개인화에서 사용자 프로파일 구축을 위한 효과적인 규칙확인 방법

손준원*, 배기성**, 석민수*
*성균관대학교 정보통신공학부
**성균관대학교 정보통신공학부
e-mail : logos@ece.skku.ac.kr

Efficient Rule Validation Methods for User Profiling in Personalization

Jun-Won Sohn*, Kee-Sung Bae**, Min-Su Suk*
*School. of Information and Communication Engineering, Sung-Kyun-Kwan University
** School. of Information and Communication Engineering, Sung-Kyun-Kwan University

요 약

추천 시스템에서부터 1:1 마케팅에 이르는 전자 상거래의 다양한 응용 영역에서, 개별 사용자로부터 개인화된 사용자 프로파일을 구축하는 것은 매우 중요하다. 이러한 프로파일들은 사용자들의 구매 행위와 같은 개인별 행동들을 설명해주며, 특히 다양한 데이터 마이닝(Data Mining) 기술들을 이용해 사용자의 거래 기록으로부터 학습된 규칙들을 발견해낼 수 있다. 발견된 규칙들 중에는 거짓이거나 연관 없거나 또는 하찮은 것들도 존재하기 때문에, 가장 중요한 문제 가운데 하나는 발견된 규칙들을 처리후-분석을 어떻게 수행하느냐이다. 예를 들어, 발견된 규칙을 사용자 프로파일에 적합한 것인지를 확인할 때 좋은 규칙과 나쁜 규칙을 어떻게 판명하는가 하는 문제이다. 이 논문에서는 규칙을 확인하는 과정에서 객관적 척도를 이용하는 방법을 제안하였다.

1. 서론

개인화(Personalization)는 개인의 선호도와 행동에 관한 지식을 기초로 개인에게 맞춘 콘텐트와 서비스를 제공하는 능력이기 때문에 비즈니스, 특히 1:1 마케팅과 같은 마케팅 분야에 있어서 매우 중요한 도구로 인식되어 왔다. 앞서 언급했듯이 인터넷 기반의 e-비즈니스, eCRM 환경에서 고객 관리는 핵심을 차지하며, 고객 관리의 중요성이 커짐에 따라 개인화에 대한 기술적 관심도와 발전도 동반 상승하고 있다. 개인화의 응용은 개인화된 웹 콘텐트 표현으로부터 책, CD, 주식 구입 추천에 이르고 있다.

개인화의 많은 이슈들 가운데 중요하게 다루어지는 것들은 크게 아래와 같이 분류될 수 있다. 첫째는 고객들이 누구이며, 그들은 어떻게 행동하며, 다른 고객들과 어떻게 유사성을 보이는지에 관한 포괄적인 지식들을 기반으로 어떻게 개인적인 추천을 제공할 것인가 이다. 다양한 추천 시스템들이 이 문제에 부딪혀 있다. 문제를 해결하기 위해 협업필터링(collaborative-filtering)이나 콘텐트 기반 접근을 사용하고 있으며, 어떤 시스템은 두 가지 방법을 합쳐서 이를 해결하기도 한다.

둘째로 고객들의 거래 기록들로부터 학습한 정보를 이용하여 정확하고도 광범위한 개인 프로파일들을 구현하는 것이다. 프로파일이 담고 있는 정보는 사실 데이터와 규칙 데이터로 이루어져있다. 사실 데이터는 인적 정보와 같이 고객에 관한 데이터들을 의미하며, 규칙 데이터는 'A 는 빵을 사면 우유를 산다'와 같이 고객의 행동 양식을 의미하는 데이터라 할 수 있다. 데이터 마이닝(Data mining) 기법을 이용하여 데이터들로부터 행동 규칙을 이끌어 낸다.

따라서, 개인화의 핵심 이슈인 개인 사용자의 거래 기록에서부터 개인화된 프로파일(Profile)을 작성하는 것은 매우 중요하다. 또한 사실 데이터와 규칙 데이터들로 구성된 고객에 관한 복잡하며 대용량의 데이터들로부터 지식을 추출하는 데이터 마이닝 기법은 사용자 프로파일링(user profiling)에 있어 최적의 신뢰성을 제공한다.

