2002년 한국정보처리학회 춘계학술발표논문집 제9권 제1호

추천 시스템을 위한 웹 로그 분석

강태기, 김준태
동국대학교 컴퓨터공학과
e-mail : {ghappy,jkim}@dgu.ac.kr

Web Log Analysis for Recommendation Systems

Tae Ki Kang, Jun Tae Kim
Dept. of Computer Engineering, Dongguk University

요 약

협동적 추천은 사용자의 상품에 대한 구매 데이터를 이용하여 상품을 추천하는 방법이다. 그러나 구매 데이터가 희소한 경우 추천의 정확도가 떨어지는 문제점이 있다. 이러한 희소성 문제를 해결하기 위해서 클러스터링, SVD 등 다양한 방법이 제시되었으나, 근본적으로 사용자의 성향을 파악하기에는 부족한 점이 있다. 구매 데이터만을 이용했을 때의 문제점을 해결하기 위해서는 이를 보완할 수 있는 데이터의 활용이 필요하다. 웹 로그 분석을 통해서 구매 데이터의 희소성을 보완할 수 있으며, 사용자의 상품에 대한 부정적 반응을 구매 데이터에 반영할 수 있다. 본 논문에서는 웹 사이트에 접근 하는 사용자들에 의해서 만들어진 웹 로그를 분석하여 추천 시스템의 성능을 개선하였다.

1. 서론

인터넷의 발달과 함께 탄생한 인터넷 쇼핑몰은 이제 태동기를 지나 성숙기에 접어들고 있다. 미국의 온라인 서점인 아마존(www.amazon.com)의 경우 그동안의 닷컴 기업에 대한 우려를 불식시키는 흑자전환을 통해서 인터넷 쇼핑몰의 성공 가능성을 확인시켜주었다. 아마존의 경우 사용자의 기호에 적합한 상품을 추천해 주는 여러 가지 추천 방법을 사용하고 있으며, 대표적으로 협동적 추천을 사용하고 있다.

협동적 추천은 사용자의 상품 구매 성향과 비슷한 다른 사용자의 구매 내역을 비교하여 상품을 추천 하는 방법이다. 사용자에 대한 다른 정보가 없더라도 상품에 대한 선호도별로 사용자 집단을 분류하고 사용자와 선호도가 비슷한 사람이 구매한 상품을 추천한다. 협동적 추천은 사용자의 구매 데이터가 많을수록 추천 정확도를 높일 수 있으나, 구매 데이터가 희소하거나 없는 경우 추천이 불가능하거나 정확도가 떨어지는 단점이 있다. 실제 Amazon 이나 CDnow 의 경우 전체 상품의 개수에 대한 사용자 구매 데이터의 비율이 1%미만으로 사용자의 성향을 반영하는데 미흡한 점이 있다. 따라서 추천 정확도를 높이기 위해서는 사용자의 명시적인 정보 입력으로 이루어진 구매 데이터뿐만 아니라 묵시적인 정보를 활용하는 방법이 필요하다.

사용자가 인터넷 쇼핑몰의 웹 페이지에 접근하면 자동적으로 웹 로그 파일이 생성되게 된다. 이를 활용하면 사용자에 대한 의미 있는 행동 패턴을 찾아 낼 수 있다. 사용자가 웹 페이지를 방문하는 일정 순서를 순회라고 하며, 순회 패턴(Traversal Pattern)이란 일정 수 이상의 사용자들이 공통적으로 순회하는 웹 페이지의 순서를 말한다.

본 논문에서는 웹 사이트에 접근하는 사용자에 의해서 만들어진 웹 로그 데이터를 활용하여 구매 데이터에서는 얻을 수 없었던 순회 패턴 정보와 특정 상품에 대한 부정적 반응을 추출하여 구매 데이터와 함께 협동적 추천에 사용한다.

2. 관련 연구

2.1. 협동적 추천

협동적 추천이란 사용자들 사이의 유사도를 구하여 유사도가 높은 사용자들이 선호한 상품들을 추천하는 방법이다.

협동적 추천에 관한 연구로 Breese 는 피어슨 관계계수와 벡터 유사도(vector similarity)와 같은 메모리 기반(memory-based)방법과 확률적 방법인 베이시안(Bayesian)방식의 모델 기반(model-based)방법을 협동적

추천에 응용하는 연구를 수행하였다.

