서비스 연구 창간호, 2011. 9

다차원 연관 분석을 이용한 인터넷 이용자의 특징 분석

이수은* · 정용규**

목 차

요 약

데이터 마이닝은 대용량의 데이터베이스로부터 기존에 알려지지 않은, 즉 단순한 질의어로 추출할 수 없는 형태의 '유용한' 정보를 찾아내고 이를 바탕으로 데이터에 대한 통찰(insight)을 얻는 것으로 정의할 수 있다.

본 논문에서는 웹에서 발생하거나 웹 사이트에 저장한 데이터를 대상으로 유용한 패턴을 찾아내기 위하여 인터넷을 이용하는 이용자의 특징을 분석하기 위해 시도되었다. 즉 인터넷 사용자에 대한 일반적인 통계 정보 데이터에 연관성 분석을 적용하여 인터넷 사용 시간에 영향을 미치는 인터넷 이용자의 특징을 분석하였다. 실험을 통하여 데이터로부터의 연관 규칙을 추출해내었으며, 최적의 결과를 도출하기위한 데이터 전처리 및 알고리즘을 적용하여 웹 마이닝을 위한 인터넷 사용자의 특징을 분석한 결과 그 유용성을 확인할 수 있었다.

표제어: 연관 규칙, 데이터마이닝, 웹 마이닝

접수일(2011년 8월 30일), 수정일(2011년 9월 5일), 게재확정일(2011년 9월 20일)

* 을지대학교 의료산업학부 의료전산학전공, sueun1874@naver.com
** 을지대학교 의료IT마케팅학과 교수, ygjung@eulji.ac.kr

1. 서론

비 트랜잭션 데이터들을 대상으로 신뢰도가 높은 연관 규칙들을 도출하기 위해서는 유사한 연관 규칙을 보이는 데이터들에 동일한 항목을 부여하는 것이 중요하다. 각 데이터들이 부여받는 항목은 각 속성의 구간화 단계에서 결정되므로, 결과로 도출된 연관 규칙들의 신뢰도는 각 속성의 구간 설정 방법에 절대적인 영향을 받는다고 할 수 있다. 따라서 속성의 구간 설정 방법에 대한 연구는 활발하게 논의되어 온 주제이다[3-6].

데이터 마이닝 기법이란, 대량의 데이터로부터 이들 내에 존재하는 관계, 패턴, 규칙 등을 탐색하고 모형화함으로써 유용하고 새롭고 의미 있는 정보들을 추출하는 일련의 과정을 말한다. 최근에 자료의 효율적 저장을 위한 기술(데이터베이스, 통신 등)의 발달에 의한 데이터양의 급속한 팽창과 컴퓨터의 성능의 향상과 더불어 거대한 데이터의 실시간 분석이 가능해지고 거대한 데이터로부터 분석을 통한 새로운 지식의 발견이 가능해 짐에 따라 21세기 지식 정보화 사회에서는 새로운 지식의 습득이 경쟁력의 원천이 되고 있다. 데이터를 분석하여 얻어진 정보는 새로운 마케팅 전략 구축을 위한 기반이 되고, 개인의 신용평가, 통계적 품질관리(Statistical Process Control), 유통이나 의료등 여러 분야에서 활용이 되고 있다[7, 1]. 일반적으로 데이터 마이닝을 위해서는 다섯 단계의 반복된 절차들(Data Selection, Cleansing, Transformation, Mining, Interpretation)[2]이 필요로 된다. 이 중에서 Data Selection, Cleansing의 과정은 데이터를 분류하기 위해서 수행되어 진다.

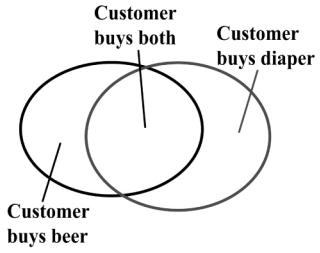
본 논문에서는 웹에서 발생하거나 웹 사이트에 저장한 데이터를 대상으로 유용한 패턴을 찾아내기 위하여 인터넷을 이용하는 이용자의 특정을 분석하고자 한다. 인터넷 사용자에 대한 일반적인 통계 정보 데이터를 가지고 연관성 분석을 통하여 인터넷 이용자의 특징을 분석하고 한다.

