개인화를 위한 추천시스템 알고리즘에 관한 연구

강현철1) · 한상태2) · 신연주3)

요 약

개인화된 추천시스템(recommendation system)은 자동화된 정보 필터링 기술을 적용하여 고객의 취향에 맞는 아이템(상품, 기사, 컨텐츠 등)을 추천하는 시스템이다. 이러한 추천시스템에서 가장 중요한 것은 고객의 특성을 정확히 파악하여 가장 적절한 아이템을 추천해 줄 수 있는 능력이라고 할 수 있다. 본 연구에서는 추천시스템을 위해 제안된 여러 알고리즘들을 소개하고 그 특징들을 비교하였으며, 연관성규칙발견과 군집분석을 이용한 추천시스템 알고리즘을 실제 자료에 적용하여 그 결과를 살펴보았다.

주요용어 : 개인화, 추천시스템, 연관성규칙발견, 군집분석

1. 서론

전자상거래에서 개인화된 추천시스템(recommendation system)은 자동화된 정보 필터링 기술을 적용하여 고객의 취향에 맞는 아이템(상품, 기사, 컨텐츠 등)을 추천해 주는 시스템이다. 추천시스템에서 가장 중요한 것은 고객의 선호도를 정확하게 분석하고 정제하여 정확한 예측력으로 고객이 원하는 가장 적절한 아이템을 추천해 줄 수 있는 능력이다. 추천시스템에 관한 기존의 연구 방법은 크게 인구통계학적 추천기법, 내용기반(contents-based) 추천기법, 선호도에 의한 추천기법, 웹로그(web log) 분석에 의한 추천기법 등이 있으며, 최근에는 여러 가지 형태의 추천 기법을 결합한 방법도 연구되고 있다(금종경, 2001; 서지현, 2002).

그러나 기존의 추천 기법들은 다음과 같은 몇 가지 문제를 가지고 있다. 첫째, 사용자들이 신상 정보의 유출을 염려하여 인적사항을 비워두거나 부정확한 내용을 입력하는 관계로 정확한 사용자 유형 분석이 어렵다. 둘째, 영화나 게임등과 같이 개인별 선호도의 차이가 큰 컨텐츠에 대한 개인별 평가정보를 온라인상에서 수동적으로 수집하는 관계로 정확한 개인별 취향 분석이 어렵다. 셋째, 시간에 따라 변화하는 사용자의 개인별 선호도를 동적으로 반영할 수 있는 알고리즘 개발이 어렵다. 넷째, 디지털 컨텐츠의 종류가 다양해지고 대량으로 생산됨에 따라 개인별로 선호하는 컨텐츠를 사용자에게 제공하기 위한 자동화된 통합 솔루션이 부족하다(김영지 등, 2002).

본 연구에서는 개인화된 추천시스템을 위해 제안된 여러 알고리즘들을 소개하고 그 특징들을 비교하였으며, 연관성규칙발견과 군집분석을 이용한 추천시스템 알고리즘을 실제 자료에 적용하여 그 결과를 살펴보았다.

2. 개인화 및 추천시스템의 개념

2.1 개인화

1) 호서대학교 자연과학부 정보통계학전공 교수, (336-795) 충남 아산시 배방면 세출리 산 29-1

2) 호서대학교 자연과학부 정보통계학전공 교수, (336-795) 충남 아산시 배방면 세출리 산 29-1

3) 호서대학교 대학원 통계전공 석사과정, (336-795) 충남 아산시 배방면 세출리 산 29-1

개인화(personalization)란 웹사이트 방문자의 개별적인 요구에 대한 개인화된 컨텐츠, 상품, 커뮤니티를 제공하는 것을 의미한다. 이는 개인별로 맞춤서비스를 제공하기 위한 핵심적인 전략 중 하나이며, 개인화를 통해 고객들에게 차별화된 서비스를 제공함으로써 고객들과 친밀한 관계를 유지하여 충성도를 높일 수 있다. 또한 컨텐츠 사이트에서 개인화 전략은 사용자들이 컨텐츠를 검색하는 시간과 경비를 절약하고 사이트에 대한 만족도를 높여 매출을 극대화하기 위한 전략으로 사용할 수 있다.