Helocker 는 다양한 방식의 유사도 계산과 여러 가지 방식의 유사도 가중치 실험을 하였다. 유사도 계산에는 피어슨 관계 계수, 스피어만 관계 계수, 벡터 유사도, 엔트로피(entropy)를 이용하였고, 선호도 값을 구하는 방법으로는 *평균 가중치(average rating), 유사 사용자의 상품 선호도 가중치 합*(deviation from mean), z 평균 점수(z score average)방법을 이용하여 실험하였다. 그 결과로서, 유사도를 구할 때 평가하는 선호도 값의 범위가 이산적인 경우에는 스피어만 관계 계수, 평가하는 선호도 값의 범위가 연속적인 경우에는 피어슨 관계 계수를 이용하는 것이 높은 정확도를 나타냄을 보였다. 또한 선호도 값을 구할 때는 전체적으로 유사 사용자의 상품 선호도 가중치 합을 이용하는 것이 높은 정확도를 보인다고 하였다.

Billsus 는 추론에 소요되는 계산 시간을 단축하기 위해서 특이 행렬 분해(Singular Value Decomposition)방법을 사용하여 상품의 차원을 축소하고, 유효 범위를 개선하고자 했다.

협동적 추천 방법에 적용되는 알고리즘은 메모리 기반(Memory-based)방식과 모델 기반(Model-based)방식으로 구분된다. 메모리 기반 방식은 전체 사용자 데이터 베이스를 이용하여 실제 사용자가 평가하지 않은 상품을 예측하는 것으로, 식(1)을 적용하여 상품의 선호도 예측값을 구한다.

$$P_{a,j} = \overline{v_a} + K\sum_{i=1}^{n} w(a,i)(v_{i,j} - \overline{v_i}) \qquad (1)$$

$P_{a,j}$ 는 사용자 a 의 상품 j 에 대한 선호도 예측값, $\overline{v_a}$ 는 사용자 a 의 선호도 평균, K 는 표준화 요소, $w(a,i)$ 는 사용자 I 와 실제 사용자 a 사이의 유사도 가중치를 나타낸다.

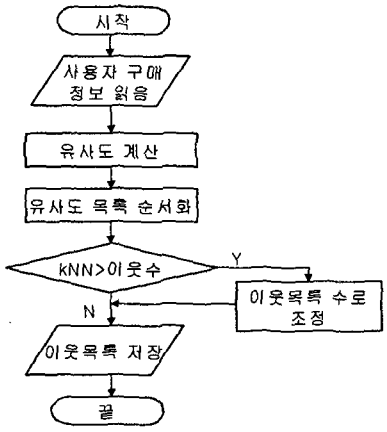
모델 기반 방식은 사용자 데이터베이스에서 모델을 학습하거나 추정하기 위하여 사용되며, 확률적인 관점에서 협동적 추천에 기존 고객의 평가를 이용하여 평가 기대값을 구하는 방식으로 크게 클러스터링과 베이시안 네트워크 모델(Bayesian Network Model)로 구분할 수 있다. 기대값을 구하는 방법으로 식(2)를 적용한다.

$$p_{a,j} = E(v_{a,j}) = \sum_{i=0}^{m} \Pr(v_{a,j} = i \mid v_{a,k}, k \in I_a) \times i \qquad (2)$$

I_a 는 사용자 a 가 평가한 상품 집합, m 은 평가 크기의 최대값, $v_{a,k}$ 는 사용자 a 가 상품 k 에 평가한 값, $P_{a,j}$ 는 사용자 a 가 상품 j 에 평가할 예측값을 나타낸다.

협동적 추천은 사용자간의 유사도를 계산하는 단계와 유사한 사용자들의 구매 이력을 바탕으로 상품에 대한 예측값을 계산하는 단계로 나눌 수 있다.

사용자간의 유사도를 계산하는 흐름은 [그림 1]과 같다.



