한국의학도서관 : 제 33 권 제 1, 2 호 2006

의학도서관에서의 웹 마이닝 기술의 적용 가능성

가톨릭대학교 성의교정 도서관

정 소 나

서 론

인터넷의 확산으로 웹사이트 활용이 각 분야에 걸쳐 활발하게 이루어지고 있고 정보기술의 급격한 발전으로 정보환경도 변화하여 도서관도 소장에서 접근의 관점으로 변화하고 있다. 의학도서관도 예외는 아니다. 대표적인 의학정보센터인 National Library of Medicine (NLM)에서는 CD-ROM 혹은 상업적 정보벤더를 통해 제공해오던 MEDLINE을 2001년부터 인터넷을 기반으로 한 PubMed 방식으로 인터페이스를 개발하였다. 또한 많은 문헌정보학, 의학, 정보과학 관련 연구자들이 이용자 연구를 통해 이용자의 행태를 측정하고 미래의 정보요구를 예측하여 시스템 및 서비스에 반영함으로써 이용자의 만족을 극대화시키고자 노력해오고 있다. 국내 의학도서관에서도 최신의 정보기술을 적용하여 웹기반 도서관시스템을 구축하여 인터넷상의 웹사이트를 이용하여 많은 정보서비스를 보다 다양하고 폭넓게 그리고 시공간적 제약 없이 제공하는 환경이 되었다.

그러나 도서관 서비스에 대한 조사 및 분석 방법의 경우 설문지법이나 실험이 많은 부분을 차지하여 개선이 요구된다. 물론 데이터베이스나 전자저널의 이용통계 등을 출판사나 대행사를 통해 제공받아 이용자 분석에 사용하고 있으나 도서관 웹페이지를 통해 입력되는 대용량의 데이터분석을 지속적이고 체계적으로 측정하여 시스템에 반영할 필요가 있다. 즉, 도서관 운영과 장서구축의 개발에 필요한 의사결정을 위하여, 세분화된 이용자 집단에 적합한 맞춤 서비스를 제공하기 위하여 데이터마이닝과 같은 최신 정보기술을 통해 대량의 데이터를 효율적으로 분석하는 것이 필요하다. 이용자에 대한 중요성을 인식하고 이용자의 가치에 따른 차별화된 서비스를 제공하기 위해서는 다양하고 풍부한 이용자 관련 데이터가 가장 필요하고 이를 통한 분석 및 예측이 중요하다.

최근 주목받기 시작한 데이터마이닝 기술은 대량으로 축적되어 있는 데이터를 통해 이용자의 개입 없이 웹페이지의 분석을 통한 이용자들의 이용패턴을 측정할 수 있고 정보요구를 예측할 수 있게 한다. 본고에서는 데이터마이닝에 관한 이론의 고찰을 통해 도서관 이용자의 로그파일을 분석하고 예측할 수 있는 방법들에 대하여 살펴보고자 한다. 문헌연구를 통하여 데이터마이닝의 정의와 관련기술, 데이터마이닝 과정, 수행결과에 대하여 살펴보고 웹마이닝을 위한 로그분석의 실제적인 측정요소를 분석하고 의학 도서관내에서의 웹마이닝 적용분야를 파악하는데 의의가 있다.

이론적 배경

1. 데이터마이닝의 정의

데이터마이닝(Data mining)이란[1] 대용량의 데이터로부터 이들 데이터 내에 존재하는 패턴, 관계, 규칙 등을 탐색하고 모형화함으로써 유용한 지식을 추출하는 일련의 과정들을 의미한다. 자동화되고 지능을 갖춘 데이터베이스 분석기법으로 90년대 초반부터 지식 발견[Know-

ledge Discovery in Databases (KDD)], 정보발견, 정보수확 등의 이름들과 함께 등장하였는데 일반적으로 대량의 데이터로부터 새롭고 의미 있는 정보를 추출하여 의사결정에 활용하는 작업과 관련된 일련의 행위라 할 수 있다.

2. 데이터마이닝의 관련 분야

데이터마이닝은 새롭게 등장한 개념이 아니고 통계, 의사결정지원시스템, 데이터베이스 관리, 데이터웨어하우징, 기계학습 등에서 이론적 배경을 찾을 수 있다. Data mining, knowledge extraction, information discovery, information harvesting, data archaeology, data pattern precessing 등이 데이터에서 유용한 패턴을 발견한다는 개념을 포함하는데 이들의 공통목표는 대용량의 데이터집합으로부터 지식을 추출하는 것이다.

1) 데이터마이닝과 지식 발견[2]: 지식 발견은 인공지능이나 전문가시스템 관련연구에 주로 등장한다. 데이터로부터 유용한 정보를 발견하는 전 과정이라 정의할 수 있다. 데이터마이닝은 지식발견 프로세스 중에서 데이터로부터 정보를 추출하기 위해서 기법을 적용하는 특정단계이다. 즉, 지식발견 프로세스 중에서 데이터의 의미를 찾는 과정으로 보는 것이 타당하다.

2) 데이터마이닝과 통계학: 통계학은 집단현상을 수량적으로 관찰하고 분석하는 방법을 연구하는 학문으로 데이터마이닝과 가장 밀접한 관계를 가지고 있다. 실제로 데이터마이닝의 주요 기법인 사례기반추론방법, 의사결정나무, 군집화 방법, 인공신경망에서의 함수 등은 전통적인 통계학 이론을 배경으로 하고 있고 데이터마이닝 상용 프로그램에서도 선형회귀분석, 로지스틱 회귀분석 등의 통계기법이 포함된다.

일반적으로 통계프로그램만으로도 데이터마이닝 작업을 수행할 수 있다. 그러나 대용량의 데이터를 다루는 경우에는 통계기법보다는 데이터마이닝의 기법을 적용하는 것이 양질의 결과를 얻을 수 있다. 또한 예측정보를 생산하는 데는 사례기반추론, 의사결정나무, 인공신경망 등의 예측기법들을 적용하는 것이 더욱 좋은 결과를 얻을 수 있다.