개인화를 통해 달성하고자 하는 목표는 해당 사이트에 따라 매우 다양하며 이를 크게 세 가지 기능으로 요약할 수 있다. 첫째, 웹컨텐츠(web contents)의 개인화로 고객의 관심사에 맞춘 웹페이지를 구성한다. 둘째, 실시간 구매추천(real-time recommendation)으로 고객의 구매 패턴을 실시간 예측하여, 구매 가능성이 높은 제품을 추천한다. 셋째, 캠페인 관리(campaign management)로 고객의 선호에 따른 광고 및 이벤트를 제공한다.

2.2 추천시스템

개인화를 위한 핵심 기술 중의 하나는 고객들의 취향과 구매 이력을 분석하여 개인별로 차별화된 정보를 자동적으로 필터링하기 위한 추천시스템이다. 추천기법에 관한 기존의 연구 동향은 크게 인구통계학적 추천기법, 내용기반 추천기법, 선호도에 의한 추천기법, 웹로그 분석에 의한 추천기법 등으로 분류되어지는데, 이 절에서는 그 중 몇 가지 기법들을 간단히 소개하고자 한다.

인구통계학적 정보에 의한 추천기법은 사용자의 성별, 연령, 직업 등과 같은 인구통계학적 요소에 의해 사용자 유형별 특징을 분석하여 상품을 추천하는 방법이다. 이 기법은 전통적인 추천기법의 하나로, 단순한 형태의 정보필터링 기법을 이용하는 타겟 마케팅 전략의 하나로 널리 사용되고 있다. 특히 이 기법은 사용자의 피드백 정보가 없이도 상품에 대한 추천이 가능하여 시스템 초기 구축 단계나 처음 방문한 사용자에 대해서도 적용할 수 있다.

내용기반 추천기법은 개인의 요구나 개인으로부터 입력된 모든 정보와 상품에 포함된 텍스트 정보를 이용하여 필터링하는 방식이다. 이 기법은 사용자 프로파일을 통해 과거 구매나 추천 결과를 쉽게 반영할 수 있는 장점이 있으며 추천 속도가 빠르다.

선호도에 의한 추천기법은 상품 및 컨텐츠에 대해 고객 또는 평가자들로부터 수집된 선호도에 기초하여 추천하는 방법으로 사례기반추론(case-based reasoning) 기법, 사용자 기반 협업적 필터링(user-based collaborative filtering), 아이템 기반(item-based) 협업적 필터링 등의 기법들이 있다(김영지 등, 2002; 박지선 등, 2000).

3. 웹로그 데이터 분석을 통한 추천시스템

웹로그(web log) 분석에 의한 추천기법은 웹사이트 방문자의 기록인 웹로그에 대한 분석을 통해 얻어진 방문자의 패턴에 기초하여 추천하는 방법이다. 여기에는 주로 연관성규칙발견 또는 군집분석과 같은 데이터마이닝 기법이 사용되며, 본 연구에서는 사례를 통해 이러한 과정을 소개하고자 한다.

3.1 연관성규칙발견에 의한 추천 프로세스

연관성규칙발견(association rule discovery)은 아이템 간의 연관성을 분석하는 기법으로, 이때 연관성을 재기 위해 주고 사용되는 측도로는 지지도(support), 신뢰도(confidence), 향상도

(lift) 등이 있다. 즉, n을 전체 사용자 세션의 수, $n(A)$를 A라는 페이지를 방문한 세션의 수, $n(A, B)$를 A와 B를 모두 방문한 세션의 수, $n(A \rightarrow B)$를 A를 먼저 방문한 후에 B를 방문한 세션의 수라고 할 때, 이들은 다음과 같이 계산된다(강현철 & 정병철, 2001).

$$Support(A \rightarrow B) = n(A, B)/n$$
$$Confidence(A \rightarrow B) = n(A, B)/n(A)$$
$$Sequence(A \rightarrow B) = n(A \rightarrow B)/n(A)$$

<표 3.1> 연관성규칙발견의 결과 예

(a) 아이템의 수가 1인 연관성규칙발견의 결과

# of pages	Support	Confidence	Lift	Rule
1	56.94	100.00	1.76	A
1	55.56	100.00	1.80	B
1	54.86	100.00	1.82	C
1	54.17	100.00	1.85	D
1	54.17	100.00	1.85	E
...	...	...	...	...

