웹 로그와 구매 DB를 이용한 개인화 시스템에 관한 연구

김 영 태 , 이 성 주
조선대학교 전자계산학과
kyttie@hanmail.net

A Study on Personalization System Using Web Log and Purchasing Database

Abstract

In this paper, a methodolgy for customizing web pages for indivisual users is suggested. It shows an efficient way to personalize web pages by predicting one's site access pattern. In addition, the prediction can reflect one's tendency after actual purchase. By using the APRIORI algorithm, one of the association rule search methods, the associativity among the purchase items can be inferred. This inferrence is based on the log data in a web server and database about purchase. Finally, a web page which contains the relationship, relative links on other web pages, and inferred items can be generated after this process.

Keyword : personalization, association rule, Apriori algorithm

1. 서 론

이제 인터넷은 완전히 대중화되었으며 또한 전자상거래 및 인터넷 쇼핑몰도 대중화되어 가고 있는 추세이다. 그러나 이러한 전자상거래에서 고객과의 효과적인 커뮤니케이션의 실현이 이루어지지 않는다면 결코 고객들에게 좋은 호응을 얻지 못할 것 이다. 그러므로 사용자의 선호도나 관심, 구매경험과 같은 정보를 이용하여 사용자에게 알맞은 정보를 제공한다면 고객들에게 좋은 호응을 얻을 수 있을 것이다. 따라서 고객의 정보와 행동 패턴을 분석하고 이를 이용하여 고객의 다음 행동이나 구입할 물품을 예측해 추천하는 등의 고객별로 각자의 구미에 맞는 정보를 제공하는 개인화(Personalization)는 매우 중요하다. 개인화를 통해 쇼핑몰 운영자는 사용자의 지속적인 이용이나 구매를 얻어낼 수 있으며 사용자 또한 자신에게 가장 알맞은 정보를 편리한 방법으로 얻을 수 있게 되는 것이다.

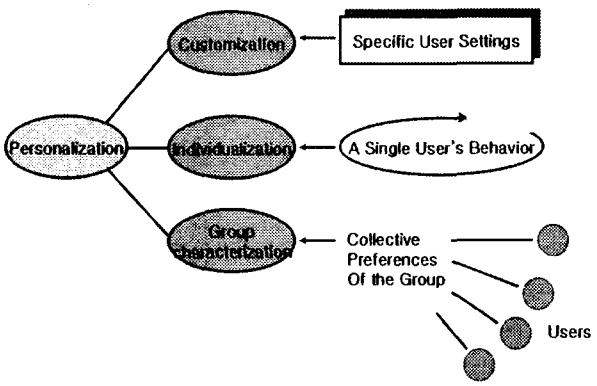
웹사이트에서 사용자들의 접근 로그 데이터는 고객의 접근 패턴을 분석하는데 중요한 자료를 제공한다. 또한 고객들이 웹사이트에서 실제 구매했던 물품들 또한 고객의 향후 구입할 수 있는 물품을 예측하는데 중요한 정보가 된다. 따라서 본 논문에서는 이러한 개인화된 컨텐츠를 제공하기 위하여 웹 서버의 로그 데이터들과 데이터베이스의 구매정보를 이용하여 데이터마이닝의 기법중의 하나인 연관규칙을 사용하는 방법을 제시하였다. 이러한 두 가지의 정보를 이용하여 예측함으로써 보다 신뢰성 있는 웹 페이지의 개인화를 제공할 수 있다.

