Support Vector Regression을 이용한 연속성 피드백 정보의 협동 추천 시스템

임민택 전성해 오경환
서강대학교 컴퓨터학과
{taigi@ailab, shjun@ailab, kwoh@ccs}sogang.ac.kr

Collaborative Recommendation System of Continuous Feedback Information Using Support Vector Regression

Min-Taik Lim Sung-Hae Jun Kyung-Whan Oh
Dept. of Computer Science, Sogang University

요 약

인터넷으로부터 필요한 정보를 얻기 위하여 무의미한 탐색을 반복하는 경우가 자주 나타나고 있다. 이러한 Dizzy Web에서 사용자와 관련 있는 정보를 추천해 주는 방법에 대한 연구가 많이 진행되고 있다. 특히 협동 추천시스템에 대한 연구가 활발히 진행되고 있다. 이 시스템의 구현 알고리즘 중에서 기존의 메모리 기반은 수행 시간에 대한 부담이 매우 크며, 모델 기반은 연속성 데이터에 대한 처리가 어렵거나 불가능하다는 문제가 있다. 본 논문에서는 특히 웹 사용자 모델에서 효과적인 연속성 피드백 데이터를 이용한 사용자 모델링 방법을 제안하고 이를 통해 웹 페이지 예측을 수행하는 시스템을 구현하였다. 논문에 사용된 연속성 데이터는 사용자의 웹 페이지 방문시간이고 이 데이터를 분석하기 위해 기존의 모델 기반 알고리즘에 Support Vector Regression 기법을 결합하는 알고리즘을 설계하였다. 실험에서는 제안 모델의 정확성과 예측 능력에 대하여 기존의 Pearson 알고리즘과 비교하였다. 논문에서 제안하는 방법이 매우 적은 시간 비용을 요구하면서도 유의할 수 있는 수준의 결과가 얻을 수 있음이 확인되었다.

1. 서 론

인터넷 사용자들이 그들이 원하는 정보를 보는 시간이 전체 소비 시간 중 42%에 지나지 않고 전체 인터넷 사이트의 51%는 웹 페이지에서 제시하는 내용을 쉽게 알 수 없으며 90%이상이 적절치 못한 구조를 가지고 있다는 연구결과가 발표되었다[1]. 따라서 비효율적인 인터넷에서 사용자가 경제적으로 정보를 수집할 수 있는 방법이 필요하게 되었다. Personalized Web은 이와 같은 문제를 해결하려는 연구 분야 중 하나로서 각종 정보로부터 사용자의 성향을 파악하고 이를 기반으로 웹사이트를 적응, 변화시키며 서비스를 제공하는 방법이다. 즉, 이에 대한 연구는 해당 사이트로부터 효과적으로 사용자에게 적절한 정보를 제공하고자 하는 것이 일차 목표이고 사용자에게 특정 정보만을 추려서 제공함으로써 시스템의 부하를 줄이고 성능 향상을 추구하는 것이 두 번째 목표이다. 궁극적인 목표는 웹사이트에서 사용자의 여러 반응을 분석하여 최적의 추천 시스템을 구축하는 것이다. 웹 개인화를 위한 연구 중에서 사용자 모델링(user modeling)은 관심을 갖게 되는 분야이다. 사이트를 찾아온 사용자가 어떤 부류에 속하고 이용 패턴 및 전반적인 성향은 어떤지를 구체화하여 이를 시스템에서 이용할 수 있는 형태로 모델링하는 연구이다. 추천 시스템의 사용자 모델링에서 중요한 요소 중 하나는 사용자로부터 얻어지는 피드백이다. 주어진 컨텐트에 대한 사용자의 반응으로부터 사용자의 성향을 파악하고 사용자에 맞는 상품, 정보, 페이지를 제공한다. 일반적으로 피드백은 명시적 피드백(explicit feedback)과 암시적 피드백(implicit feedback)으로 구분되며 명시적 피드백은 컨텐트, 상품 등에 대해 사용자로부터 직접 얻어지는 정보를 의미하고 암시적 피드백은 마우스의 움직임, 페이지에 머문 시간, 페이지간의 이동 등과 같이 사용자의 행동으로부터 간접적으로 관찰될 수 있는 정보를 말한다. 현재 구현되고 있는 대부분의 추천시스템은 명시적 피드백 중 사용자로부터의 등급평가 정보만을 이용하고 있으며, 이 경우 전체 사용자로부터 반응을 얻기가 어렵기 때문에 데이터의 희소성 문제를 유발한다. 등급평가 정보와 같은 이산 데이터가 아닌 많은 경우의 피드백들이 나타내는 연속성 데이터를 처리할 수 있는 방법이 현재 거의 없는 실정이다. 본 논문은 추천시스템에 적용될 수 있는 사용자 모델링을 구현에 있어서 어느 웹사이트에서나 손쉽게 얻어질 수 있는 로그 데이터를 기반으로 사용자로부터 연속성 피드백을 이용하여 최근 빠른 학습 속도와 비교적 높은 정확성으로 패턴인식 분야에서 많은 연구가 되고 있는 Support Vector Regression(SVR)을 사용한 모델링 방법을 제시하였고 객관적인 데이터를 통하여 그 성능을 실험하고 검증하였다. 2절에서는 제안하는 연속성 피드백 데이터 기반의 웹 페이지 예측 모형의 구축에 대한 이론적 배경에 대해 알아보고 다음으로 비선형 회귀 모형을 통한 웹 페이지 예측 시스템을 제안하고 4절에서는 구현 및 실험결과를 알아보고 마지막 5절에서는 결론 및 향후 연구에 대해서 논의한다.