3) 데이터마이닝과 의사결정지원 시스템[3]: 의사결정지원시스템(Decision Support System)은 기업의 의사결정을 보다 쉽게 할 수 있도록 관련 자료를 분석하여 정보를 제공한다. 데이터마이닝은 주관적인 견해를 수반하여 미래지향적인 정보전달을 이용하여 이용자의 행동을 예측함으로 전략을 수립한다.

4) 데이터마이닝과 데이터웨어하우징: 데이터마이닝을 위한 데이터는 보통 "데이터웨어하우스(data warehouse)[4]"로부터 추출하게 된다. 데이터웨어하우스는 축적된 많은 데이터를 사용자 관점에서 주제별로 통합하여 별도의 장소에 저장한 데이터베이스이다. 데이터웨어하우징은 데이터웨어하우스를 구축하고 활용하는 일련의 과정으로 의사결정을 지원하기 위하여 환경을 구축하는 것을 목적으로 한다. 데이터웨어하우스로부터 데이터를 추출하면 데이터 통합 및 정제에 대한 문제가 해결되고 유지 보수의 과정을 거치기 때문에 데이터마이닝을 위한 통합 및 정제 과정이 필요 없게 된다. 일반적으로 데이터마이닝에서 50~80%의 노력과 시간을 데이터의 통합과 정제과정에 소모하는 것을 비춰볼 때, 데이터웨어하우징을 데이터마이닝의 선행 단계로 보는 견해도 있다. 그림 1은 데이터와 데이터웨어하우스 그리고 지식베이스의 관계를 도식화하여 데이터마이닝의 시스템 구조를 파악할 수 있도록 한 그림이다.

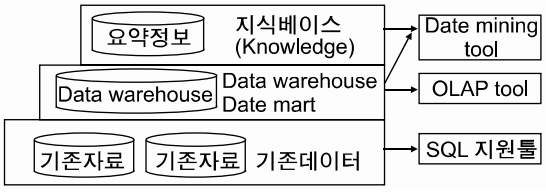

그림 1. 데이터마이닝과 데이터 웨어하우징.

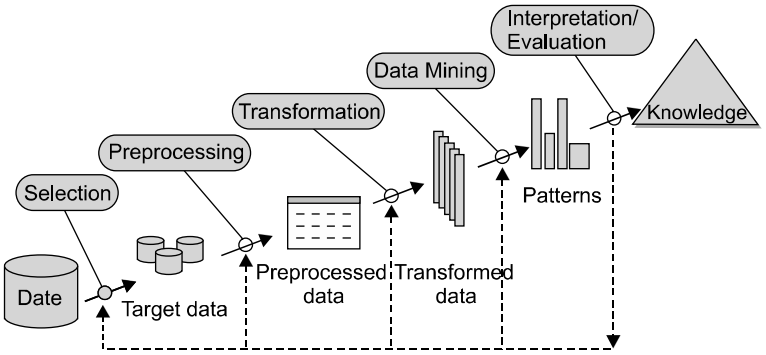

그림 2. 지식발견(KDD) 프로세스의 단계.

3. 데이터마이닝 과정

1) 지식발견 프로세스와 데이터마이닝: 데이터마이닝은 지식발견 프로세스의 한 단계로 데이터 분석과 계산 가능한 한도 내에서 데이터에 대해 특정계산을 생성할 수 있는 발견 알고리즘을 적용하는 것이다. 전체적인 지식발견 프로세스는 그림 2[5]와 같은데 패턴 가운데 새로운 "지식"으로 생각할 수 있는 것을 알아내기 위해 "마이닝"된 패턴을 평가하고 가능한 해석을 하는 작업을 포함한다. 데이터선택, 정제, 변환 등의 데이터 준비과정은 반자동화된 많은 시간을 요구하는 작업이다.[6]

데이터 준비과정이 필요한 이유는 웹에서 발생하는 원시데이터를 정제하는 작업을 통해 데이터 분석 수행을 방해하는 문제를 해결하고 데이터의 특성을 이해하여 더욱 의미 있는 데이터 분석을 수행하기 위해서이다. 이러한 데이터 정제과정이 수행됨으로 주어진 데이터 집합에서 좀 더 의미 있는 지식을 추출할 수 있다.

지식발견 프로세스는 대화식으로, 반복적으로 이루어지는 작업으로 여러 단계를 포함한다. Brachman & Anand (1996)는 프로세스의 대화식 특성을 강조하는, 지식발견 프로세스에 대한 실용적인 관점을 제시하였다. 여기에 그 기본적인 단계의 대략적인 순서를 보면 다음과 같다.[7]

① 데이터 선택: 가장 중요한 데이터를 선택하는 과정이다. 응용 도메인과 관련된 선행 지식에 대한 이해를 개발하고 고객의 관점에서 지식발견 프로세스의 목표를 확인한다.

② 목표 데이터 집합 생성: 지식 발견이 수행될 데이터 집합을 선택하거나 변수나 데이터 샘플의 부분집합에 초점을 맞춘다.

③ 데이터 정제(손질, 전처리): 필요하다면 부정확한 값, 결손 값, 불일치, 잡음을 제거하고 데이터의 범위를 벗어난 데이터 및 특이 값을 추출하는 단계이다. 잡음을 설명하거나 모델링하는데 필요한 정보 수집, 잃어버린 데이터 필드 처리에 대한 전략 결정, 시계열 정보와 변경 사항을 설명하는 작업을 한다.

④ 데이터 축소와 프로젝션: 작업의 목표에 의존하는 데이터를 표현하기 위한 유용한 특성을 발견한다. 고려해야 할 유효 변수의 수를 줄이거나 동일한 데이터 표현을 발견하기 위해 차원 축소나 변형 기법을 사용한다.

⑤ 지식발견 프로세스의 목표와 특정 데이터 마이닝 기법 대응: 예를 들어 요약, 분류, 회귀, 클러스터링 등이 있다.