이와 같이, 데이터 마이닝 기법을 이용해 발견된 규칙들은 사용자의 프로파일에 저장되기 전에 규칙들의 신뢰성을 판단해야만 한다. 대용량의 데이터 베이스로부터 발견된 수많은 규칙들 중에는 위조이거나 연관성이 떨어지는 것들도 많이 존재하기 때문이다. 프로파일에 저장할 수 있는 "좋은" 규칙의 판단 기준은 통계적으로 확인할 수 있고 해당 응용 분야의 전문가가 받아들일 수 있어야 하는 것이다. 본 논문에서는 먼저 개인화의 의미와 구성 기술들, 효과적인 사용자 프로파일링을 위한 규칙들간의 연관성과 신뢰성을 판단하기 위해 사용되는 다양한 메트릭(metric)들을 비교 연구해 보고자 한다.

2. 개인화

2.1. 개인화의 정의

표 1 은 Oakana-dos-Reis & Zahedi(1999) 연구에서 개인화(Personalization)와 맞춤화(Customization)를 비교한 것을 편집한 것으로서, Financial DSS 를 개발하는데 있어서의 개인화의 개념을 소개하고 있는데, 이를 통해 보면, 개인화는 이용자의 특성을 대상으로 하여 이와 관련하여 다양한 통계 분석 기법 및 데이터 마이닝 기법을 통해 개별화 할 수 있는 지식과 규칙을 찾아내는 과정으로 볼 수 있다.

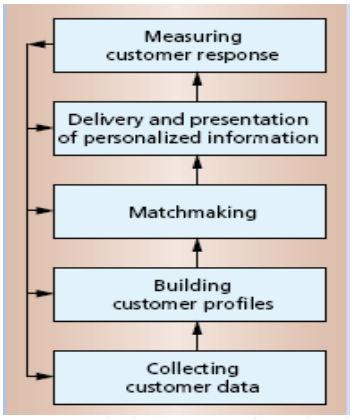
<표 1> 개인화와 맞춤화의 비교

	개 인 화	맞 춤 화
대 상	이용자 특성	이용자 경험
개인 정보 변화	느 림	상황에 따라 신속히 변화
개인 정보 수집	개인의 지식과 규칙은 퍼스널리티의 다양한 영역에서 수집된 자료를 통계적으로 분석하여 얻어짐	개인에게 직접 얻어짐
인지 여부	개인에게 정보수집 사실을 알리지 않음	특정 개인(customer)의 선호에 근거함
S/W 개발	동일한 제품의 다양한 버전 또는 다수의 S/W 제품	특정 버전의 S/W 제품
S/W 생산성	높 음	낮 음

개인화의 개념은 분야에 따라 여러 가지 의미들이 혼용되어 사용되고 있어 개념상의 혼동을 주고 있는 것이 사실이지만, 개인을 위해 사용자에게 옵션을 제공하는 것(Customization), 개별적으로 제품 및 서비스를 제공하는 것(Individualization), 비슷한 선호의 그룹의 특성에 근거하여 개별적 기능 및 서비스를 제공하는 것(Group Characterization)의 3 가지 접근 방법과 개별적 고객에게 제품을 포함하여 서비스를 개별적으로 제공하는 것 등의 개념들을 포괄한 개념으로 이해될 수 있다.

2.2. 개인화 기술의 구성

개인화 기술(Personalization Technology)은 그림 1 에서 보여지듯이 5 가지의 단계로 구성되어있는 반복적인 프로세스라고 할 수 있다.



<그림 1> 개인화 프로세스의 단계

고객 데이터 수집 단계는 데이터 수집의 단계로써 이 데이터는 인적 정보뿐만 아니라, 구매와 브라우징(browsing) 행위까지도 포함한 광범위한 것을 일컫는다. 데이터들이 수집된 후에는, 데이터 웨어하우스에 전처리(prepare), 정제(clean), 저장(store) 된다. 수집된 데이터를 기반으로 정확하고 포괄적인 고객 프로파일의 구성이 필수적이며, 이 단계는 개인화 구축에 있어서 핵심 이슈이다. 또한, 개인화 시스템은 반드시 적합한 콘텐트와 서비스를 개별적으로 고객에게 맞춰주어야 하므로, Matchmaking 단계는 개인화된 추천을 제공한다. 이러한 추천 시스템들은 콘텐트 기반이나 협업필터링(collaborative-filtering)과 같은 기술들을 사용한다. 콘텐트 기반 시스템은 과거에 고객이 선호했던 제품과 유사한 제품을 추천하게 되며, 협업필터링(collaborative-filtering) 시스템은 과거에 유사한 성향과 선호도를 보였던 다른 고객들이 제품을 추천한다. 협업필터링을 사용하는 솔루션이나 회사의 경우 조금씩 다른 특징들을 가지고 있지만, 공통적으로 사람의 입력 정보 즉, 프로파일에 의존한다. 이것의 단점은 입력 정보가 주관적이고 서술적 정보이기 때문에 부정확할 수 있다는 점과 프로파일은 정적이어서 그 정보가 수집된 후의 일정 시점이 지난 뒤에는 프로파일 정보가 과거 정보가 되어 수행 성과가 낮아진다는 점이다. 전자 상거래 회사들은 몇 가지 방법을 통해 개인화된 정보를 고객에게 전달한다. 전달 방법 분류들 중 한가지는 pull, push, passive 의 3 가지이다. Push 방법은 현재 시스템과 상호작용하고 있지 않은 고객에게 접근하는 것이다. Pull 방법은 고객의 분명한 요구가 있을 때에만 개인화 정보를 이용할 수 있도록 고객에게 공지하는 것이다. Passive 전달은 전자 상거래 응용 환경에서만 개인화 정보를 보여준다. 그림 1 에