[그림 1]

사용자간의 유사도를 계산하기 위한 대표적인 방식으로는 피어슨 관계 계수와 벡터 유사도(Vector Similarity)가 있다. 피어슨 관계 계수는 여러 분야에 많이 사용되지만 사용자의 평가수가 적은 경우, 사용자의 평가와 평균 평가의 차이가 0 이 되면 다른 사용자의 평가에 상관없이 사용자의 유사도는 *0* 이 되는 단점이 있다. 본 논문에서는 피어슨 관계 계수의 이러한 단점을 피하기 위해서 벡터 유사도를 사용하여 사용자 유사도를 구하는데 적용하였다. 벡터 유사도 공식은 유사도 관계를 차원에서의 거리가 아닌 각도를 이용하여 구하는 방식으로 다음의 식(3)으로 표현한다.

$$Sim(P,D) = \frac{P \cdot D}{\|P\| \|D\|} \qquad (3)$$

위 식에서 P,D 는 사용자의 상품에 대한 평가 데이터 벡터를 나타낸다.

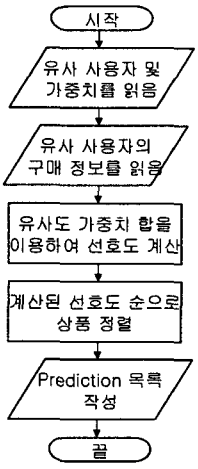
사용자들의 유사도를 기반으로 사용자의 상품 선호도 값을 예측하는데 사용되는 식은 유사 사용자의 상품 평가값의 유사도 가중치의 합이다.

구해진 사용자간의 유사도를 바탕으로 상품의 추천 목록을 작성하는 단계는 [그림 2]와 같다.

사용자들의 유사도를 기반으로 사용자의 상품 선호도 값을 예측하는데 사용되는 식(4)은 유사 사용자의 상품 평가 값의 유사도 가중치 합이 사용된다.

$$P_{a,i} = \overline{r_a} + \frac{\sum_{u=1}^{n} (r_{u,i} - \overline{r_u}) * w_{a,u}}{\sum_{u=1}^{n} w_{a,u}} \qquad (4)$$

$w_{a,u}$: a 사용자의 현 사용자에 대한 피어슨 관계 계수의 값



[그림 2]

2.2. 웹 마이닝

웹 마이닝은 웹 사이트와 관련된 데이터를 분석하여 유용한 정보나 지식을 발견해 내는 기술을 말한다. 웹 문서나 멀티미디어 자료를 분석하는 웹 컨텐트 마이닝(web content mining)은 대표적으로 STRUDEL, GroupLens 등의 시스템이 있다. 웹 로그 데이터를 분석하는 웹 유시지 마이닝(web usage mining)은 대표적으로 WebViz, WEBMINER 등의 시스템이 있으며, 탐사하려는 지식의 형태에 따라서 분류(classification), 클러스터링(clustering), 연관 규칙(association rule), 순회 패턴 탐사 등이 있다.

웹 로그 데이터로부터 사용자의 접근 패턴을 추출하는 기존의 연구는 패턴 추출시 룰을 만들거나, 패턴의 개수나 길이의 조절, 페이지별 헤더 테이블의 사용 등이 있다. 이러한 연구의 공통점은 웹 접근 경로를 그래프 혹은 트리로 표현한 후 접근 횟수를 기준으로 패턴을 추철한다는 점이다. 사용자가 자주 방문하는 페이지 일수록 사용자의 관심도가 높다는 가정하에 패턴을 추출한다. 그러나 이러한 방법은 몇가지 단점이 있다. 웹 설계상의 문제로 인해 특정 페이지를 방문하기 위해서 특정 경로를 거쳐야 하는 경우에는 특정 페이지가 패턴에서 제외될 가능성이 있다. 또한 페이지당 접근 횟수만으로 패턴을 추출할 경우 사용자가 해당 페이지에 접속했으나 곧이어 다른 페이지로 이동하거나 사이트를 나가는 킬러 페이지인 경우에도 패턴에 포함될 수 있다.

3. 웹 로그 분석 과정

웹 로그 데이터를 협동적 추천에 사용하기 위해서는 적절한 전처리 과정이 필요하다. 웹 로그 데이터가 갖고 있는 여러 가지 정보 중에서 협동적 추천에 이용할 수 있는 정보만을 추출하는 과정이 필요하며, 얻어진 정보를 구매 데이터에 반영하는 작업이 필요하다. 본 논문에서 웹 로그 데이터를 통해서 얻고자 하는 점은 다음과 같다.