⑥ 데이터마이닝 알고리즘 선택: 데이터에서 패턴을 탐색하는 데 사용되는 기법을 선택한다. 여기에는 어떤 모델과 인수가 적합하고 지

식발견 프로세스의 종합적인 기준과 함께 어떤 데이터 마이닝 기법에 대응시킬지 결정한다.

⑦ 데이터마이닝: 특정한 표현 형식이나 형식의 집합에서 관심이 있는 패턴을 탐색한다. 여기에는 분류 규칙이나 트리, 회귀, 클러스터링 등이 있다. 사용자는 이전 단계를 올바르게 수행함으로써 데이터마이닝 기법에 크게 도움이 될 수 있다.

⑧ 마이닝된 패턴을 해석하고, 반복하기 위해 1단계에서 7단계로 돌아갈 수 있다. 이 단계에서 추출된 패턴/모델을 시각화하거나 추출된 모델에 주어진 데이터를 시각화하는 작업이 포함될 수 있다.

⑨ 발견된 지식 통합 정리: 이 지식을 사후처리를 위해 다른 시스템에 통합하거나 필요로 하는 사용자에게 문서화 또는 보고한다. 여기에는 사전에 발견(추출)된 지식과 잠재적인 충돌을 검사하고 해결하는 작업이 포함된다.

지식발견 프로세스는 위의 프로세스를 여러 번 반복할 수 있고 어떤 두 단계만 반복할 수도 있다. 지식발견에 대한 이전 작업들의 대부분은(단계 7)의 데이터마이닝에 중점을 두고 있다. 하지만 다른 단계는 실질적으로 지식발견 프로세스의 응용을 성공시키는 데 역시 중요하다.[2)]

지식발견 프로세스의 데이터마이닝 요소는 특별한 데이터마이닝 방법들을 반복하는 응용을 포함한다. 지식발견의 목적은 시스템의 의도된 용도에 따라 두 가지 유형으로 정의된다. 첫째, 시스템이 사용자의 가정을 검증하는 데 제한적으로 사용되는 확인(verification)과 둘째, 시스템이 자율적으로 새로운 패턴을 찾아내는 발견(discovery)이다. 패턴발견은 더 나아가 어떤 개체의 미래 동작을 예측하기 위해 시스템이 패턴을 찾아내는 예측(prediction)과 사용자가 이해할 수 있는 형태로 나타내어 주기 위해 패턴을 찾아내는 묘사(description)로 세분할 수 있다.

2) 데이터마이닝 기법[8,9]: 데이터마이닝 기법들과 목표작업은 다음과 같다.

① 연관성 규칙발견(Association Rule Discovery, Market Basket Analysis)

② 사례기반추론(Case-Based Reasoning, Memory-Based Reasoning)

③ 군집분석(Cluster Analysis)

④ 연결분석(Link Analysis)

⑤ 판별분석(Discrimination Analysis)

⑥ 의사결정나무(Decision Tree, Rule Induction)

⑦ 인공신경망(Artificial Neural Network)

⑧ 유전자알고리즘(Genetic Algorithm)

⑨ OLAP (Online Analytic Processing)

자세히 살펴보면 다음과 같다.

① 연관성 규칙발견(Association Rule Discovery, Market Basket Analysis): 연관성 규칙발견은 고객의 거래기록 데이터로부터 상품구매의 연관성을 파악하여 연관성이 있는 것들을 그룹화하는 클러스터링의 한 종류이다. 이러한 기법을 통하여 동시에 구매될 수 있는 상품들을 찾아냄으로 시장바구니분석(Market Basket Analysis)을 효과적으로 적용할 수 있다. 여기에 사용되는 연관 규칙은 구체적 상품들을 지목하여 분석하므로 이해가 쉽고 실제 업무에 적용이 용이하다.

② 사례기반추론(Case-Based Reasoning, Memory-Based Reasoning): 주어진 새로운 문제를 과거의 유사한 사례를 바탕으로 주어진 문제의 상황에 맞게 응용하여 해결해 가는 기법이다.

③ 군집분석(Cluster Analysis): 클러스터링 기법은 속성이 비슷한 대상을 묶어서 몇 개의 군집으로 나누는 것을 목적으로 한다. 이러한 기법을 통하여 고객을 고객의 경제력, 고객의 연령과 같은 변수를 기준으로 묶어 볼 수 있다. 복잡한 대용량의 데이터를 분석하려면, 이러한 기법을 적용하여 원본 데이터를 몇 개의 그룹으로 나누어 우선적으로 분석하게 된다. 몇 개의 그룹화를 통한 분석은 전체 데이터에 대한 윤곽을 우선적으로 파악하는 데 도움이 된다.

⑥ 의사결정나무(Decision Tree, Rule Induc-

tion): 의사결정나무는 데이터마이닝의 분류 작업에 주로 사용되는 기법으로, 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 부류별 특성을 속성의 조합으로 나타내는 분류모형을 나무의 형태로 만드는 것이다. 이렇게 만들어진 분류모형은 새로운 레코드를 분류하고 해당 부류의 값을 예측하는 데 사용된다.

의사결정나무는 순환적 분할(recursive partitioning) 방식을 이용하여 나무를 구축하는 기법으로, 나무의 가장 상단에 위치하는 뿌리마디(root node), 속성의 분리기준을 포함하는 내부마디(internal nodes), 마디와 마디를 이어주는 가지(link), 그리고 최종 분류를 의미하는 잎(leaf)들로 구성된다.

⑦ 인공신경망(Artificial Neural Network): 신경망은 인간 두뇌의 신경세포를 모방한 개념으로 마디(node)와 고리(link)로 구성된 망구조를 모형화 하고, 의사결정나무와 마찬가지로 과거에 수집된 데이터로부터 반복적인 학습과정을 거쳐 데이터에 내재되어 있는 패턴을 찾아내는 모델링 기법이다.