A : /index_hnc B : /index_shopping C : /index_contents D : /index_new_login E : /index_adclick	F : /index_wiselognews G : /wiselog H : /index/community I : /index_recom

(b) 아이템의 수가 2인 연관성규칙발견의 결과

# of pages	Support	Confidence	Lift	Rule
2	48.61	87.50	1.59	B → C
2	47.92	86.25	1.51	B → A
2	46.53	83.75	1.55	B → E
2	46.53	83.75	1.55	B → D
2	45.14	81.25	1.58	B → F
...	...	...	...	...

(c) 아이템의 수가 3인 연관성규칙발견의 결과

# of pages	Support	Confidence	Lift	Rule
3	44.44	95.52	1.68	B & D → A
3	42.36	91.04	1.68	B & D → E
3	41.67	89.55	1.92	B & D → H
3	39.58	85.07	1.75	B & D → I
3	39.58	85.07	1.68	B & D → G
...	...	...	...	...

연관성규칙발견에 의한 추천 프로세스는 먼저 배치처리(bath process)에 의해 아이템들 간의 연관성을 측정한 후, 웹 사용자의 방문시 연관성에 기초하여 실시간으로 아이템을 추천하는 과정으로 이루어진다. 이와 같은 프로세스를 사례를 통해 설명하면 다음과 같다.

먼저 배치처리에서는 일정한 기간(일주일 또는 한달)동안에 수집된 웹로그 데이터를 방문자의 성, 연령 등 인구통계학적 정보를 기준으로 나눈 후에, 각 데이터에 대해 최대 k개의 아이템을 가지는 연관성규칙발견을 실시해 놓는다.

실시간 추천과정에서는 배치처리에 의해서 얻어진 아이템들 간의 연관성을 기초로 추천이 이루어진다. 예를 들어, 새로운 고객이 30대의 여자라면 동일한 인구통계적 속성을 가진 사용자들이 가장 많이 방문하는 웹페이지들을 추천한다. 그 다음으로 이 고객이 특정한 웹페이지를 방문하였다면 그 웹페이지와 가장 연관성이 높은 웹페이지들을 추천한다. 예를 들어, <표 3.1>의 (a)에서와 같이 새로운 고객에게 30대 여자의 방문비율이 높았던 A, B, C, D, E를 추천하였고 이 고객이 그 중 B를 방문하였다면 (b)와 같은 결과를 기초로 C, A, D, E, F를 추천한다. 계속해서 이 고객이 웹페이지 B를 방문한 후 웹페이지 D를 방문하였다면 (c)와 같은 결과를 기초로 A, E, H, I, G를 추천한다. 이러한 과정을 반복하여 계속적으로 실시간 추천이 이루어지며, 만약 이 고객이 이 번 세션에서 k개 이상의 웹페이지를 방문하였다면 가장 최근에 방문한 k개의 웹페이지에 기초하여 추천이 이루어진다.

3.2 군집분석을 통한 추천 프로세스

웹 사용자의 패턴을 파악하고 분류하기 위해 주로 사용되는 다른 기법으로는 군집분석을 들 수 있다. 웹로그 데이터에 대한 군집분석에서는 보통 세션이 분석대상이 되며 세션 프로파일 행렬에 대해 유클리드 거리를 이용한 통상적인 군집분석 알고리즘(k-평균 군집분석 등)을 적용하는 것이 일반적이지만, 웹페이지들 간의 구조를 고려한 가중거리를 사용하기도 한다. 군집분석을 통한 추천 프로세스도 배치처리와 실시간 추천으로 이루어지는데, 그 과정을 사례를 통해 설명하면 다음과 같다.

