찾아내어 의료검진정보를 제공하게 된다.

'홍'씨 가족 발병 case	발병 패턴
A case	복통, 구토, 토혈, 위암
B case	복통, 토혈, 위암
C case	복통, 위암

<표 1>'홍'씨 가족의 빈번 발병패턴(혈연)

'축산업' 지역 발병 case	발병 패턴
A case	설사, 구토, 콜레라
B case	고열, 설사, 구토, 콜레라
C case	고열, 설사, 구토, 탈수, 콜레라

<표 2>'축산업' 지역인의 빈번 발병패턴(집단)

<표 1>과 <표 2>는 집단별 의료검진시스템에서 축적된 패턴 결과의 한 부분을 예로 나타낸 것이다. '축산업' 지역에 거주하는 '홍길동'씨는 잦은 복통과 구토로 인해 건강 검진을 받고자 한다. '홍길동'씨는 자신의 진단 및 치료 내역을 의료 검진 정보시스템에 제출하고 의료검진정보시스템은 '홍길동'씨가 속한 집단의 패턴 중에서 <표 1>의 패턴에서 '홍길동'씨의 진단패턴과 일치하는 패턴을 찾는다. 의료검진정보시스템은 '홍길동'씨에게 '위암'이 발병될 수 있다는 의료검진 정보를 제공한다.

4. 결 론

본 논문에서는 집단별로 축적된 의료정보를 데이터마이닝 기법으로 개인 맞춤형 의료검진정보를 도출하여 제공하는 시스템에 대하여 제안하였다. 기존의 건강 검진 정보는 개인의 현재 건강상태에 초점을 맞추어 제공 되어왔으나 본 논문에서 다루어진 의료 정보시스템의 데이터 마이닝 기법으로 발견된 새로운 검진정보는 혈연, 지역 등의 외부적 요인까지 고려하여 각 종 질환, 질병의 조기 발견과 치료까지 가능하게 할 것으로 예측한다.

본 논문에서 제안한 의료검진정보 시스템은 개인의 의료 검진, 치료정보에 대한 선수집이 필요하며 정보들에 대한 접근 권한이 필요하다는 단점이 있다. 이러한 단점들을 보완하기 위해서는 집단구성원들의 정보제공 동의가 필요하다. 개인의 신상보호를 위해서 의료진단 정보보안기술이 추가적으로 뒷받침 되어야 할 것이다.

감사의 말

본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었으며 (IITA-2008-C1090-0801-0028), 21세기 프론티어 연구개발사업의 일환으로 추진되고 있는 지식경제부의 유비쿼터스 컴퓨팅 및 네트워크 원천 기반기술 개발사업의 08B3-B1-10M 과제로 지원된 것임.

참고문헌

[1] 한창환; 이규백; 김호성; 이병채, "의료정보(Medical Informatics); 건강정보(Health Information); 데이터베이스(Database)",1p ,2000
[2] Jiawei Han; Micheline Kamber, "Data Mining concepts and Techniques", p5-p9
[3] Dunja Mladenic, Nada Lavrac, Marko Bohanec, and Steve Moyle, "Data Mining and Decision Support

Integration and Collaboration", Kluwer Academic Publishers Boston/Dordrecht/London
[4] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rule," Proc. Int'l Conf. Very Large Data Bases, pp. 487-499, Sept. 1994
[5] Rakesh Agrawal; Ramakrishnan Srikant, "Mining Sequential Patterns", 1p, 1995
[6] Jiawei Han; Micheline Kamber, "Data Mining concepts and Techniques", p285-286
[7] 권은희; 이승철; 이주창; 김응모, "개인 맞춤형 의료진단 서비스 제공을 위한 효율적인 데이터마이닝 기법", 2p-3p, 2007
[8] 이현정, "논문 데이터 마이닝을 이용한 질병관련 유전자의 발굴",이화여자대학교, 1p, 2003
[9] Rakesh Agrawal; Ramakrishnan Srikant, "Fast Algorithms for Mining Association Rules",2p ,1994