2. 관련연구

2.1 개인화(Personalization)

개인화는 초기의 웹 페이지의 내용과 화면구성을 웹 이용자에게 맞도록 제작하는 것을 말한다. 사용자가 자신의 선호, 관심, 구매경험등과 같은 정보를 웹사이트에

제공하고 웹사이트에서는 그 자료를 바탕으로 사용자에게 알맞은 정보를 제공함으로써 운영자는 사용자에 대한 정보를 얻고 사용자는 지속적인 이용이나 구매를 하게 되고 자신에게 필요한 정보를 편리하게 얻을 수 있게 되는 것이다. 현재 기술관점에서 웹사이트를 보았을 때, 정적인 방식의 일방적인 사이트(HTML)에서 출발하여 DB 와 연동되는 동적 사이트(CGI, ASP, PHP, JSP) 그리고 고객의 프로파일 정보와 분석, 학습엔진 등이 완전하게 통합하여 개별 고객에게 적용되는 personalization 사이트의 구현이 가능하게 발전하였다[9].



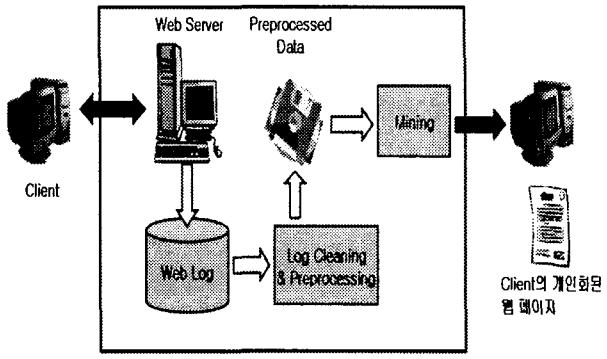
[그림 1] 개인화의 개념

개인화의 개념은 여러 가지가 혼동되어 사용되고 있어 개념상의 혼동을 주고 있는데, Rubricsoft 사의 정의에 따르면, 개인화라는 것은 고객화(Customization)와 개별화(Individualization), 그룹특성화(Group Character-ization)의 3가지 형태를 포함하는 광의의 개념이다.

본 논문은 이 개념 중 개별화(individualization)에 초점을 둔 논문이다. 고객에게 특별한 컨텐츠를 제공하기 위해 이용 형태를 패턴화하고, 설정하는 것을 말한다. 여기서는 사용자의 이용 형태를 패턴화하고자 웹 로그와 사용자의 구매 DB 를 이용하였다.

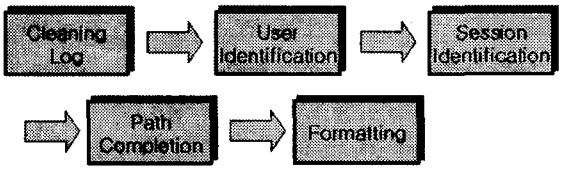
2.2 웹 로그 마이닝(Weblog Mining)

일반적으로 웹 마이닝은 웹 페이지의 내용을 마이닝하는 Web content mining 과 웹의 이용경로를 마이닝하는 Web usage mining 으로 나눌 수 있다. 그런데 기업의 입장에서 볼 때에는 web usage mining 이 전자상거래 등에서 고객의 행동패턴분석 같은 정보를 제공해 줄 수 있으므로 주로 연구되어 오고 있다[4]. Web usage mining 에서 일반적으로 많이 사용하는 방법이 웹 서버의 로그를 이용하는 것이다. 웹 서버에 접근한 사용자의 로그 파일을 분석함으로써 단순한 웹사이트를 방문한 사용자의 수를 아는 거 이상으로 기간별분석, 사용자분석, 페이지분석 등 다양한 분석을 할 수 있다. 일반적인 웹 로그 마이닝의 방법은 아래의 그림과 같이 생각할 수 있다[4].



[그림 2] 웹 로그 마이닝의 순서

먼저 로그데이터의 수집이 이루어지면 그 데이터를 전처리 하는 과정을 필요로 한다. 이 과정은 로그 데이터를 분석 가능한 데이터로 변환하는 작업이라 할 수 있다. 로그에 저장되는 데이터는 원시적인 형태의 데이터이기 때문에 분석에 적합한 형태로 데이터를 변환하고 정제하는 과정은 필수적이라 할 수 있다.