먼저 배치처리에서는 일정한 기간(일주일 또는 한달)동안에 수집된 웹로그 데이터를 방문자의 성, 연령 등 인구통계학적 정보를 기준으로 나눈 후에, 각 데이터에 대해 별도의 군집분석을 실시해 놓는다.

실시간 추천과정에서는 배치처리에 의해서 얻어진 군집분석 결과를 기초로 추천이 이루어진다. 예를 들어, 새로운 고객이 30대의 여자라면 동일한 인구통계적 속성을 가진 사용자들이 가장 많이 방문하는 웹페이지들을 추천한다. 그 다음으로 이 고객이 특정한 웹페이지 C를 방문하였다면 그 웹페이지에 대한 평균 방문비율이 가장 높은 군집을 찾고 그 군집에 속한 고객들이 가장 많이 방문하였던 웹페이지들을 추천한다. 계속해서 그 고객이 다른 웹페이지 B를 방문하였다면 두 개의 웹페이지 C와 B에 대한 평균 방문비율이 가장 높은 군집을 찾고 그 군집에 속한 고객들이 가장 많이 방문하였던 웹페이지들을 추천한다. 이러한 과정을 반복하여 계속적으로 실시간 추천이 이루어지게 된다.

<표 3.2> 군집분석의 결과 예: 평균 프로파일

	전체	군집1	군집2	군집3	군집4
A	0.99	0.98	1	1	1
B	0.99	0.98	1	1	1
C	0.99	1	1	0.97	1
D	0.99	1	1	0.97	1
E	0.99	1	0.98	1	1
F	0.99	1	0.98	1	1
G	0.99	1	1	1	0.67
…	…	…	…	…	…
군집크기	144	52	49	37	6

A : /marvins
B : /1004angel
C : /login
D : /event/qlogin/event1
E : /so905
F : /contents/KMTV/asx/live
G : /bit801

4. 결론 및 토의

최근 웹로그 데이터 분석을 이용한 개인화 및 추천시스템에 많은 연구가 진행되고 있는데, 이는 방문자의 사용패턴에 근거하여 특정 아이템(상품, 기사, 컨텐츠 등)을 사용자마다 다르게 구성해 주거나 추천해주고자 하는 것이다. 그러나 웹로그 데이터에 대한 분석은 주로 전산학 및 경영과학 분야에서 주로 연구되어 왔기 때문에 통계학 분야에서 널리 알려진 많은 분석방법들이 아직은 제대로 응용되지 못하고 있다. 웹로그 데이터가 통계학 분야에서 기존에 다루어 왔던 데이터와는 다소 독특한 특성을 가지고 있으나 주성분분석 및 인자분석, 판별분석, 의사결정나무분석 등 통계적 자료분석 기법들은 웹로그 데이터에 대해서도 적절히 변형되어 사용될 수 있을 것이며, 따라서 이를 위한 지속적인 연구가 필요할 것으로 생각된다.

참고문헌

[1] 강현철 · 정병철 (2001). A Stedy of Web Usage Mining for eCRM, 『한국통계학회논문집』, 제8권 제3호, 831-840.

[2] 금종경(2001). eCRM을 위한 Web Personalization의 자동화에 관한 연구, 경희대학교 기술경영학과 석사학위 논문.

[3] 김영지 · 문현정 · 옥수호 · 우용태 (2002). 사례기반추론 기법을 이용한 개인화된 추천시스템 설계 및 구현, 『정보처리학회논문지D』, 제9권 제6호, 1009-1016.

[4] 박지선 · 김택헌 · 류영석 · 양성봉 (2000). 추천시스템을 위한 2-WAY 협동적 필터링 방법을 이용한 예측 알고리즘, 『정보과학회 논문지 : 소프트웨어 및 응용』, 제29권 제9호, 669-675.

[5] 서지현 (2002). 이중 소비자 구조를 고려한 개인화 시스템 설계 방안에 대한 연구, 경희대학교 산업공학과 석사학위 논문.