첫째, 사용자의 순차 패턴을 탐사하여 순서별 가중치를 부과하여 상품 구매 데이터에 반영한다. 사용자가 인터넷 쇼핑몰에 접속해서 여러 상품을 보았다면 처음 본 상품이 가장 관심도가 높다고 할 수 있다. 다만, 같은 종류의 상품을 여러 개 보았다면 다른 가중치를 부과해야 한다.

둘째, 사용자가 방문한 웹 페이지 내역을 확인해서 특정 상품 정보 페이지에 접속한 후, 짧은 시간 이내에 다시 이전 페이지로 돌아가거나, 다른 페이지로 이동했다면, 방문했던 상품에 관심도가 떨어진다고 판단할 수 있다. 따라서 이러한 정보를 구매 데이터에 적용하면 상품에 대한 사용자의 부정적 관심도를 반영할 수 있다.

셋째, 사용자의 순회 패턴을 이용하여 사용자간 유사도를 구할 수 있다. 구매 데이터가 일치하지 않더라도 웹 사이트의 접근 경로가 비슷하다면 유사도가 높은 사용자라고 판단할 수 있다. 단, 웹 사이트의 구조상 모든 사용자가 공통적으로 지나가야 하는 페이지는 순회 패턴 유사도 비교에서 제외해야 한다.

웹 로그 파일 안에는 사용자의 IP 주소, ID, 웹 페이지에 접근한 날짜와 시간, 요청 방법, 접근한 페이지의 URL, 데이터 잔송에 사용된 프로토콜, 에러 코드, 전송 바이트 수 등에 대한 정보가 들어 있다. 이러한 웹 로그 파일을 추천 시스템에 사용하기 위해서는 다음과 같은 절차를 거쳐야 한다.

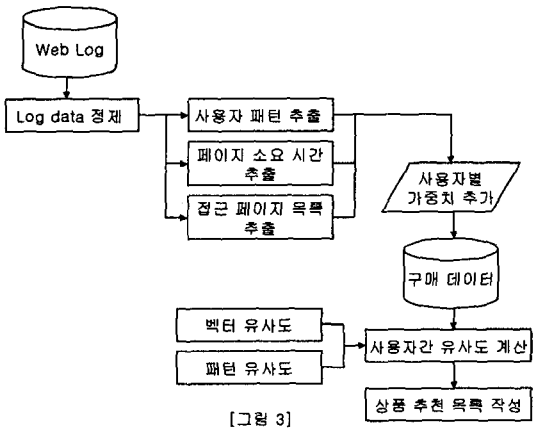
1. 웹 로그 데이터를 대상으로 파싱(parsing)하는 과정과 이용자의 IP 주소, 이용자의 ID, 웹 문서에 접근한 날짜와 시간, 접근한 문서의 URL 을 제외한 다른 불필요한 항목들을 정제(cleaning)한다.
2. 사용자가 접근한 웹 페이지 각각을 트랜잭션으로 간주하고, 적용 가능한 접근 시간 내에 사용자별로 한 세션 동안 거쳐간 웹 페이지 운행 경로를 이용하여 트랜잭션 시퀀스를 형성한다.
3. 트랜잭션 시퀀스 안의 운행 웹 페이지 각각을 대상으로 데이터 시퀀스가 될 최후 전진 문서(last forward document)집합을 추출한다.
4. 사용자별 페이지 소요 시간을 측정하고, 짧은 시간이내에 페이지를 이탈한 경우 구매 데이터에 부정적 관심도를 반영한 가중치를 부여한다.

이러한 웹 로그 분석과정이 끝나면 웹 로그 분석에서 얻어진 사용자의 부정적 반응, 페이지 접근 데이터, 사용자의 순회 패턴을 이용한 사용자간 유사도의 정보가 추천 시스템에 적용된다.

4. 시스템 구현

추천 시스템은 크게 웹 로그 분석 과정과 유사도 계산 상품 추천의 3 단계로 구성된다.

전체 시스템 구성도는 아래 [그림 3]과 같다.



[그림 3]

웹 로그 분석은 데이터의 정제 과정을 거쳐서 사용자의 패턴, 페이지 소요 시간, 접근 페이지 목록을 추출하게 된다. 이러한 정보의 추출은 각각 아래와 같은 방법으로 사용된다.