[그림 3] 로그 데이터의 전처리 과정

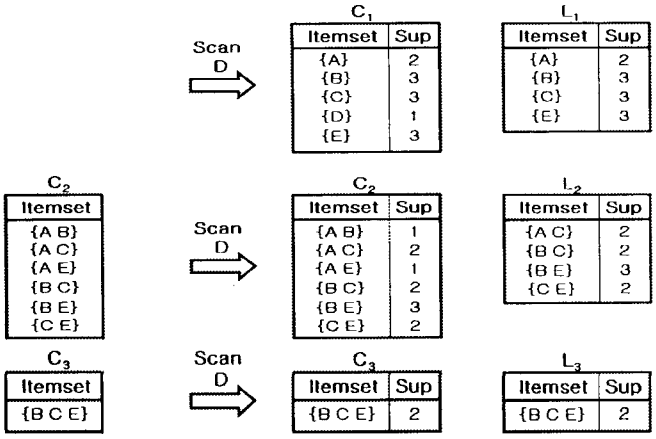
로그에서 분석에 필요 없는 부분들을 정제하는 Clean-ing Log, 개별 사용자 패턴에 관한 정보를 필요로 하기 때문에 반드시 필요한 User Identification, 사용자의 시간 초과 유무를 점검하여 정제하는 과정으로 Session Identification, 로그에 기록되지 않은 궤적을 연결하는 Path Completion, 데이터 마이닝 분석에 필요한 데이터 포맷으로 전환하는 Formatting 과정이 있다.

2.3 Apriori 알고리즘

Apriori 알고리즘에서는 각 패스에서 빈발 항목집합의 후보 집합을 구성하고 난 후에 각 후보 항목집합의 발생 빈도수를 계산하고, 사용자가 정의한 최소 지지도를 기초로 하여 빈발 항목집합을 결정한다.

TID	Items
100	A C D
200	B C E
300	A B C E
400	B E

트랜잭션 데이터베이스



[그림 4] 후보 항목집합의 생성과 빈발 항목집합

[그림 4]는 빈발 항목집합을 찾는 과정을 설명한다. 처음에는 각 항목의 발생 빈도수를 세기 위해 단순히 모든 트랜잭션들을 스캔하여 읽어 후보 1-항목집합들의 집합 C1 을 얻는다. 최소 지지도를 2 라고 가정하면(s_{min}=2) 빈발 1-항목집합들의 집합 L_1 이 결정될 수 있다. 다음으로 후보 항목집합 C_2 를 생성하기 위해 접합연산자를 이용해 L_1*L_2 를 구한다. C_2 는 |L1|/2 개의 2-항목집합들로 이루어진다. 여기서 다시 데이터베이스를 스캔하여 후보 항목집합 C_2 에서 s_{min}>= 2 를 만족하는 빈발 2-항목집합 L_2 를 구한다. 이러한 과정을 품목 수를 늘려가면서 최소지지도를 만족하는 빈발품목 집합이 존재하지 않을 때까지 계속 진행한다.