신경망은 분류, 군집, 연관규칙 발견과 같은 작업에 널리 사용되는 데이터마이닝 기법으로 신용평가, 카드 도용패턴 분석, 수요 및 판매예측, 고객세분화(customer segmentation) 등 여러 가지 목적으로 다양한 산업분야에 폭 넓게 적용되고 있다. 그림 3은 신경망 구조의 예이다.

⑧ 유전자알고리즘(Genetic Algorithm): 선택적 도태나 돌연변이 같은 생물 진화의 원리로부터 착안된 알고리즘으로서 확률적 탐색이나 학습 그리고 최적화를 위한 한 가지 기법이다.

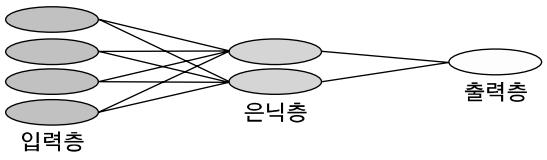

그림 3. 신경망 구조 예.

⑨ OLAP (Online Analytic Processing)[10]: 분석과 관리 목적을 위해서 다차원데이터를 모으고, 관리하고, 프로세싱하고 표현하기 위한, 응용프로그램 및 기술들의 종류이다. OLAP은 위의 정의에서와 같이 최종 사용자가 다차원 정보－사용자들에 의해 이해되는 기업의 실제차원을 반영하는 정보로 예를 들면 '이번 달 매출액이 지난달에 비해서 얼마나 상승했는가 혹은 하락했는가?' '지난 해 같은 달에 비해서 어떠한가?' '목표치를 달성했는가?' '경쟁사의 매출액과 비교해서는 어떠한가?' 등과 같은 정보－에 직접 접근하여 대화식으로 정보를 분석하고 의사결정에 활용하는 과정이다. 최종사용자는 온라인상에서 직접 데이터에 접근하며, 대화식으로 정보를 분석하므로 최종사용자가 기업의 전반적인 상황을 이해할 수 있게 하고 의사결정을 지원하는 데 그 목적이 있다고 할 수 있다.

가장 많이 사용되는 모델 표현은 의사결정나무(decision tree)와 규칙, 비선형 회귀와 분류, 사례기반추론(example-based)기법(근접-이웃(nearest-neighbor)과 사건-기반(case-based) 추론 기법이 포함하여), 확률적 그래픽 의존 모델(Bayesian 네트워크를 포함하여)과 관계형 학습 모델(귀납적 논리 프로그래밍을 포함하여)을 포함한다.[2]

3) 데이터마이닝 수행결과: 데이터마이닝을 통해 분류, 추정, 예측, 군집, 설명 등의 결과를 알 수 있다.[7]

① 분류(classification): 데이터 아이템을 여러 개의 미리 분류된 클래스로 매핑(분류)할 수 있다.

② 회귀(regression): 데이터 아이템을 실제값의 예측 변수로 매핑하는 함수의 학습과 변수들 간의 기능적인 관계를 발견할 수 있다.

③클러스터링(clustering): 범주의 한정된 집합과 데이터를 기술하는 클러스터를 인식할 수 있다.

④요약(summarization): 데이터의 부분 집합에

대한 간단한 기술을 할 수 있다.

⑤의존성 모델링(dependency modeling): 변수들 간의 중요한 의존 관계 기술하는 모델을 발견할 수 있다.

⑥ 변화, 변동 탐색(change and deviation detection): 이전에 측정되거나 규범적인 값에서 얻은 데이터 중에서 가장 중요한 변화를 발견해 낼 수 있다.

4. 웹 마이닝[11]

웹마이닝[5]은 인터넷 사용자의 사이트 방문시 남게 되는 로그데이터를 양적으로 분석하여 이를 기반으로 한 모델과 이론을 개발하는 데 중점을 둔다. 즉, 웹 마이닝이란 웹서버의 로그데이터로부터 웹 이용자의 의미 있는 접속 패턴을 발견하는 과정으로 웹으로부터 유용한 정보를 발견하고 분석하는 것으로 웹 로그의 데이터마이닝이라고 할 수 있다.

웹서버는 웹사이트 사용자의 행동 패턴에 대한 정보를 웹 서버 로그파일에 저장한다. 저장되는 정보는 IP 주소, HTTP 요청페이지, 요청에 대한 응답시간 등으로 분석에 사용될 수 있는 귀중한 정보이다. 이를 통해 이용자의 행동 패턴에서 일정한 패턴을 추출하여 이를 모델화하여 가시적으로 제시할 수 있다.

웹마이닝의 분류는 다음의 그림 4와 같이 나눌 수 있다. 웹 컨텐츠 마이닝은 사이트에 관련된 데이터만을 분석함으로써 유용한 정보를 얻어 내는 것이고 웹 사용 마이닝은 웹서버에 접속한 사람들의 접속 패턴을 발견하고 분석하는 것을 뜻한다.

웹 사용 마이닝은 웹 서버로그, 부라우저 로그, 사용자 프로파일, 쿠키 등 부차적인 데이터를 이용하여 웹 사용자의 세션과 행동으로 생성된 데이터로부터 정보를 발견하는 기법이다. 웹 사용 마이닝을 통해 다음을 수행할 수 있다.

① 웹 접근 패턴(Web Access Pattern) 분석과 웹 사용자의 네비게이션 행위를 분석한다.

② 웹상에서 사용자에 따른 최적화된 링크를 동적으로 설정한다.

③ 적응 웹 사이트를 구축한다. 즉, 이용자의 필요에 맞게 실시간으로 웹 페이지를 수정하는 개인화작업을 의미한다.

④ 웹 페이지와 사용자의 모델링, 웹페이지와 사용자의 카테고리화, 웹페이지와 사용자의 매칭을 수행하는 웹의 개인화를 설정할 수 있다. 연관규칙이나 트랜잭션 군집화 등의 기법을 이용하여 네비게이션 패턴에 기반한 URL 사이의 관계나 사용자와 비슷한 성향을 가지는 군집을 찾아내 개인화에 적용한다.[11]

5. 웹 사용 마이닝을 위한 웹 로그 분석

1) 웹 로그화일의 분석: 사용자가 어떤 사이트를 방문하는 경우 서버의 로그파일에 방문자의 기록이 남게 된다. 이러한 방문자의 정확한 데이터를 기반으로 웹사이트 내에서의 이용이 빈번한 페이지나 사용자의 이용패턴을 파악할 수 있다.