2. 협동 추천시스템과 SVR

2.1 협동 추천(Collaborative Recommendation)

협동 추천은 기본적으로 사용자들의 아이템에 대한 평가 정보를 기반으로 하여 특정 사용자의 특정 아이템에 대한 유용성과 선호도에 대한 예측을 목적으로 한다[2]. 협동 추천시스템이 공통적으로 지니는 문제점 중 하나는 결여 데이터(missing data) 문제이다. 대부분의 경우에 전체 아이템에 대하여 모든 사용자들의 평가를 기대하기는 어려울 뿐만 아니라 암시적 피드백 기반보다 명시적 피드백 기반 시스템에 있어 더욱 문제가 된다. 명시적 피드백 시스템의 대표적인 문제점은 사용자에 의해 부여되는 이산 데이터의 점수만을 이용하는 것이며 연속성

데이터를 처리할 수 있는 방법을 제시하지 못한다는 단점을 지니고 있다. 현실적으로 많은 시스템에서 사용자에 의해 정확히 부여되는 이산데이터를 얻는 것은 매우 제한되어 있으며 오히려 시스템을 통해 얻어지는 연속형 데이터를 사용해야 하는 경우가 많다. 사용자의 편의 측면에서 보았을 때에도 명시적 피드백을 얻는 시스템은 궁극적으로 사용자에게 불편함을 주게 된다.

2.2 SVM과 SVR

SVM(Support Vector Machine) linear classifier

Vapnik은 주어진 데이터들을 이분법적으로 나눌수 있는 이상적인 선형평면을 구하는 방법을 제시하였다[3]. 평면방정식이 주어졌을 때 분류문제를 해결하는 함수식은 다음과 같다.

$$f(x) = sign\ (w \cdot x + b) \qquad (1)$$

식(1)의 함수식 부호에 의해 부류가 결정된다. 그림 1은 실제 문제 공간에서 이 평면과 방정식이 어떻게 표현되고 적용될 수 있는지 보여준다.