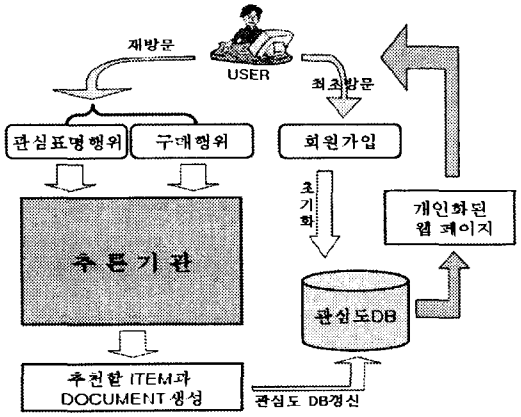
3. 개인화 시스템

3.1 시스템의 개요

본 연구에서는 개인화의 방법으로 기본적으로 사용자의 구매데이터와 로그데이터를 이용하고자 하였다. 따라서 고객의 신상정보와 구매정보를 바탕으로 추론이 이루어지는 부분과 로그데이터를 통해 고객의 웹 문서의 운행경로를 파악해 추론에 이용하는 두 가지 부분으로 이루어져 있다. 추론에는 Apriori 알고리즘이 사용되는데 구매데이터를 이용해 추론하는 부분에서 Apriori 알고리즘은 데이터베이스 검색 대상이 품목(item set)들인 반면 로그데이터를 이용한 추론에서 Apriori 알고리즘은 데이터베이스 검색 대상이 문서집합(document set)이라는 차이점이 있다.

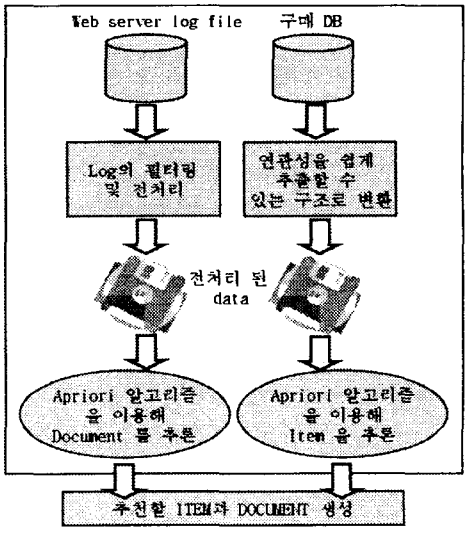
3.2 세부적 내용

기본내용에서 설명한 것을 좀더 세분하여 나타내면 [그림 5]와 같이 나타낼 수 있다.



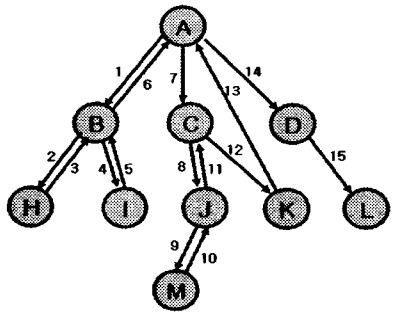
[그림 5] 시스템 순서도

사용자가 처음 해당사이트를 방문해서 회원가입을 하게 되면 고객의 정보를 저장하는 DB 와 함께 고객의 관심도에 따라 추천할 ITEM 들과 웹 페이지들을 저장할 DB 를 초기화한다. 그리고 사용자가 웹사이트에서 여러 가지 제품에 대한 관심표명행위나 구매행위를 하게 되는데 이러한 것들은 서버의 웹 로그와 구매데이터베이스에 기록이 남게 된다. 이것을 이용하여 추론기관에서 다음에 사용자가 로그인을 했을 때 추천할 ITEM 과 웹 페이지를 추론하게 되고 추론된 후에 관심도 DB 에 갱신되게 된다. 이것을 이용하여 사용자가 다시 재방문을 했을 때 이 정보를 바탕으로 사용자에게 개인화된 웹 페이지를 제공하게 된다. 사용자가 방문할 때 마다 웹 로그와 데이터베이스의 기록이 남게 될 것이므로 관심도 DB 도 계속 갱신되어 보다 신뢰도 높은 정보를 제공할 수 있게 될 것이다. [그림 5]에서 추론기관을 구체적으로 나타낸 것이 [그림 6]이다.