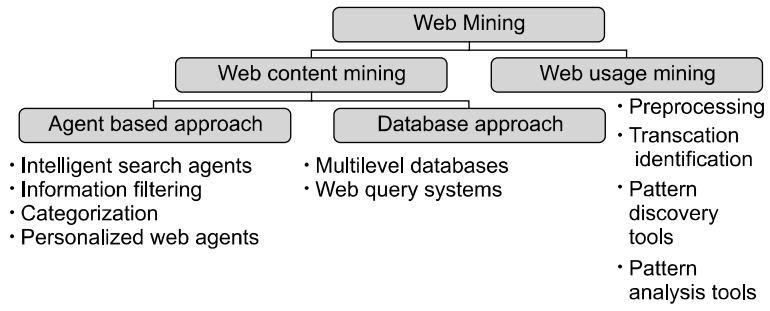

그림 4. 웹 마이닝의 분류.

로그 분석 데이터는 서버에 기록된 사이트 방문자 데이터를 로그분석 툴을 활용하여 각 조건(시간별, 월별, 일별)에 의하여 정리한 것으로 웹서버에 따라 한 개가 아닌 여러 개의 로그파일이 다양한 형태로 생성된다. 여기에는 액세스 로그, 에러로그, 레퍼럴 로그, 에이전트 로그 등이 있다.

2) 웹 로그파일 종류[11]

(1) 엑세스 로그; 부라우저가 서버에 파일을 요청한 기록을 시간과 IP 등의 정보와 함께 남긴 것이 엑세스 로그이다. 서버로부터 브라우저에 파일을 전송한 기록이므로 트랜스퍼 로그(Transfer Log)라고도 한다.

(2) 에러 로그; 웹 서버의 오작동에 대한 모든 정보를 포함한다. 파일이나 이미지들을 잘못 링크한 경우, CGI 프로그램이 정상적으로 작동하지 않는 경우, 서버의 Permission설정을 제대로 부여하지 않은 경우 등을 알 수 있다.

(3) 레퍼럴 로그; 페이지를 보기 위해서 어떤 경로를 거쳐 왔는지에 대한 기록을 알 수 있다. 사용자가 해당 사이트에서 어떤 페이지를 보았는지 알 수 있는 것은 엑세스 로그의 경우이다.

(4) 에이전트 로그(Agent Log); 사이트를 접속하는 방문자의 웹 브라우져 타입 및 웹 브라우저 타입 및 버전, 운영체제의 종류, 등에 대한 정보를 제공해 웹 사이트를 구성할 수 있는 단서를 제공한다.

3) 웹 로그파일 측정요소[12]: 로그파일 분석을 통해 측정할 수 있는 요소는 다음과 같다.

(1) 히트(Hit); 방문자가 웹사이트에 접속하여 한 페이지를 전송할 때 그 안에 포함된 그래픽, html 등의 모든 파일의 숫자를 말한다.

(2) 페이지 뷰(Page view); 방문자가 요청한 웹페이지 숫자로 보통 htm, html, asp, php 파일 등을 의미한다. 예를 들면 소장자료 검색화면에서 상세화면의 간략정보를 열어보았다면 페이지 뷰 1회가 된다.

(3) 체류시간; 한 방문자가 웹페이지에 머무른 시간이다.

(4) 세션(Session); 세션은 한 방문자가 특정 웹사이트에서 다른 사이트로 이동하는 것을 측정한 것이다.

(5) 사용자(visitor); 특정 웹사이트에 한번 이상 접속한 사용자 수를 파악하는 방법이다. 이는 방문자의 증가추이 및 자주 접속하는 이용자등을 파악할 수 있게 한다. 각 측정요소들은 다음의 표 1 항목 간의 연관성 규칙분석을 위한 정제된 데이터의 예와 같이 정보를 수집할 수 있다. 표 1은 웹페이지의 항목 간 연관성 규칙을 분석하기 위하여 실제로 수집한 로그파일로 고객 개개인이 웹페이지를 통해 수행한 각종 검색에 관한 정확한 데이터이다. 그러나 단순한 방문자의 방문기록정보로써 이용자 그룹 간 이용분석을 하기 위해서는 좀 더 상세한 정제 방법과 데이터의 정제 이후 로그데이터를 데이터 마이닝 분석툴을 사용하여 연관성 규칙분석을 수행하여야 한다. SAS 8.1 Enterprise Miner 혹은 SPSS(사)의 clementime 등이 분석에 사용되는 마이닝 툴이다.

의학도서관에서의 웹 마이닝 기술 적용

1. 의학도서관의 특성

의학도서관은 의학 관련 정보를 수집, 정리, 분석, 축적, 보존하는 일련의 활동을 통하여 병의원, 의과대학, 의학 연구소 등의 모기관의 연구자와 직원, 학생에게 최적의 정보를 제공하는 전문도서관이다. 의학도서관의 특성은 첫째, 교수, 연구원, 학부생 및 대학원생, 전공 수련의들의 학습과 교육지도를 지원해주어야 한다. 둘째, 연구 활동 지원 측면으로 이용자들의 학술 및 연구에 필요한 정보를 조직하고 연구 활동에 필요한 각종서비스를 지원하여야 한다. 셋째, 진료업무 측면에서 의학에 종사하는 사람들이 환자진료에 전념할 수 있도록 의학자가 임상에 필요로 하는 정보를 수집하고 제공할 수 있어야 한다.[13]

2. 의학도서관에서의 웹 페이지를 이용한 정보서비스[14)]

정보서비스는 의학도서관에서 제공하는 다양한 도서관 서비스 가운데 핵심을 이룬다. 의학도서관에서는 특히 인터넷을 통한 정보의 공유와 정보서비스를 기반으로 하여 이용자에게 필요한 각종 자료와 정보를 축적하여 효율적으로