서 보듯이, 고객 반응을 측정한 것은 다른 4 개의 단계에서 각각의 단계가 향상될 수 있도록 피드백을 제공해준다. 부분적으로, 시스템 디자이너들은 고객 반응을 보고 추가적인 데이터 수집, 더 나은 Matchmaking 알고리즘 개발, 정보 표현에서의 개선 등의 여부를 결정하게 된다. Scalability 는 매우 중요한 이슈이다. 기업들은 수백만의 고객과 수 만개의 상품을 다루어야 하며, 게다가 실시간 개인화 서비스는 효과적인 프로파일링과 Matchmaking 방법을 요구하고 있다.

3. 사용자 프로파일링

프로파일 구현 과정에는 규칙 발견과 규칙 확인의 두 단계가 있다. 규칙 발견 과정을 통해 발견된 결합 규칙, 연관 규칙, 분류 규칙 등의 다양한 형태의 개별 고객 행동을 모델화할 수 있다. 개인 고객의 행동을 설명하는 규칙을 발견하기 위해서는 Apori, CART 등 다양한 데이터 마이닝 알고리즘들을 사용할 수 있다. 데이터 마이닝 기법들은 통계적으로 수용할만함에도 불구하고 평범하거나 거짓 또는 관련 없는 많은 수의 규칙들을 산출해내기도 한다. 따라서, 발견된 규칙들의 확인(Validation) 과정은 매우 중요한 요구로 부각되고 있다. 발견된 규칙을 확인하는 방법 중 하나는 해당 분야의 전문가가 그것들을 조사하고 고객들의 실제 행동을 얼마나 잘 나타내었는지를 결정하는 것이다. 전문가는 몇 개의 규칙들은 받아들이고 나머지는 버리게 된다.

규칙 확인 과정에서 중요한 이슈는 Scalability 이다. 개인화의 응용 분야에서 많은 경우 사용자의 수는 엄청나게 크다. 전문가인 사람이 규칙들을 하나씩 확인하는 것은 불가능하다. 규칙 발견과 달리 규칙 확인은 각 사용자에게 분리하는 프로세스가 아니라 모든 사용자들을 한번에 처리하는 프로세스이다. 결론적으로 전문가는 유사하거나 심지어 확인된 규칙들도 다른 사용자들에게 확인해 본다.

4. 규칙 확인(Rule Validation)

4.1. 연관 규칙 마이닝(association rule mining)

연관 규칙 마이닝은 대규모 데이터 항목의 집합 사이에서 유용한 연관성과 상관관계를 찾는 방법이다. 연관 규칙 마이닝은 가장 최근에 개발된 기법으로서 다른 기법에 비해 기본 개념이 간단하고 결과가 이해하기 쉬우며 현장에 직접적인 적용이 가능하다는 장점이 있다.

연관 규칙 탐색은 크게 두 단계로 구성된다고 볼 수 있다. 첫 단계는 빈발 항목 집합(frequent itemset) 마이닝 과정이다. 즉 데이터베이스에서 동시에 발생하는 횟수가 기대 이상으로 많이 발견되는 항목들의 집합을 찾는 것이다. 어떤 항목집합이 포함된 트랜잭션들의 수가 데이터베이스 전체 트랜잭션의 수에서 차지하는 비율을 항목집합의 지지도(support)라고 정의하며, 사용자가 지정한 최소 지지도(minimum support)를 만족시키는 항목집합을 빈발 항목집합이라고 한다.