2. 관련 연구

2.1 연관 규칙 마이닝

연관 규칙 마이닝은 대용량 데이터베이스에서 아이템이 같이 유용한 연관 패턴을 찾아내는 것으로 장바구니 분석, 교차 마케팅, 카탈로그 디자인 등에 적용된다. 연관 규칙 마이닝을 통하여 연관성 있는 것끼리 배치가 가능하게 된다.

연관 규칙 마이닝은 기본적으로 아이템과 트랜잭션으로 구성되게 되는데, 트



[그림 2-1] 트랜젝션의 밴다이어그램
[Fig. 2-1] Venn Diagram of the Transaction

랜잭션으로 구성된 데이터베이스로부터 지지도와 신뢰도를 이용한 모든 규칙들을 발견하게 된다. [그림 2-1]에서 연관성 평가 척도인 최소 지지도(Support)도와 신뢰도(Confidence)는 식 (1)과 식 (2)와 같이 구할 수 있다.

$$S = P(beer \cap diaper) \quad (1)$$

$$C = P(diaper|beer) \quad (2)$$

연관 규칙 마이닝으로 발견된 규칙들 즉, 최소 지지도와 최소 신뢰도 임계값을 만족하는 규칙들이 모두 실제 적용가능할 정도로 유용하지 않으므로 향상도(Lift value)를 이용한다. 항목 집합 A의 발생은 항목집합 B와 독립한다고 가정하면 식 (3)을 만족한다.

$$P(A \cap B) = P(A)P(B) \quad (3)$$

식 (3)에 의해서 A와 B의 상관관계(향상도)를 구하게 되면 식 (4)와 같다.

$$corr_{A,B} = \frac{P(A|B)}{P(A)} = \frac{P(A \cap B)}{P(A)P(B)} \quad (4)$$

향상도에 따라서 값이 1보다 작으면, A의 발생은 B의 발생에 부정적으로 상관관계에 있으며, 값이 1이면 A와 B의 발생은 서로 독립 즉, 상관관계가 없다고 한다. 그리고 값이 1보다 크면, A의 발생은 B의 발생에 긍정적으로 상관관계에 있다고 한다. 연관 규칙 마이닝은 연관 규칙에서 다루는 데이터 타입과 연관 규칙에 포함된 데이터의 차원에 따라서 <표 2-1>과 같이 분류 할 수 있다.

〈표 2-1〉 연관규칙 마이닝의 분류
〈Tab. 2-1〉 Types of Associated Mining

구 분	분류
데이터 타입에 따른 분류	불리언(boolean) 연관 규칙
	정량적(quantitative)
데이터의 차원에 따른 분류	일차원 연관 규칙
	다차원 연관 규칙

연관 규칙 마이닝의 후처리 단계에서 연관 규칙들을 병합함으로써 규칙이 포함하는 속성들을 구간화하는 방법을 제안하였다. 여기서 연관 규칙의 우변(RHS)은 단일 속성으로 표현되고, 좌변(LHS)은 n 개의 속성들로 표현되는 형태 즉, A1 ^ A2 ^ … ^ An ⇒ B와 같은 규칙들을 도출하는 방법을 제안하였다. 이러한 기법의 기본적인 아이디어는 다음과 같다. 우선, 좌변에 포함되는 n개의 속성들을 각각 일정한 크기를 갖는 작은 구간들로 분할한다. 이러한 방법으로 좌변에 포함되는 속성들을 이용하여 n차원의 격자를 생성한다. 생성된 격자에서 좌변과 우변의 조건을 동시에 만족하는 칸들을 표시한다. 이렇게 표시된 칸들은 각각 하나의 작은 규칙을 의미하게 된다. 이후, 표시된 칸들 중에서 인접 하고 있는 칸들을 병합하여 하나의 규칙으로 만듦으로써 도출된 규칙의 수를 줄인다.