[그림 6] 추론기관

정보가 축적되지 않은 새로운 사용자들이라면 회원가입시의 프로파일을 이용하거나 사례기반추론을 이용해 비슷한 경력이나 관심을 가진 고객이 어떤 제품을 구매했는지를 추론해 추천하는 방법들이 있겠으나 여기서는 고려하지 않고 기본적 페이지만을 제시하게 한다. Apriori 알고리즘을 적용해 추론하기에 앞서 로그파일이나 구매정보의 DB를 전처리 하는 과정이 필요한데 로그파일을 전처리 하는 과정에서는 일반적인 전처리 과정을 수행하는 동시에 사용자 별로 운행경로 P 를 알아내고 MFR(Maximal Forward Reference)를 알아내는 과정을 수행한다. 예를 들어 그림[7]에서 웹 문서의 이동한 경로는 P = {A, B, H, B, I, B, A, C, J, M, J, C, K, A, D, I}가 된다. 후진한 경우를 제외하고 전진한 경우에 거쳐간 웹 문서(MFR)은 이 그림을 예로 들어 MFR = { {ABH}, {ABI}, {ACJM}, {ACK}, {ADL} }이 된다.



[그림 7]

이 과정은 DocumentSet 을 이용해 Apriori 알고리즘을 적용하기 위해 필수적으로 선행되어야 한다. 구매정보의 DB 의 전처리는 필요 없는 데이터의 필터링, 마이닝을 적용하기 위한 포맷팅 등 일반적인 과정을 적용한다. 이러한 과정이 끝나면 각각의 추론기관을 통해 추론과정을 수행한다. 이 과정을 통해 각각 웹 페이지들과 물품들이 추론기관을 통해 추천되는데 이러한 정보들은 관심도 DB 에 갱신된다. 그리고 사용자가 로그인 했을 때 이 관심도 DB 를 바탕으로 사용자의 초기 웹 페이지에 추천된 웹 문서를 링크시키고 또한 추천된 물품들을 소개한다.

4. 결 론 및 향후연구

본 논문은 개인의 특성에 맞게 웹사이트를 구축하기 위한 개인화의 한가지 방법을 제시하였고 Apriori 알고리즘을 사용해서 웹 로그와 사용자의 구매정보를 동시에 이용함으로써 개인화에 대한 신뢰성을 높이고자 하였다. 사용자가 로그정보와 이전의 구매정보를 바탕으로 웹 문서들과 아이템들을 동시에 추천받음으로써 보다 만족도가 높은 개인화를 이룰 수 있을 것이다. 그러나 여기서는 정보가 충분하지 못한 신규가입자의 문제에 대해서는 생각하지 않았다. 앞으로 신규가입자의 경우 사례기반추론을 이용해 추천하는 방법이나 가입 양식에서 개인 취향을 고려한 정보를 추출하여 본 시스템에 적용하는 연구가 필요하다. 또한 다른 쇼핑몰과의 연계를 통한 정보의 공유로 신뢰도 높은 개인화를 이루는 문제도 연구가치가 있다.

5. 참고문헌

[1]정원석, 이극 " 지능형 에이전트를 이용한 개인화 된 웹 정보 서비스" 한남대학교 컴퓨터공학과

[2]박종수, 유원경 " 연관 규칙 탐사와 그 응용" 성신여자대학교 전산학과

[3]고경자, 진훈, 김인철 " 웹 로그 마이닝에 기초한 적응형 웹사이트에 관한 연구" 경기대학교 전자계산학과

[4]이경우, 최덕원" 전자상거래에서 상품 추천을 위한 웹 개인화 방안에 관한 연구" 성균관대학교 시스템경영공학부

[5]김석기, 안정용, 한경수" 웹 로그 데이터 분석 방법에 관한 연구"

[6]Harris Kravatz " Design Web Personalization Features"

[7]R.Agrawal and R.Srikant, Fast Algorithms for Mining Association Rules, In Proc. Of the 20th VLDB Conference, pp, 487-499

[8]C-H.Lee, Y-H.Kim, P-K.Rhee " Web personalization expert with combining collaborative filtering and association rule mining technique"

[9]김기수 " 인터넷 맞춤시대를 여는 퍼스널라이제이션" 월간 마이크로소프트 2001.1 월호

[10]넷스루 기술연구소 데이터마이닝팀 " 웹마이닝" 월간 마이크로소프트 2001.5 월호