- ✓ 사용자의 패턴 : 사용자의 웹 페이지 접근 패턴을 추출하여 사용자별 테이블에 저장하게 된다. 구매 데이터를 이용하여 벡터 유사도를 구하는 것과 같은 방법으로 사용자의 패턴을 이용하여 벡터 유사도를 구하면 구매 데이터에서 추출하지 못했던 이웃 목록을 만들 수 있게 된다.
- ✓ 페이지 소요 시간 : 페이지 소요 시간을 추출하여 사용자의 상품에 대한 부정적 반응을 가중치에 반영한다. 상품 정보에 해당하는 페이지의 소요 시간이 현저하게 적거나 다른 페이지로 곧이어 이동하는 경우 해당 상품에 대한 가중치를 마이너스로 두어서 추천 시스템에서 부정적 반응을 반영하게 된다.
- ✓ 접근 페이지 목록 : 사용자의 웹 페이지 접근 목록을 통해서 구매 데이터에는 나타나지 않으나 사용자가 접근한 페이지에 대해서 가중치를 부여하게 된다. 이를 통해서 구매 데이터에 반영된 명시적인 사용자 반응과 함께 묵시적인 사용자 반응을 추천 시스템에 반영하게 된다.

본 논문의 추천 시스템의 성능을 평가하기 위한 데이터는 ACM 의 6th SIGKDD 에서 주최한 KDD Cup 2000 대회의 데이터를 이용하였다.

실험 데이터는 Gazelle.com(legwear, legcare 전자상거래 업체)의 클릭 스트림과 구매 데이터로 이루어져 있으며 2000 년 1 월 30 일부터 3 달간의 웹 로그 데이터와 구매 데이터로 이루어져있다. 데이터에 포함된 내용은 다음과 같다.

[웹 로그 데이터]
- session : date/time, cookie, browser, visit count, referrer
- page views : URL, processing time, product, assortment

[구매 데이터]
- order header : customer, date/time, discount, tax, shipping
- order line : quantity, price, assortment

5. 결론 및 향후 과제

웹 로그 분석을 통해 추출된 정보를 사용하여 추천 시스템에 적용하면 기존의 추천 시스템과 비교하여 다음과 같은 장점을 얻을 수 있다. 첫째, 사용자의 웹 페이지 접근 패턴 유사도 비교를 통해서 구매 데이터를 이용한 벡터 유사도 계산에 비해서 정확한 이웃 목록을 추출할 수 있다. 둘째, 페이지 소요 시간을 통해서는 기존의 추천 시스템에서 반영되지 못했던 사용자의 부정적 반응을 반영할 수 있게 된다. 셋째, 구매 데이터의 희소성을 보완할 수 있다. 구매 데이터만을 이용하는 경우 명시적인 사용자 반응만을 사용하게 되지만, 웹 로그를 통해서 얻어진 페이지 뷰를 이용할 경우 사용자가 접근한 페이지 목록의 가중치 부여를 통해서 데이터 희소성 문제를 보완할 수 있다. 또한 순차 패턴을 통하여 페이지 방문 순서에 따른 서로 다른 가중치 부여를 통해 사용자의 관심도를 보다 정확히 반영할 수 있다.

이러한 웹 로그 분석을 통해서 구매 데이터가 갖고 있는 한계를 극복 할 수 있으며, 추천 시스템의 추천 정확도를 개선 할 수 있다. 향후 과제로 웹 로그 데이터의 순회 패턴을 통한 유사도 비교 방법의 최적화를 위한 알고리즘 개선과 사용자의 웹 페이지 스크롤, 출력, 복사 등의 정보를 활용할 수 있는 방법을 생각해 볼 수 있다. 또한 구매 데이터의 희소성을 해결하기 위한 Default voting 방법에 대한 연구가 필요하다.

참고문헌

[1] Daniel Billsus, Michael J. Pazzani, "Learning Collaborative Information Filters", proceedings of workshop on recommender system, 1998.

[2] John S. Breese, David Heckerman, Carl Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", 14th conference of UAI-98, 1998.

[3] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl, "An Algorithmic Framework for Performing Collaborative Filtering", Proceedings of conference on Research and development in information retrieval, 1999.

[4] Ramakrishnan Srikant, Yinghui Yang, "Mining Web Logs to Improve Website Organization", The 10th international World Wide Web conference, 2001.

[5] Karuna P Joshi, Anupam Joshi, Yelena Yesha, Raghu Krishnapuram, "Warehousing and mining Web logs", Proceedings of the second international workshop on Web information and data management, 1999.