연관 규칙에서 우변은 분석가가 관찰을 통하여 이미 알고 있는 현상이고, 좌변은 분석가가 알고자하는 우변 의 원인을

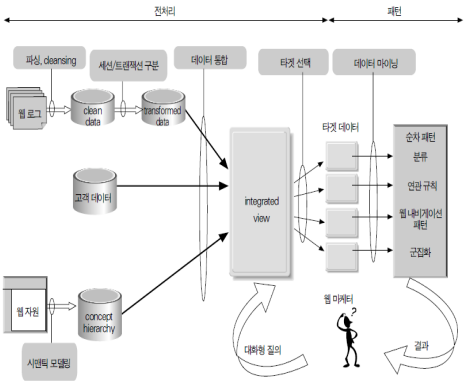
의미한다. 이런 기법은 우변이 단일 속성으로 이루어져 있어야만 한다는 제약사항을 갖고 있다. 그러나 실제 세계에서 관찰되는 현상들 중에서 단일 속성으로 표현할 수 있는 경우는 매우 드물다. 우변은 단일속성 보다는 다수의 속성으로 표현될 때에 현상을 보다 정확하게 반영할 수 있다.

2.2 웹기반 마이닝시스템

웹 마이닝은 웹에서 발생하는 모든 데이터를 분석 대상으로 삼는다. 이러한 데이터로는 서버 접속 로그 데이터(server access log data), 사용자 등록 정보(user registration data 또는 profile), 사용자 세션(session), 또는 트랜잭션(transaction), ERP 데이터(enterprise resource planning data)가 있다. 웹 마이닝은 데이터마이닝의 한 분야이기도 하지만, 기존의 데이터 마이닝 알고리즘, 웹 데이터의 전처리(preprocessing)를 위한 데이터 웨어하우징 기술, 그 외에 웹 환경 관련 기술이 연관된 데이터 마이닝을 포함하는 개념으로도 이해할 수 있다.

웹 마이닝은 대상이 되는 웹 데이터에 따라 웹 구조 마이닝(web structure mining), 웹 내용 마이닝(web content mining), 웹 사용 마이닝(web usage mining)으로 나눌 수 있다.

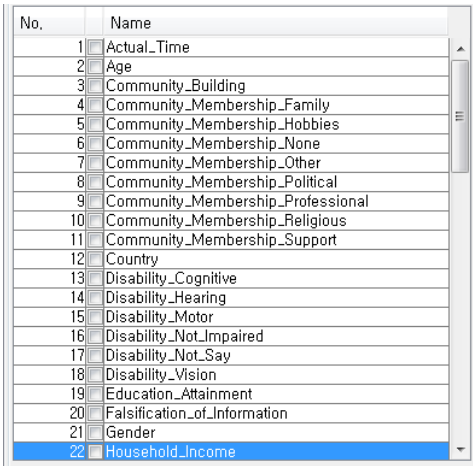
또한 서비스 관련 연구를 체계적인 이론에 기반하여 접근하기 위한 방법적인 프레임워크에서 구분하여 보면, 서비스를 대상으로 한 주체들의 실험을 통해서 사전 오류를 검출해보고 실험적으로 접근하는 측면, 이에 대한 최적의 합리성 및 타당성을 갖춘 대안을 분석하기 위한 분석적 측면, 그리고 이를 모형화하고 많은 요소들을 대입하여 공학적인 측면과, 관련 활동과 인터페이스까지를 이론적인 측면까지를 도입하는 등 체계적 및 학술적 연구 활동이 필요함을 보여주고 있다.



[그림 2-2] 웹 기반 마이닝 시스템
[Fig. 2-2] Web Based Mining System

3. 실험

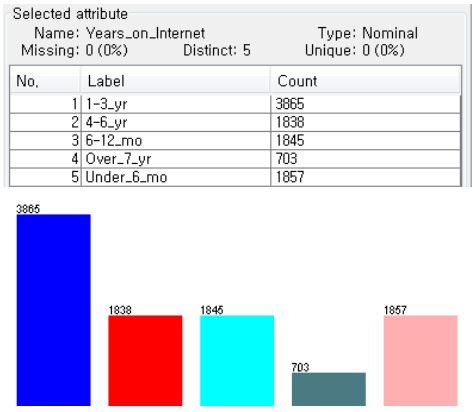
인터넷 이용자의 특징 분석을 위한 데이터는 UCI repository에서 제공한 Internet Usage dataset을 이용하였다. 이 Internet Usage dataset은 1997년 인터넷 사용자에 대한 일반적인 통계 정보로써 72개의 속성을 가지며 10,801개의 데이터를 포함하고 있다. [그림 3-1]은 속성의 일부를 보여주는데 사용자의 나이, 사는 곳, 성별, 가계수입, 교육수준, 정보위조 횟수. 결혼유무, 사용 언어, 인터넷에서 직면하는 가장 중요한 이슈, 전공분야, 전공 지역, 인터넷 사용시간 등 총 72개의 속성이 있다.