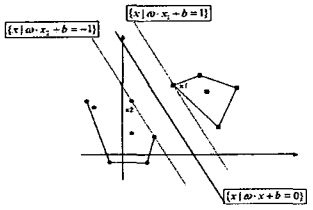

그림 1: 이상 평면(Optimal Hyperplane)의 도식화 표현

그림 1에서 중앙의 굵은 직선을 구해내는 것이 SVM의 최종 목표이다. 이러한 이상 평면은 각 인스턴스들과의 폭을 최대로 하고 분류기로서의 몇 가지 조건들을 만족한다. 따라서 주어진 인스턴스들로부터 간격(margin) 폭을 최대화하고 몇 가지 조건식을 만족하는 평면의 방정식을 구해야 한다. 이상평면은 다음의 식을 만족해야 한다.

$$y_i(\langle w \cdot x_i \rangle + b) \geq 1, \quad i = 1,\ldots,l \qquad (2)$$

식 (2)에서 점 x 와 평면과의 거리는 다음과 같이 정의된다.

$$d(w,b,x) = \frac{|\langle w,x_i \rangle + b|}{\|w\|} \qquad (3)$$

이상 평면(optimal hyperplane)은 위의 식 (3)을 만족하고 점과 평면 사이의 간격(margin)을 최대화 하는 w와 b를 구한다.

SVR(Support Vector Regression)

SVM은 손실함수(Loss function)를 이상 평면 방정식에 포함시킴으로서 회귀(regression) 문제에 적용 될 수 있다[4]. 손실함수란 기대값과 측정값의 오차를 정의하는 함수식이다. 본 논문에서는 희소데이터에 대해 우수한 성능을 갖는 ε-Insensitive 손실함수를 사용한다.

$$D = \{(x^1,y^1),\Lambda,(x^l,y^l)\}, \qquad x \in R^N, y \in R \qquad (4)$$

$$f(x) = \langle w \cdot x_i \rangle + b. \qquad (5)$$

식 (4)와 같은 데이터 구조를 식 (5)의 직선식으로 근사하는 최적의 회귀 함수는 SVM에서 다음 문제로 표현된다.

$$\text{minimize} \quad \frac{1}{2}\|w\|^2,$$

$$y_i - \langle w \cdot x_i \rangle - b \leq \varepsilon,$$

$$\text{subject to} \quad \langle w \cdot x_i \rangle + b - y_i \leq \varepsilon \qquad (6)$$

ε-Insensitive 손실함수와 Lagrange 함수를 이용하여 식 (6)의 문제를 풀면 다음과 같은 해를 얻게 된다.

$$w = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i$$

$$b = -\frac{1}{2}\langle w, (x_r + x_s) \rangle \qquad (7)$$

3. 연속성 피드백 정보를 이용한 웹 페이지 예측 모형

3.1 웹 페이지 예측 시스템 절차

본 논문의 전체 시스템은 그림 2와 같이 5단계의 절차로 이루어진다.

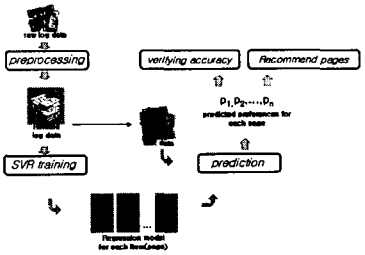

그림 2: 전체 시스템의 5단계 절차

로그 파일은 로그를 생성해 낸 서버의 종류와 웹 사이트의 성격 그리고 사이트 제공 주체에 따라 다양한 형태를 보인다. 최초의 로그 데이터는 전처리(preprocessing)과정[5]을 통해 모델 구축에 필요한 데이터로 표현된다. 정제된 로그 정보는 사용자가 방문한 페이지와 머문 시간을 하나의 인스턴스로 이용하고 적절한 수의 인스턴스를 추출하여 학습데이터로 사용하는 SVR 모델을 구축한다. 특히 이 모델은 모든 웹 페이지들에 대하여 개별적으로 작성된다. 다음으로 테스트 데이터의 각 페이지를 SVR 모델에 적용시켜 각 페이지에 대한 사용자의 선호도를 비교하여 모델에 대한 타당성을 조사한다. 테스트 데이터는 정제된 로그 파일에서 학습에 사용되지 않은 데이터로서 단순 임의 추출에 의해 구성된다. 마지막은 위 과정들을 통해 얻어진 각 페이지들에 대한 예측 값에 각 사용자의 평균 관심도와 각 페이지의 평균관심도를 고려하여 선호도를 계산하여 우선순위가 높은 페이지를 추천하는 단계이다. 전체 시스템에 대한 평가 방법으로서 실제 주어진 선호도와 예측 선호도의 오차정도를 MSE로 측정하였고 우선순위가 높은 페이지와 낮은 페이지를 추출하여 실제로 해당 페이지에 대하여 어떠한 반응을 보였는지를 측정함으로서 본 시스템의 성능을 평가하였다.