둘째 단계는 첫 단계에서 생성된 빈발 항목집합을 이용해 항목간의 규칙을 만들고 흥미도 척도를 적용해서 연관 규칙의 흥미도를 평가하는 것이다. 물론 '생성된 연관 규칙이 흥미로운 규칙이다'라고 말하는 것에는 궁극적으로 사용자의 주관적인 판단이 개입된다. 하지만 통계적인 혹은 논리적인 방법에 의해 제안하는 객관적 흥미도 척도(objective interestingness measure)는 사용자에게 규칙들을 전지(pruning)할 수 있는 근거를 제시해주기 때문에 유용하다고 할 수 있다. 처음 제안된 객관적 흥미도 척도는 신뢰도(confidence)이다.

4.2. 흥미도 척도(interestingness measure)

데이터 베이스에서 지식 발견은 사전에 알려져 있지 않거나 새롭고 이해할 수 있는 패턴들을 발견하는데 매우 효과적이다. 데이터 자체로서의 지식보다 데이터의 패턴을 설명해주는 지식인 것이다. 데이터의 패턴은 분류 규칙(classification rules), 연관 규칙(association rules), 군집(cluster) 등의 다양한 형태로 표현될 수 있다. 전형적으로 매우 많은 패턴이 산출되기 때문에 이들 중 소수의 패턴들만이 패턴을 분석하는 전문가들에게 흥미를 끌게 된다. 또한, 연관성과 활용도가 높은 패턴의 숫자가 늘어나면서 패턴의 수를 감소시킬 수 있는 기술이 필요하게 된다. 이러한 목표를 만족시키는 광범위하게 참조하는 기술이 흥미도 척도(interestingness measure)이다. 흥미도 척도는 연관 규칙 마이닝을 통해 발견된 규칙들의 유용성과 실용성을 평가하는데 매우 중요한 기준을 마련해 준다.

Agrawal 과 Srikant 는 항목집합 척도(itemset measure)를 사용한다. 항목집합 척도는 대용량 데이터베이스의 항목들의 집합들로부터 연관 규칙이 일어나는 빈번도를 확인하는데 쓰여왔다. 항목집합 X, Y 가 주어졌을 때, 연관규칙 X→Y 의 신뢰도(confidence)는 식(1)과 같이 정의한다.

$$\text{Confidence}(X \rightarrow Y) = \text{Supp}(XY)/\text{Supp}(X) \quad (1)$$

여기서, supp(XY)와 supp(X)는 각각 항목집합 X∪Y 와 X 의 지지도이다. Confidence(X→Y)는 X 가 포함된 트랜잭션의 비율을 나타내며 X 에 대한 Y 의 조건부 확률이다. 기준이 되는 신뢰도를 최소 신뢰도(minimum confidence)라고 부르며 X→Y 라는 규칙이 최소 신뢰도를 만족하면 강한 연관 규칙을 가진 규칙으로서 흥미로운 규칙이 된다. 그러나 신뢰도에 의해 강한 연관규칙임을 나타내어도 반드시 흥미 있는 규칙으로 볼 수는 없다.

흥미도는 소매 거래의 항목집합들 사이에서 연관성의 세기를 평가할 때 쓰인다. 지지도와 신뢰도가 연관 규칙 특성 결정에 사용된다면, 흥미도는 전항(antecedent)과 후항(consequent)의 독립성을 지시하는 식별항을 포함하게 된다. 식 (2)에서는 Gray 와 Orlowska 가 제안한 흥미도가 정의되어 있다.

$$I = \left(\left(\frac{P(X \cap Y)}{P(X) \times P(Y)} \right)^k - 1 \right) \times \left(P(X) \times P(Y) \right)^m \quad (2)$$

$P(X \cap Y)$ 는 신뢰도이고, $P(X) \times P(Y)$ 는 지지도,

$\frac{P(X \cap Y)}{P(X) \times P(Y)}$ 는 식별항이다. k 와 m 은 식별항과 지지항의 연관된 중요성을 각각 가중치하는 파라미터이다. 흥미도의 값이 높을수록 더 흥미롭다고 할 수 있다.