표 1. 항목 간의 연관성 규칙분석을 위한 정제된 데이터의 예

TIME	IP	URL	접속자 IP	Status code/ Transfer vol.	참고사항
4:20:23 AM	61.78.109.61	GET/Default.asp - 80 -	221.148.21.190	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+InfoPath.1) 200 0 0	kornet 이용자 (성남분당)
4:20:24 AM	61.78.109.61	GET/majin.htm - 80 -	221.148.21.190	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+InfoPath.1) 404 0 2	kornet 이용자 (성남분당)
4:20:24 AM	61.78.109.61	GET/css/maestro.css - 80 -	221.148.21.190	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+InfoPath.1) 200 0 0	kornet 이용자 (성남분당)
4:20:24 AM	61.78.109.61	GET/header.js - 80 -	221.148.21.190	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+InfoPath.1) 200 0 0	kornet 이용자 (성남분당)
4:20:25 AM	61.78.109.61	GET/search/kor/detail_query.asp page_id=1 80 -	221.148.21.190	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+InfoPath.1) 200 0 0	kornet 이용자 (성남분당)
4:20:38 AM	61.78.109.61	GET/search/kor/simple_query.asp - 80 -	221.148.21.190	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+InfoPath.1) 200 0 0	kornet 이용자 (성남분당)
4:20:46 AM	61.78.109.61	GET/search/kor/search_connect.asp backdepth=1&max_srch=1000000&page=1&area0=all&material=all&history=key_word&graph=%B4%DC%BC%F8%B0%CB%BB%F6&query0=%C0%DB%B9%AE&range=20 80 -	221.148.21.190	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+InfoPath.1) 302 0 0	kornet 이용자 (성남분당)
4:20:46 AM	61.78.109.61	GET/search/kor/search_result.asp backdepth=1&history=key_word&query0=작문&range=20&page=1&area0=all&material=all&graph=단순검색&max_srch=3 80 -	221.148.21.190	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+InfoPath.1) 200 0 0	kornet 이용자 (성남분당)

표 1. Continued

TIME	IP	URL	접속자 IP	Status code/ Transfer vol.	참고사항
4:21:00 AM	61.78.109.61	GET/search/kor/DetailInfo.asp history=key_word&query0=작문 &range=20&max_srch=3&area0=all& material=all&graph=단순검색&control_ no=7129919&backdepth=2&mctp= km&page=1&recno=2&artlist=0&r_title= 작문신공%20%20:%20나의%20생각 을%20디자인하는%20 글쓰기 80 -	221.148.21.190	Mozilla/4.0+(compatible; +MSIE+6.0;+Windows+ NT+5.1;+SV1;+InfoPath. 1) 200 0 0	kornet 이용자 (성남 분당) 상세 화면
4:21:11 AM	61.78.109.61	GET/search/kor/search_connect.asp backdepth=3&history=key_simple& range=20&page=1&graph=단어검색 &area0=all&yearto=&material=all&y=5 &x=56&query0=글쓰기 80 -	221.148.21.190	Mozilla/4.0+(compatible; +MSIE+6.0;+Windows+ NT+5.1;+SV1;+InfoPath. 1) 302 0 0	kornet 이용자 (성남분당)
4:21:11 AM	61.78.109.61	GET/search/kor/search_result.asp backdepth=3&history=key_simple& query0=글쓰기&range=20&page= 1&graph=단어검색&area0=all&yearto =&material=all&y=5&x=56&max_srch =10 80 -	221.148.21.190	Mozilla/4.0+(compatible; +MSIE+6.0;+Windows+ NT+5.1;+SV1;+InfoPath. 1) 200 0 0	kornet 이용자 (성남분당)
≈≈≈					
3:47:14 AM	61.78.109.61	GET/search/kor/ cirlist_result.asp - 80 -	160.1.7.43	Mozilla/4.0+(compatible; +SIE+6.0;+Windows+ 98) 200 0 0	내부 이용자 대출 도서 조회 및 연기 화면
3:47:14 AM	61.78.109.61	GET/css/ maestro.css - 80 -	160.1.7.43	Mozilla/4.0+(compatible; +MSIE+6.0;+Windows+ 98) 200 0 0	내부 이용자
3:47:14 AM	61.78.109.61	GET/header.js - 80 -	160.1.7.43	Mozilla/4.0+(compatible; +MSIE+6.0;+Windows+ 98) 200 0 0	내부 이용자
3:47:14 AM	61.78.109.61	GET/Men_info/men_info/ men_info.htm - 80 -	160.1.7.43	Mozilla/4.0+(compatible; +MSIE+6.0;+Windows+ 98) 200 0 0	내부 이용자 개인 정보 로그인 화면
3:47:22 AM	61.78.109.61	POST/include/kor/ logincheck.asp - 80 -	160.1.7.43	Mozilla/4.0+(compatible; +MSIE+6.0;+Windows+ 98) 302 0 0	내부 이용자
3:47:22 AM	61.78.109.61	GET/search/kor/ cirlist_result.asp page_id=3 80 -	160.1.7.43	Mozilla/4.0+(compatible; +MSIE+6.0;+Windows+ 98) 200 0 0	내부 이용자 대출 도서 조회 및 연기 화면
3:47:33 AM	61.78.109.61	GET/search/kor/ rcustlist_result.asp - 80 -	160.1.7.43	Mozilla/4.0+(compatible; +MSIE+6.0;+Windows+ 98) 200 0 0	내부 이용자 예약 도서 조회

이용할 수 있게 하고 정보를 시간과 장소를 불문하고 실시간으로 제공하고 있다.