3.2 SVR 모델 학습

SVR 모델은 각 페이지에 대하여 구축되며 해당 페이지를 제외한 나머지 페이지들에 대한 선호도를 축으로 하는 회귀 모형이 된다. 그림 3에 이를 개념적으로 표현하였다.

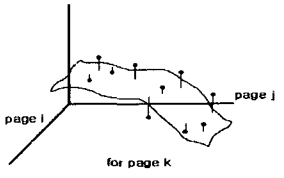

그림 3: 연속성 피드백 데이터를 갖는 웹 사용 예측

그림에서 각 점은 사용자 한 명을 나타내며 평면에 있는 축들은 사용자가 지나온 페이지 브라우징 시간, 세로축은 해당 페이지에 대한 브라우징 시간이다. 따라서 그림은 해당 페이지를 제외한 다른 페이지의 브라우징 시간에 따라 사용자의 해당 페

이지에 관한 브라우징 시간을 예측하는 과정을 나타내고 있다. 평면은 학습된 회귀함수를 나타내며 사용자를 나타내는 각 점은 학습 시에 사용된 인스턴스가 된다.

3.3 선호도 예측

테스트 데이터를 생성하여 각 페이지별로 구축된 SVR 모델에 입력하면 해당 페이지의 브라우징 시간에 대한 예측값이 출력된다. 이 값은 제안 방법에서 두 가지에 이용된다. 하나는 테스트 데이터의 실제 해당 페이지 브라우징 시간과의 비교를 위한 데이터로서 이용되고, 다른 하나는 이를 다시 선호도로 변환하여 실제 사용자에게 추천하는 페이지 선택에 사용된다. 예측된 브라우징 시간을 선호도로 변환하기 위해 페이지에 대한 특성과 각 사용자에 대한 특성을 식 (6)으로 반영하였다.

$$PREF_{u,k} = \frac{\mu_u + \mu_k}{2} + \frac{(P_{u,k} - \mu_u)}{\sigma_u} \cdot \frac{(P_{u,k} - \mu_k)}{\sigma_k} \qquad (8)$$

where, $P_{u,k}$: Predicted time browsing page k for user u

μ_u ,μ_k : the mean browsing time for user u and page k

4. 구현 및 실험

4.1 Data 및 전처리

본 실험에서는 인터넷 쇼핑몰(Gazelle.com, KDD Cup 2000, 1.2GB)의 2개월 동안의 로그 데이터를 이용하였다[3]. 한 개의 로그 정보는 217개의 attribute로 구성되어 있다. 전처리는 유효 사용자 추출, 특성 attribute 데이터 추출, 시간 데이터 변환의 3가지 과정을 거쳤다. 사용자의 구분은 Cookie ID를 이용하였고 페이지 방문시간은 페이지 요청 처리시간부터 다음 클릭 스트림이 발생할 때까지의 시간 간격으로 계산하였다. 생성시킨 인스턴스의 수는 각 페이지 별로 생성 가능한 데이터의 66%를 무작위로 추출하여 학습데이터로 사용하고 나머지 34%는 학습의 성능 확인을 위해 사용하였다.

4.2 모델의 정확도와 시간 비용에 관한 실험

제안하는 SVR기반의 연속성 데이터를 이용한 웹 사용자 모델의 성능을 MSE(mean squared error)를 통해 Pearson 방법과 비교하였다[6].

	SVR	Pearson
MSE(전체)	1.75	1.37
MSE(상위 50%)	1.19	1.01

표 1: 모델의 정확성 비교

표 1은 전체 모델에 대하여 페이지 별로 각각 생성된 테스트 데이터에 대한 결과들의 MSE값이다. Pearson의 MSE가 약간 작음을 알 수 있다. 하지만 그림 4에서 볼 수 있듯이 계산 시간은 제안 방법이 훨씬 빠름을 알 수 있다. 특히 데이터의 크기가 커질수록 그 차이는 더욱 커짐을 알 수 있다.