5. 실험 및 분석

규칙 확인을 위해서는 먼저 연관규칙 마이닝을 이용해 연관 규칙들을 찾아내는 것이다. 앞에서 언급한 바와 같이, 연관규칙 마이닝의 두 가지 단계를 이용하게된다. Apriori(Agrawal, R. *et al*., 1994)와 같은 대표적인 빈발 항목집합 마이닝 알고리즘들 중에서 하나를 사용해서 빈발 항목집합을 찾아낸다. 다음 단계에서 찾아낸 빈발 항목집합을 기초로 하여 연관규칙을 생성하고 흥미도 척도를 적용하여 평가한다. 4.2 절에서 언급한 흥미도 척도는 객관적 흥미도 척도(objective interestingness measure)에 해당한다. 이러한 객관적 흥미도 척도들은 서서히 연구들이 진행 중이므로, 본 실험에서는 Agrawal & Srikant's Itemset Measure, Gray & Orlowska's Interestingness, Liu et al. Reliable Exceptions, Zhong et al. Peculiarity 등 11 개의 객관적 척도들과 Klementtien et al. Rule Templates, Dong & Li's Interestingness 등 6 개의 주관적 척도들을 적용할 것이다. 규칙에 동일하게 객관적 척도와 주관적 척도를 적용함으로써 척도들 사이의 순위(ranking)와 전지(pruning)를 비교해 볼 수 있고, 이를 통해 척도들의 상관관계도 발견할 수 있다.

6. 결론

발견된 규칙들의 연관성 패턴에 대해서 올바른 흥미도 척도 문제에 대해 논의가 가능하다. 이들 흥미도 척도들은 연관 규칙뿐만 아니라 분류기법, feature 선택 등의 문제에서 사용되는 상관관계 평가 척도를 포함시켜 17 개 척도들의 특성을 비교할 수 있다. 대표적인 흥미도 척도인 신뢰도는 상호 배반적인 규칙을 모두 흥미로운 규칙으로 평가하는 오류를 범할 수 있다. 객관적 흥미도 척도들은 발견된 패턴들의 구조에 기반한 것이며, 주관적 척도들은 사용자의 신뢰(beliefs)나 데이터들간의 관계에 주목한 치우침들에 기반하는 것임을 알 수 있다.

참고 문헌

[1] Gediminas Adomavicius, Alexander Tuzhilin, New York University, "Using Data Mining Methods to Build Customer Profiles", Research Feature IEEE 74-82, Feb. 2001.

[2] P. Hagen, "Smart Personalization", *The Forrester Report*, Forrester Research, Cambridge, Mass., July1999.

[3] *Comm. ACM*, Special Issue on Recommender Systems,vol. 40, no. 3, 1997.

[4] P. Resnick et al., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," *Proc. 1994 Computer-Supported Cooperative Work Conf.*, ACM Press, New York, 1994, pp. 175-186.

[5] U. Shardanand and P. Maes, "Social Information Filtering:Algorithms for Automating 'Word of Mouth,'" *Proc.Conf. Human Factors in Computing Systems (CHI 95)*, ACM Press, New York, 1995, pp. 210-217.

[6] W. Hill et al., "Recommending and Evaluating Choices in a Virtual Community of Use", *Proc. Conf. HumanFactors in Computing Systems (CHI95)*, ACM Press, New York, 1995, pp. 194-201.

[7] G. Adomavicius and A. Tuzhilin., "Expert-Driven Validation of Rule-Based User Models in Personalization Applications", *J. Data Mining and Knowledge Discovery*, Jan. 2001, pp. 33-58.

[8] R. Agrawal et al., "Fast Discovery of Association Rules," Advances in Knowledge Discovery and Data Mining,AAAI Press, Menlo Park, Calif., 1996, chap. 12.

[9] G. Piatetsky-Shapiro and C.J. Matheus., "The Interestingness of Deviations", *Proc. AAAI-94 Workshop Knowledge Discovery in Databases*, AAAI Press, MenloPark, Calif., 1994, pp. 25-36.

[10] R. Agrawal et al., "Mining Association Rules between Sets of Items in Large Databases", *Proc. ACM SIGMOD Int. Conf. On Management of Data*, 1993, pp.207-216.

[11] A. Silberschatz and Alexander Tuzhilin., "What Makes Patterns Interesting in Knowledge Discovery Systems", *IEEE Transactions on Knowledge and Data Engineering*, 8(6), Feb. 2001, pp. 970-974.

[12] U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth., "From data mining to knowledge discovery: an overview. In: U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy." (Eds.) *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996, 1-34.

[13] R. Hilderman, H. Hamilton, and B. Barber., "Ranking the interestingness of summaries from data mining systems.", *Proc. of the 12th International Florida Artificial Intelligence Research Symposium (FLAIRS'99),* Orlando, FL, May 1999, pp. 100-106.

[14] R. Hilderman and H. Hamilton., "Evaluation of interestingness measures for ranking discovered knowledge.", *Proc. of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'01),* Hong Kong, April 200, pp. 247-259.

[15] P. Tan, V. Kumax, and J. Srivastava., "Selecting the right interestingness measure for association patterns.", *Proc. of the 8th Int. Conference on Knowledge Discovery and Data Mining*, 2002.