데이터마이닝 기법을 이용한 지능형 학회 관리 시스템 설계 및 구현

조영기*, 백성욱, 김상수, 조주상, 장인엽, 장철호
세종대학교 전자정보공과대학
joyungki@sju.ac.kr, sbaik@sejong.ac.kr {sskim, jscho, s022403, s022405}@sju.ac.kr

Design and Implementation of Intelligent Institute Management System Using Data Mining

Yung Ki Jo, Sung Wook Baik, Sang Soo Kim, Ju Sang Cho, In Yeob Jang, Chul Ho Chang
College of Electronics & information Engineering, Sejong University

요 약

본 논문에서는 학회사이트의 중요 정보들을 효율적으로 관리하기 위해 구축된 지능형 학회 관리 시스템의 설계 및 구현사례를 제시한다. 시스템 운영을 지원 하기위해 회원정보, 기업정보, 논문분야 정보 및 논문 정보 등의 데이터를 기반으로 데이터마이닝을 수행했으며 데이터마이닝 과정에서 나타난 여러 유용한 규칙들을 제시했다. 분석된 정보를 이용해 회원 위주의 학회 사이트 운영정책과 동적 인터페이스를 제공하기 위한 웹 사이트의 개인화 계획을 제시하였다.

1. 서 론

오늘날 디지털 정보기술의 발달로 정보관리와 활용에 대한 인식이 높아지면서 효과적인 정보관리와 정보 활용방안에 대한 연구가 활발해지고 있다. 기업들은 신속하고 정확한 마케팅 전략과 여러 가지 상황에 대한 적절한 의사결정을 위한 의미있는 고급 정보 혹은 지식들이 필요할 수 밖에 없다. 이러한 디지털 정보 욕구를 만족시키기 위해 다양한 연구 및 활동을 유도하는 곳이 학회 사이트이다.

웹상에서는 사이트 간에 이동이 쉽고 사용자들은 유동성이 강함으로 사용자들의 요구사항과 특징들을 잘 이해해서 사용자에 따른 개인화 방안을 강구하는 것은 매우 중요하다. 이는 실시간으로 축적된 데이터를 데이터마이닝 기법으로 패턴을 분석함으로서 가능하다[1]. 데이터마이닝 기법을 이용한 지능형 개인화 엔진을 학회 관리 시스템에 활용하기 위해 회원들의 성향을 파악하여 회원들에게 필요한 콘텐츠와 정보를 제공한다면 더욱 적극적인 관심을 보일 것이다.

본 시스템은 학회 회원 정보 관리, 회원과 연관된 기업 및 논문 정보 관리, 회원 등급 및 회원 데이터를 데이터마이닝 기법에 의한 회원 개개인의 개인화된 서비스 제공 등이 목적이다. 이를 통해 활발한 학회 활동 유도 및 다양한 분야와 연계하여 학회에 풍부한 컨텐츠를 보유하고 효율적으로 관리하고자 한다.

본 논문에서는 학회 사이트의 지능형 웹 콘텐츠 관리 시스템의 설계 및 구현 사례를 제시한다. 학회 관리 시스템을 통해 수집한 회원들의 사용기록에 대한 데이터들과 학회 및 기업 등의 데이터베이스를 통해 얻은 회원의 세부 정보들을 데이터마이닝 과정에 따라 분석하여 여러 유용한 패턴을 발견했다. 분석된 정보를 기반으로 회원 위주의 학회 사이트 운영정책과 동적 인터페이스를 제공하기 위한 웹 사이트 개인화(Personalization) 계획을 제시한다.

2. 시스템 설계 및 구현

2.1 시스템의 개요

지능형 웹 콘텐츠 관리 시스템은 사용자들이 인터넷을 통해서 손쉽게 각종 콘텐츠를 접할 수 있도록 구현되었다. 주요기능으로는 회원관리 기능, 논문 열람기능, 작업흐름 관리 기능, 개인화 기능 등이 있다. 회원은 등급별로 나누어 정회원, 준회원, 학생회원, 종신회원, 특별회원, 단체회원, 임원 등으로 나누어 관리하며 각각 회원 별로 제공되는 기능과 콘텐츠가 상이하다. 이처럼 각 회원별로 차별화된 서비스를 제공함으로서 Contents 열람자, 저작자, 편집자, 승인자들과 어플리케이션 사이의 작업과정을 자동화할 수 있도록 설계했다. 이를 통해 각 사용자들의 역할과 책임 기반의 관리 시스템을 구축하였다. 논문에 대한 정보를 XML형태로 구조화 시켜 파싱(parsing)을 통해 제공하며 콘텐츠가 제공되는 메뉴 구조, 논문이 표현되는 디자인 요소 그리고 실제 제공되는 논문내용을 분리해서 관리하기 때문에 동일한 논문을 상이한 View를 통해 출판할 수 있도록 디자인 요소와 콘텐츠를 분리했다.