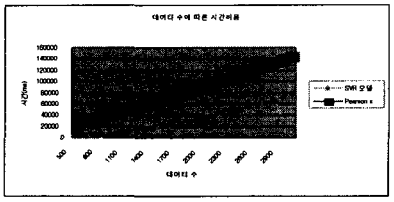

그림 4: 데이터 수에 따른 시간 비용

따라서 SVR 기반의 웹 사용자 모델 방식을 통하여 실시간 예측이 가능하게 된다. 이는 인터넷 기반의 여러 예측 시스템에서 효과적으로 사용할 수 있다.

4.3 웹 페이지 추천 시스템의 성능 실험

제안된 SVR 기반 사용자 모델링을 이용한 추천시스템의 성능을 평가하기 위하여 20 페이지 이상을 방문한 사용자 150명에 대하여 10개의 페이지는 이미 방문한 페이지로 설정하고 나머지 10개의 페이지에 대하여 예측한 선호도를 이용해 순위도(raking rate)를 구하고 순서대로 예측 페이지의 30%인 3개씩의 HIGH preference item과 LOW preference item을 선택한 후, 실제 데이터에서 보이는 선호도와 비교하였다. 결과는 표 2와 같다.

	SVR	Pearson
Pr(high/high)	0.31	0.35
Pr(low/low)	0.29	0.31
Pr(high/low)	0.18	0.16
Pr(low/high)	0.15	0.13

표 2: 웹 페이지 예측 시스템 정확도

즉, SVR은 Pearson알고리즘과 비교 할 때 성능의 차이를 크게 보이지 않으면서도 빠른 시간 내에 학습과 예측이 가능하였다.

5. 결 론

본 논문은 SVR 기법을 이용하여 연속성 피드백 기반의 웹 페이지 사용자 모델링을 제안하여 웹 페이지 추천시스템에 적용하였다. SVR 기반의 사용자 모델링 방법의 장점은 기존의 모델기반 방식의 협동 추천시스템에서는 제한되었던 연속성 피드백 정보를 다룰 수 있는 것이다. 실험을 통한 제안 방법의 성능 평가 결과 간단한 피드백 정보인 방문시간 만을 기초한 모델링을 통해서도 정확도는 유의할 만한 수준을 보였고 기존의 Pearson 방법에 비하여 무시할 만한 정확도의 차이에서 시간비용의 성능이 매우 우수함을 알 수 있었다. 우수한 성능의 시스템을 구축하기 위해서 사용자의 선호도를 보다 잘 표현할 수 있는 정보를 선택할 수 있는 연구가 필요하고 본질적으로 협동 추천시스템이 가지고 있는 결여데이터에 대한 해결방안도 고려될 때 제안 모델은 더 나은 성능을 기대할 수 있을 것이다. 제안 방법은 웹 페이지 예측 시스템 뿐만 아니라 개인화 웹 사이트의 구현, 전자상거래의 추천 시스템, 사용자 인터페이스 에이전트의 개발, 개인에게 특화된 정보검색 시스템 등 다양한 시스템 구축에 유용하게 사용될 수 있을 것이다.

감사의 글

본 연구는 과학 기술부 주관 뇌신경정보학 사업에 의해 지원되었음.

참 고 문 헌

[1] Forrester Research, "educorner.com/courses", 2002.

[2] Basu,C et al., Recommendation as classification: Using Social and Content-based Information in Recommendation, Proceedings of the Workshop on Recommendation system. AAAI Press, Menlo Park California, 1988.

[3] V. Vapnik et al. " Support vector networks" Machine Learning 20, 273-297 1995.

[4] Vapnik., "Statistical Learning Theory", Wiley, N.Y., pp.445-448, 1998.

[5] Recardo Baeza-Yates et. al., "Modern Information Retrieval", ACM Press, pp. 6-8, 1999

[6] J. Han ,M. Kamber, "Datamining" pp.435-436, 2001