No.	Name
1	Actual_Time
2	Age
3	Community_Building
4	Community_Membership_Family
5	Community_Membership_Hobbies
6	Community_Membership_None
7	Community_Membership_Other
8	Community_Membership_Political
9	Community_Membership_Professional
10	Community_Membership_Religious
11	Community_Membership_Support
12	Country
13	Disability_Cognitive
14	Disability_Hearing
15	Disability_Motor
16	Disability_Not_Impaired
17	Disability_Not_Say
18	Disability_Vision
19	Education_Attainment
20	Falsification_of_Information
21	Gender
22	Household_Income

[그림 3-1] Internet Usage 데이터 속성

[Fig 3-1] Attributes of Internet Usage Dataset



[그림 3-2] Year__on__Internet의 클래스 분포

[Fig. 3-2] Distribution of Year__on__ Internet Classes

데이터는 애뮬레이터를 CfsSubsetEval를 이용하였고, 탐색방법은 BestFirst로 한 Attribute Selection을 통해 72개의 속성을 41개로 전 처리하였다. 전 처리한 후 Apriori Associations Rule을 적용하여 연관규칙을 발견했다. [그림 3-2]는 데이터의 일부 속성의 클래스 분포를 나타낸다.

4. 실험결과 및 고찰

실험은 데이터 마이닝 문제들을 해결하기위한 기계학습 알고리즘들을 구현하는 WEKA로 하였다. Internet Usage dataset에 연관성 규칙을 적용한 결과는 <표 4-1>과 같다. <표 4-1>에는 신뢰도가 0.98인 최적의 규칙들을 나타낸다.

〈표 4-1〉 최적의 규칙들

〈Tab. 4-1〉 Optimized Rules

1. Not_Purchasing_Cant_find = 0 9290
   ==> Disability_Hearing = 0 9136
2. Not_Purchasing_Bad_experience = 0
   Not_Purchasing_Company_policy = 0 9590
   ==> Disability_Hearing = 0 9431
3. Not_Purchasing_Bad_experience = 0 9914
   ==> Disability_Hearing = 0 9747
4. Not_Purchasing_Company_policy = 0 9771
   ==> Disability_Hearing = 0 9606
5. Disability_Hearin g = 0
   Not_Purchasing_Too_complicated = 0 9385
   ==> Not_Purchasing_Bad_experience = 0 9226
6. Not_Purchasing_Company_policy = 0
   Not_Purchasing_Too_complicated = 0 9255
   ==> Disability_Hearing = 0 9098
7. Community_Membership_Religious = 0 9408
   ==> Disability_Hearing = 0 9248
8. Not_Purchasing_Bad_experienc e= 0
   Not_Purchasing_Too_complicated = 0 9386
   ==> Disability_Hearing = 0 9226
9. Not_Purchasing_Company_policy = 0
   Not_Purchasing_Too_complicated = 0 9255
   ==> Not_Purchasing_Bad_experience = 0 9097
10. Not_Purchasing_Too_complicated = 0 9552
    ==> Not_Purchasing_Bad_experience = 0 9386

이를 통하여 Attribute를 41개로 Reduction하였고 Who 속성을 제외한 40개의 속성에서 39개의 이용자 정보를 기반으로 하여 Years on Internet 속성에 영향을 주는 주요 특징을 분석하였다.

5. 결론

웹에서 발생하거나 웹 사이트에 저장한 데이터를 대상으로 유용한 패턴을 찾아내기 위하여 인터넷 사용자에 대한 일반적인 통계 정보를 통하여 인터넷 이용시간에 영향을 미치는 인터넷 사용자의 특징을 분석하고자 했다.