2.2 기능

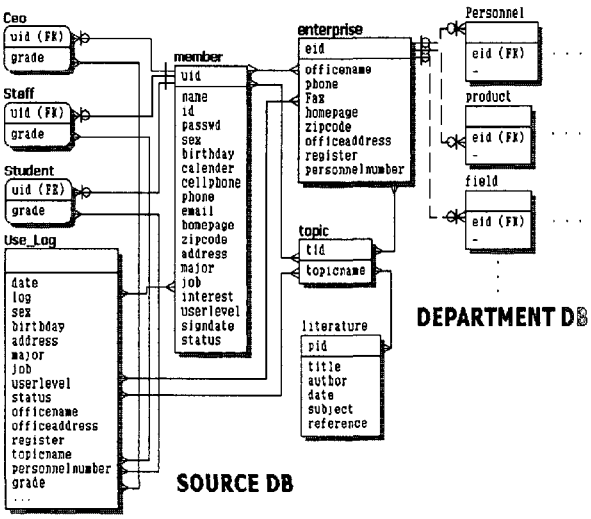
현재 구축(http://www.dcs.or.kr/memberdb/)되어 있는 기능으로는 회원가입 시 아이디 중복검사나 우편번호자동검색 기능을 다른 회원관리 시스템과 마찬가지로 추가시켰고, 로그인 기능에서는 로그인후 서버에서 회원 아이디를 바탕으로 회원등급을 확인한 후 앞서 언급한 것처럼 스크립트 처리기를 통해 회원의 등급별로 상이한 메뉴구조와 차별화된 서비스를 제공한다. 예를 들어 회원의 등급이 '임원'인 회원 즉, 학회 관리자가 로그인을 하게 되면 회원이 로그인해도 볼수 없는 [관리자 페이지]라는 링크가 추가해진다. [관리자 페이지]에는 전체회원의 정보검색, 정보수정 및 삭제기능을 가지고 있어 관리자들이 회원의 정보를 관리할 수 있도록 제공하고 있다. 그 밖에도 각 회원 등급별로 특성화된 메뉴구조를 보여준다. 또한 웹 메일러를 구축하여 관리자들은 여러 회원들에게 공지사항 등을 신속하게 알릴 수 있게 구축되어 있다. 다음으로 비밀번호 분실시 대게 회원관리 시스템에서는 서버에서 자동으로 처리를 하게 되어있으나 본 시스템에서는 더욱 보안성을 강화하기 위해 회원들은 관리자와 전자우편을 통해서만

처리되도록 설계되었다. 정보 검색엔진은 원하는 논문 정보를 사용자가 빠르게 접근할 수 있도록 하기 위해 구현된 내용으로 키워드를 통한 검색이다. 정보검색 알고리즘은 다음과 같다.

```
<? # equipment 테이블에 저장된 데이터를 불러온다. #
if ($_POST[classification]=='all') {
        $query = "SELECT distinct title, s_number,
            author, classification, reference, date,
            subject, publication_date  FROM literature
        where title LIKE '%$S_POST[title]%' group by title"; }
else {  $query = "SELECT distinct title, s_number,
            author, classification, reference, date,
            subject, publication_date  FROM literature
            where title LIKE '%$S_POST[title]%'
            AND classification='$_POST[classification]'
            group by equip_name"; }
$result = mysql_query($query); ?>
```

또한 웹에 대한 사용자의 접속과 행동에 대한 기록은 Server에서 USE_LOG 테이블에 기록하여 관리자 페이지를 통해 리포팅 되도록 했다.