의학도서관의 경우 의학이라는 주제 분야 내에서 교육, 연구, 진료측면이 폭넓게 분포하고 있고 이용자 대상에 있어서도 연령 및 지적수준 등 다양한 이용자 정보추구 행태를 보인다. 따라서 이용행태에 대한 지속적이고 정확한 정량적 분석과 함께 서비스과정에 대한 평가가 필요하다. 이용행태에 대한 분석에 웹마이닝 기술을 적용하는 것은 웹페이지를 통해 발생하는 로그파일로부터 이용관련 데이터를 쉽게 대량으로 수집할 수 있어 유용하다. 또한 로그파일로 부터 정확한 데이터가 수집될 수 있으므로 서비스 과정 및 결과에 대한 평가가 전통적 정보서비스의 경우보다 용이하다. 따라서 트랜잭션 로그의 정기적 분석을 통하여 서비스의 성공정도를 가늠하고 실패요인을 구명하는 것 이외에 이용행태의 파악과 소요시간 등의 효율을 측정할 수 있다.[15]

웹마이닝 기술 적용가능성을 도서관 웹페이지 사용성 측면과 Customer Relationship Management (CRM) 기법을 적용한 이용자관련 정보의 재구축 측면에서 살펴보고자 한다.

1) 도서관 웹페이지 사용성: 도서관에서 웹페이지로 제공하고 있는 서비스는 다음과 같다.

(1) 온라인 소장 목록; 소장자료의 검색(단행본, 연속간행물, 학위논문, 멀티미디어자료)과 전자정보원(e-journal, web DB, e-book, VOD)의 검색

(2) 도서관 서비스: My library [개인정보관리, 대출현황(개인화정보)], 도서구입신청, 이용자 맞춤 주제정보서비스, 상호대차, 온라인 참고정보원, 공개자료실, 온라인 이용자 교육, 신착자료안내, 학위논문 원문제공, 지정도서 서비스

(3) 도서관 안내 및 게시판: 도서관안내, 공지사항, 자료실, 도서관 통계

도서관 웹페이지 중 온라인 소장목록은 사서가 이용자의 검색 행태를 분석하여 의사결정을 해야 하는 웹페이지 중 하나이다. 현재 자관에서 소장하고 있는 인쇄 형태의 소장 자료와 전자자원의 자료를 별도로 찾게 하기 위하여 전자자료 통합목록을 제공하고 있는 도서관유형과 웹 DB, VOD 자료, e-book 및 e-journal을 통합하여 온라인 소장목록을 통해 제공하는 도서관 유형이 있다. 도서관의 웹서버에 저장되는 로그파일에 웹마이닝 기술을 적용하여 분석하면 이용자들의 검색행태에 기반한 화면구성을 user friendly하게 구축할 수 있다. 이를 위해서는 웹페이지 구성 시 로그파일을 생산할 수 있도록 미리 디자인하는 지혜도 필요하다. 또한 도서관의 웹페이지를 통해 제공되는 서비스들의 대부분은 A-to Z의 자모순에 의한 디스플레이를 하고 있거나 게시판의 형태로 이용자가 직접 Key in을 하는 페이지들이다. 온라인 소장목록의 경우 이용자의 인적, 심리적, 교육적 특성과 상황에 따라 다양한 검색행태를 보인다. 따라서 사서들은 여러 가지의 검색기법 및 브라우징, 혹은 웹페이지의 구성을 많은 이용자가 쉽게 사용할 수 있도록 최적화시킬 필요가 있다.[15] 설문조사나 벤치마킹을 이용한 전통적인 분석방법보다는 소속기관 이용자의 특성에 근거하여 정량적이고 정확한 분석을 위해 데이터마이닝 기술을 적용하는 것이 바람직할 것으로 생각한다. 보다 적극적인 방법으로는 데이터마이닝 툴을 이용한 프로그램을 도입하여 웹페이지의 분석 및 다양한 정보검색 기법의 사용, 그리고 각종 검색을 표현하는 용어의 선택 시에 데이터마이닝 기술을 적용하여 반영할 수도 있다.

2) CRM 구현을 위한 이용자관련정보의 재구축: CRM[16]은 풍부한 고객관련 데이터를 다양한 정보기술을 활용하여 분석함으로써 기업과 고객간의 상호교류를 관리하는 고객중심의 마케팅적 경영방식을 말한다. 그림 5는 CRM의 개념도로 기업은 다양한 채널을 통하여 발생된 고객정보를 데이터 마트와 데이터 마이닝과 같

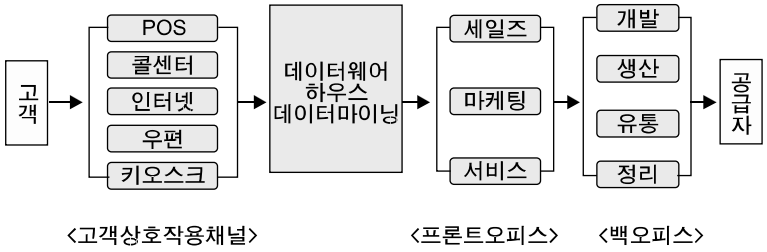

그림 5. CRM의 개념도

은 정보기술에 통합, 분석하여 고객의 기호, 가치 요구사항, 문제점을 적절하고도 차별화된 서비스를 제공한다.

도서관은 장서, 시설, 사서라는 인적, 물적 자원을 기반으로 이용자의 요구를 충족시키고자 적절하고도 차별화된 서비스를 제공해왔다. 그러나 이용자인 고객을 중심으로 체계적이고 논리적이며 근거에 입각하여 장서를 개발하고 관리하여 고객의 정보요구, 정보이용행태를 파악하는 등의 이용연구를 실제 도서관 현장에서 수행하고 적용시키는 도서관은 많지 않다.

도서관을 활발히 이용하는 고객, 이용을 중단한 고객, 이용 가능성이 있는 고객의 행태를 파악하여 각 집단에 적합한 맞춤 서비스를 제공하기 위하여 참고서비스, 웹서비스, 장서관리, 이용자 관리 등에 CRM 기법을 적용해야 한다. 다음의 3단계를 통한 CRM 시스템의 구축을 통해 얻어진 분석데이터를 도서관 정책입안 및 정보서비스 개발에 활용하는 것이 가능하다.