UCI repository에서 제공한 Internet Usage dataset을 이용하여 실험을 진행하였는데 72개의 속성을 가지고 만개 이상의 데이터를 포함하여 데이터에 대한 파악이 힘들었다. 그래서 Attribute Selection방법을 통하여 속성을 축소하여 실험했다. 여러 속성에서 Years on Internet 속성에 영향을 주는 주요 특징을 연관규칙을 통하여 분석하고자 했다. 실험 결과는 데이터로부터의 단순한 연관 규칙만을 추출해 내었으며, 추후에 최적의 결과를 내기위한 데이터 전처리 및 알고리즘을 적용하여 웹 마이닝을 위한 인터넷 사용자의 특징 분석을 하고자 한다.

참고문헌

[1] Agrawal, R., T. Imielinski, and A. Swami(1988), “Database Mining: A Performance Perspective”, IEEE Transactions on Knowledge and Data Engineering, 5(6), 914-925, Dec, 1993.

[2] Leem, Y. M. and K. J. Rogers(2001), “Modeling for Pattern Recognition of Information Flow in Manufacturing Systems”, Proceeding of EDA 2001 Conference, 511-516.

[3] Miller, R. J. and Y. Yang(1007), “Association Rules Over Interval Data”, In Proc. ACM Int'l. Conf. on Management of Data, ACM SIGMOD, 452-461.

[4] Mining Association Rules between Sets of Items in Large Databases (1998), Proceedings of the ACM SIGMOD Conference Washington DC, USA, May, 1993.

[5] Povinelli, R.(2001), Identifying Temporal Patterns for Characterization and Prediction of Financial Time Series Events, Springer Berlin.

[6] Zhang, T., R. Ramakrishnan, and M. Livny(1996), “BIRCH : An Efficient Data Clustering Method for Very Large Databases”, In Proc. ACM Int'l. Conf. on Management of Data, ACM SIGMOD, 103-114.

[7] 곽영훈 외(2001), Missing Values in Classification Trees.

[8] 김영문, 곽준구(2004), 효율적인 데이터

마이닝을 위한 데이터 범주화에 관한 방법론, 한국산업경영시스템학회 2004 춘계학술대회논문집.

[9] 김태수, 박지영(2004), 다차원 시소러스 구축에 관한 실험적 연구, 지식처리연구, 5(1/2).

[10] 송유진, 강장묵(2011), “클라우드 컴퓨팅 환경에서의 데이터의 안전한 저장을 위한 관리방법”, 한국인터넷방송학술대회.

[11] 표창균, 한경석(2011), “Web2.0 기술을 적용한 대용량자료 실시간 활용 모델 연구: 국방 사례를 중심으로”, 한국IT서비스학회, 576-581.

Analysis of Internet User Features using Multi-dimensional Association Analysis

Su-Eun Lee* · Yong-Gyu Jung**

ABSTRACT

Data mining that can not be extracted with a simple query in the form of "useful" means to find information in large databases from the existing and unknown knowledge. It is based on this insight about the data can be defined as a gain. In this paper, we use the Internet to find useful patterns on the Web or saved data to the target Web site, which is to analyze the characteristics of users. A general statistical information on Internet users to the data by applying a relevance analysis, Internet use affect the amount of time to analyze the characteristics of Internet users. Only through experiments extracting data from the association rules, producing optimal results apply for the data pre-processing and algorithm for mining the Web to Internet users' characteristics were analyzed.

Key word: Association Rules, Data Mining, Web Mining

* Eulji University Department of Medical Industry, Medical Computer Science, sueun1874@naver.com
** Eulji University IT Department of Medical Marketing, Professor, ygjung@eulji.ac.kr

저 자 소 개

이 수 은 (Su-Eun Lee)

을지대학교 의료산업학부 의료전산학전공에 재학하고 있으며, 비즈니스인텔리젼스 및 집단지성을 적용한 다양한 분야의 적용을 위한 프로그램을 연구 중이다. 또한 의료데이터를 데이터마이닝 기법을 이용하여 다양하게 알고리즘을 적용하는 실험을 진행 중이다.

정 용 규 (Yong-Gyu Jung)

서울대학교, 연세대학교, 경기대학교에서 각각 학사, 석사, 박사학위를 취득하였고, 현재 을지대학교 의료IT마케팅학과 교수로 재직 중이다. ISO/TC154, UN/Cefact의 한국대표위원으로 활동하고 있으며, 의료정보, 전자무역, 해상물류, 금융전산에 Semantic Web, Process Modelling, ebXML 등의 표준기술의 적용에 관심이 많다.