2.3 데이터베이스 설계



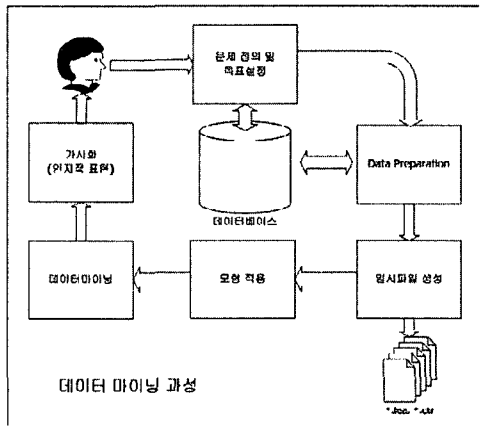
[그림 1] 데이터마이닝을 위한 데이터베이스 모델링

[그림 1]은 데이터마이닝을 위한 데이터베이스 모델링 사례이다. SOURCE DB 즉, 회원기본정보(member), 학회에서 보유하고 있는 기업정보(enterprise), 관련분야정보(topic), 논문 (literature) 등의 학회가 이미 보유하고 있는 DB이고 DEPARTMENT DB가 회원별 세부 정보를 얻기 위해 기업 데이터베이스에 있는 직원(Personnel), 생산품(product), 종사분야(field) 테이블을 기업(enterprise) 테이블에 조인한 각 기업이 가지고 있는 DB이다. Use_Log 테이블이 실제 데이터 마이닝에 쓸 data들을 가지고 있는 테이블로, 앞에서 말한 여러 테이블의 정보 중 데이터마이닝에 유효한 정보들만으로 구성된 테이블이다.

3. 데이터마이닝을 통한 규칙발견

데이터마이닝은 방대한 데이터를 기반으로 데이터웨어하우스나 데이터마트안에 저장되어 있는 대량의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다[2].

지능형 학회 관리 시스템을 통해서 얻어진 데이터를 기반으로 각각의 패턴을 분석해서 회원 관리의 효율성 증진을 위한 데이터 마이닝 사례를 제시한다.



[그림 2] 데이터마이닝 수행 과정

3.1 문제 정의 및 목표 설정

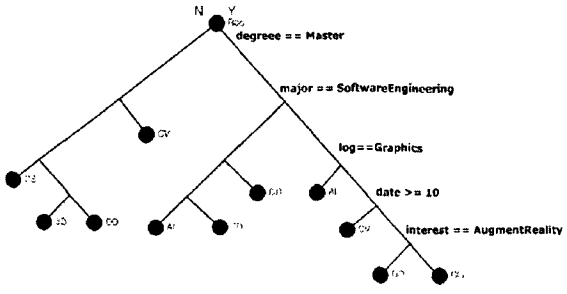
데이터마이닝을 위해 디지털 콘텐츠 학회 DB의 레코드를 사용했다. 디지털 콘텐츠 학회 사이트는(http://www.dcs.or.kr) 학회의 수많은 회원들이 방문을 하며 이들 대부분은 디지털 콘텐츠 분야에 대한 연구와 개발에 대해 큰 관심을 가진 사람들이다. 논문 투고 및 심사와 출판 그리고 웹을 통한 배포는 학회에서 매우 신중한 판단이 요구되고 있어서 제한적인 관리가 이루어지기 쉽다. 따라서 각각 사용자 특성에 맞는 사용자 위주의 적극적인 관리 시스템의 운영이 필요하다. 앞으로 제시할 데이터마이닝의 목적은 관심분야와 웹 활동 기록을 토대로 한 개인화를 구현함으로서 적극적이고 능동적인 학회 운영이 될 수 있도록 하는 것이다.

3.2 마이닝을 위한 준비 작업(Data Preparation)

데이터마이닝에 사용될 데이터의 선택은 앞서 설명한 것처럼 질의를 통해 Use_Log 테이블에 기록했다. Use_Log테이블에 적재하는 과정에서 발생한 결손 값을 제거하기에 앞서 회원들의 세부 정보와 프로필 정보가 일치하지 않는 레코드를 먼저 선별했다. 이 과정에서 많은 레코드가 줄어들었으며 결손값을 가지는 원인은 많은 필드가 필수 항목이 아닌 경우 임시로 값을 입력하거나 누락시키는 경우가 많았다. 또 데이터베이스의 메타데이터를 기반으로 한 탐색과정에서 나타난 값들을 수치화 했다. 수치화 한 데이터 타입들은 VARCAHR, CHAR, DATE 등이다. 수치화된 데이터들은 가시화 단계에서 원래의 형태로 복구했으며 데이터베이스 내의 모든 데이터 들을 임시파일(*.lea,

*.ctr)로 생성하여 마이닝을 위한 준비 작업을 끝냈다.