(1) CRM 전략 수립
- 이용자 데이터 확보방안 수립
- 외부 데이터 활용방안 수립
- CRM 구축/활용방안 수립

(2) 데이터웨어하우스 구축을 위한 방안수립
- 데이터웨어하우스 기반 구축
- 고객 DB 구축

(3) CRM 시스템 구축
- 데이터 분석지원 시스템 구축(잠재고객 확보, 제공서비스의 차별화, 이용자 성향 및 행태 분석)
- 데이터 마이닝을 이용한 지능형 분석시스템 구축
- 적정한 채널에 대한 캠페인 전개시스템 구축

CRM을 성공적으로 수행하기 위해서는 다양한 접점채널을 통하여 가능한 많은 정보를 수집하는 것이 필요하다. 즉, 이용자파일, SDI 서비스, 상호대차 신청서, 희망도서 신청서 등을 통해 고객의 정보를 수집할 수 있는 방안을 마련하여야 한다. 또한 이용자의 대출, 참고질의 등과 같은 과거이력을 이용자 집단과 연관하여 프로파일을 구축하여 활용하여야 한다. 로그파일을 통한 정제된 데이터와 각종 고객정보는 고객중심으로 재구성된 데이터웨어하우스, 데이터마크로 통합하여 고객에 대한 대응, 개선방안을 위한 분석 및 평가에 활용할 수 있다.

결 론

본 연구에서는 의학도서관의 정보서비스 개선을 위한 데이터마이닝의 적용가능성을 살펴보았다.

로그파일의 분석은 설문지법이나 관찰 등을 통해서 파악할 수 없는 이용자 그룹에 대하여 혹은 연구법과 병행하여 정보이용 및 검색행태를 조사하기에 적합하다. 이용자의 행동 패턴 추적이 어렵고 많은 수의 이용자를 지원할 인력, 비용, 설비 등의 부담이 커지는 경우, 또한 정보 수집과 유지가 어려운 경우 웹서버에 저장되는 대용량의 로그파일을 분석할 수 있다. 따라서 이용자 집단의 연령 및 학력수준이 다

양하고 직접 방문보다는 웹페이지를 통해 양질의 서비스를 원하는 의학도서관의 경우 적용가능하다.

데이터마이닝은 이용자의 필요, 요구, 태도에 더욱 근접한 서비스 모델을 모형화할 수 있게 한다. 따라서 의학 도서관에 있어서 정책을 결정하거나 서비스를 개선할 경우 이용자의 행태와 관련된 패턴을 발견하여 웹페이지 및 정보서비스 컨텐츠 모델링에 적용할 수 있다.

의학도서관 환경에 있어 로그파일의 분석을 포함한 데이터마이닝 분야의 적용은 이용자의 개인화 서비스에 긍정적인 효과를 가져 올 수 있다. 웹페이지를 통한 정보서비스의 개선을 위해 그리고 이용자 중심의 CRM 기법을 적용하기 위하여 사서는 매일매일 발생되는 많은 트랜잭션 파일들과 다양한 이용자 정보의 분석을 통해 이용자 개개인에 적합한 서비스를 개발하고 시스템 환경을 개선하여야 할 것으로 생각한다.

참 고 문 헌

1) 최종후, 한상태, 강현철 외. 데이터마이닝: 기능과 사용법. 3rd ed. 서울: 자유아카데미 2001.
2) Fayyad UM, Piatetsky-Shapiro G, Smyth P, et. al. Advances in Knowledge discovery and data mining. California: MIT Press; 1996;1-34
3) 신승중, 김순곤, 박인규. 데이터마이닝을 이용한 마케팅 전략의 분석 비교에 관한 연구. 한국통신학회 학술발표회 논문집 1997;16(1):656-9.
4) Nicholson S. Bibliomining for Automated Collection Development in a Digital Library Setting: Using Data Mining to Discover Web-Based Scholarly Research Works. Journal of the American Society for Information Science and Technology 2003;54(12):1081-90.
5) 김성희, 이수연. 데이터마이닝기법을 이용한 검색엔진의 검색효율성 측정에 관한 연구. 한국도서관·정보학회지 2000;31(4):191-212.
6) 이기욱, 성창규. 데이터마이닝 기법을 이용한 추천 시스템의 구현. 한국 컴퓨터정보학회 논문지 2006;11(1):293-300.
7) Chang CC, Chen RS. Using data mining technology to solve classification problems: a case study of campus digital library. The Electronic Library 2006;24(3):307-21.
8) 강현철, 한상태, 최종후 외. 데이타마이닝: 방법론 및 활용. 3rd ed. 서울: 자유아카데미 2002.
9) Nicholson S. Bibliomining for Automated Collection Development in a Digital Library Setting: Using Data Mining to Discover Web-Based Scholarly Research Works. Journal of the American Society for Information Science and Technology 2003;54(12):1081-90
10) Rozic-Hristovski A, Hristovski D, Todorovski L. Users' information-seeking behavior on a medical library Website. J Med Libr Assoc 2002;90(2):210-7
11) 손용배. 웹 마이닝을 이용한 도서관 홈페이지의 사용성 평가에 관한 연구. 청주: 충남대학교 대학원 2002. 60 p.
12) Cohen LB. A Two-Tiered Model for Analyzing Library Website Usage Statistics, Part 1: Web Server Logs. portal: Libraries and the Academy. 2003;3(2):315-26.
13) 김상준. 의학분야 WEB-DB 품질평가 : pubmed와 embase를 대상으로 한국문헌정보학회지 2004;38(2):161-188.
14) 조찬식, 한혜영 웹페이지를 통한 의학도서관의 정보서비스 실태조사. 정보관리학회지 2005;22(2):87-101.
15) 장혜란. 디지털참고봉사의 이용 활성화 방안. 한국도서관·정보학회지 2004;35(4):215-28.
16) 박여원. CRM기법의 전문도서관 적용 방안에 관한 연구. 정보관리연구. 2004;35(1):51-69.