3.3 모형화(Modeling)



[그림 3] 의사결정나무를 통한 분류과정

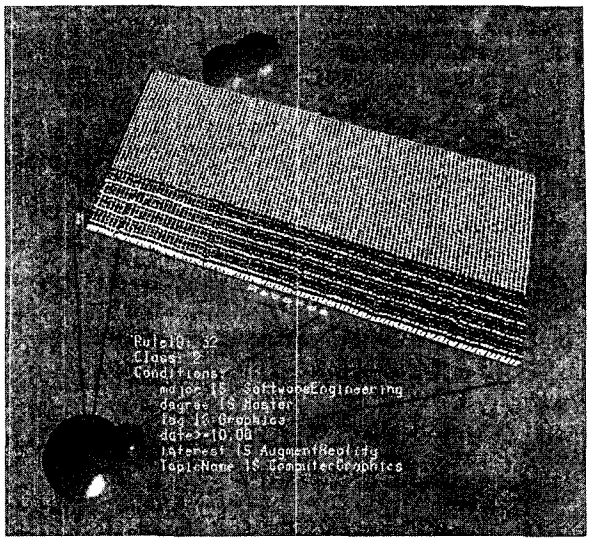
모형화 단계에서는 추출된 모델이나 지식들을 아주 잘 표현할 수 있는 장점이 있기 때문에 의사결정나무(Decision Tree)모형을 적용했다. 먼저 회원의 가입정보 및 디지털컨텐츠 학회 사이트 활동 log 정보에 대한 탐색 및 변형단계를 거친 후 모형화 하여 데이터마이닝을 수행 한다. 다음은 모형화를 통해 분류된 여러 군집중 하나의 규칙을 나타낸다.

IF(degree='Master')
AND(major=='Software Engineering')
AND(log=='Graphics')
AND (date>=10)
AND(interest=='Augment Reality')
RESULT-> 필요한정보=='Information about Computer Graphics'

분류규칙을 보면 회원의 major(전공)이 Software Engineering 이고 degree(학위)가 석사이며 디지털컨텐츠 학회 log(방문기록)에 그래픽관련 정보를 많이 이용하고 date(방문빈도)가 10회 이상이면서 또한 interest(관심분야 및 취미)가 Augment Reality(증강현실)인 회원이 필요로 하는 정보가 컴퓨터 그래픽이라고 분류된 예를 볼 수 있다. 여기서 주의할 점은 log(방문기록)에는 회원이 이용한 다양한 서비스들이 기록되는데 그 중 가장 많이 이용한 서비스를 추출한 것이고 date(방문빈도)는 다른 조건이 다 만족하더라고 방문빈도가 적으면 그러한 정보를 필요로 하지 않는 회원일 수 있으므로 매우 중요하다. 이 밖에도 여러 규칙들이 추론됐으며 그중에는 우리가 미처 기대하지 못했던 규칙들도 많이 포함되어 있었다.

3.4 가시화

[그림 4]는 모형화 과정에서 얻어진 규칙(rule)들을 사용자에게 효과적으로 전달하기 위해 InferView를 이용해 가시화(Visualization)한 예이다. 그림에서 나타난 규칙은 앞서 언급한 분류조건을 만족하는 규칙들이다. InferView에서는 3차원 구를 통해 분류조건을 만족하는 데이터의 군집을 표현했으며 구를 클릭 했을때 나타난 3차원 막대그래프는 분류조건을 만족하는 모든 데이터분포를 나타낸다.



[그림 4] InferView를 이용한 가시화(Visualization) [3]

5. 결론 및 향후 방향

본 논문에서 제시한 지능형 학회 관리 시스템의 설계 및 구현 사례는 시스템을 통해서 얻은 학회 웹사이트 사용 기록, 학회 DB에서 얻은 회원의 세부 정보를 분석했다. 추출된 데이터를 기반으로 데이터마이닝 기법을 적용해 몇 가지 중요한 패턴을 발견했으며 이를 통한 개인화 방안, 즉 각 회원의 필요 정보를 제시했다. 그러나 모형화 과정에서 의사결정나무만을 사용한 점에서는 보다 정확하고 신뢰도 높은 분석 결과를 위해, 데이터마이닝의 목적이나 데이터의 특징에 맞는 기법들이 선택되어야 하므로 신경망, 클러스터링 등의 여러 기법들을 적용하는 노력이 필요하다. 본 논문에서 데이터마이닝 과정을 통해 추천 시스템 구축을 위한 관심분야 분석 작업으로 개인화 시스템 설계를 제시했다. 이를 토대로 실시간 데이터마이닝을 이용한 개인화 솔루션을 구현하기 위한 연구와 개발이 필요하다.

참고 문헌

1. Lingras. P, "Rough set clustering for Web mining", FUZZ-IEEE Proceedings of the 2002 IEEE International Conference, Vol. 2, pp. 1039 - 1044, 2002
2. 백성욱, "데이터마이닝 분야의 연구, 개발 및 활용", KOSEN(www.kosen21.org), 2001
3. Bala J.,Baik S.,Gutta S., Hadjarian A.,Mannucci M.,and Pachowicz P, "InferView:An Integrated System For Knowledge Acquisition And Visualization", proceedings of the Fedral Data Mining Symposium & Exposition 99, McLean, Virginia, 1999