한국 퍼지 및 지능시스템 학회 2003, Vol. 13, No. 4, pp. 433-438

앙상블 Support Vector Machine과 하이브리드 SOM을 이용한 동적 웹 정보 추천 시스템

Dynamic Recommendation System of Web Information Using Ensemble Support Vector Machine and Hybrid SOM

윤경배 · 최준혁

Kyung-Bae Yoon, Jun-Hyeog, Choi

김포대학 컴퓨터계열

요 약

최근, 인터넷 쇼핑몰과 같은 웹 사이트를 대상으로 각 사용자에게 가장 필요한 정보를 제공하기 위한 웹 정보 추천 시스템에 대한 연구가 활발히 진행되고 있다. 사용자 프로파일과 명시적 피드백에 의존하는 대부분의 웹 정보 추천 시스템의 경우 사용자의 다양하고 정확한 정보를 필요로 하지만 실세계의 문제에 있어 이러한 연관 정보를 얻기란 쉽지 않다.
본 논문에서는 사용자의 명시적 피드백과 프로파일에 의존하지 않는 웹 정보 서비스를 위한 정보 예측 기법을 제안한다. 이를 위해 앙상블 Support Vector Machine과 하이브리드 SOM을 설계하고 적용하여 웹 로그 데이터의 희소성 문제를 해결하면서 대용량 웹 데이터로부터 사용자에게 꼭 필요하고 유용한 정보를 추천할 수 있는 동적 웹 정보 예측 시스템을 설계하고 구현한다.

Abstract

Recently, some studies of a web-based information recommendation technique which provides users with the most necessary information through websites like a web-based shopping mall have been conducted vigorously. In most cases of web information recommendation techniques which rely on a user profile and a specific feedback from users, they require accurate and diverse profile information of users. However, in reality, it is quite difficult to acquire this related information. This paper is aimed to suggest an information prediction technique for a web information service without depending on the users' specific feedback and profile. To achieve this goal, this study is to design and implement a Dynamic Web Information Prediction System which can recommend the most useful and necessary information to users from a large volume of web data by designing and embodying Ensemble Support Vector Machine and hybrid SOM algorithm and eliminating the scarcity problem of web log data.

Key Words : Web information recommendation system, Hybrid SOM, Ensemble Support Vector Machine

1. 서 론

웹 개인화(personalized web) 기법에서, 사용자 모델링을 수행함에 있어 가장 중요한 요소는 사용자들로부터 얻어지는 피드백 정보이다.

현재 구현되고 있는 대부분의 추천 시스템은 명시적 피드백 정보 중에서 사용자의 등급 평가 정보만을 이용하고 있다. 이럴 경우 전체 사용자로부터 안정적인 반응을 얻기가 힘들기 때문에 데이터의 희소성 문제를 유발하고, 등급평가 정보와 같은 이산적인 데이터가 아닌 많은 경우의 연속형 피드백 정보일 때 나타나는 연속성 데이터를 다룰 수 있는 방법이 없다.

본 논문은 동적 추천 시스템에 적용될 수 있는 사용자 모델링을 구현함에 있어, 웹 로그 데이터를 기반으로 사용자로부터 얻은 암시적 피드백 정보를 이용한 앙상블 Support Vector Machine(Ensemble Support Vector Machine)과 하이브리드 SOM(Hybrid Self-Organizing feature Maps)을 이용한 동적 웹 정보 예측 시스템을 설계하고 구현한다.

접수일자 : 2003년 1월 16일
완료일자 : 2003년 5월 13일
감사의 글 : 본 연구는 2003학년도 김포대학의 연구비 지원에 의해 연구되었음

2. 관련 연구

2.1 연관성 기반의 추천 시스템

기업은 추천 시스템을 통해 개별 고객에게 새로운 상품을 추천하거나 효과적인 광고를 통해 전체적인 사용자의 만족도를 높혀 매출 증가를 유도하고자 한다. 물론 기업뿐 아니라 각종 웹 서버의 주체들도 추천 시스템을 통하여 이와 유사한 효과를 거두고자 한다. 이러한 추천 시스템은 사용자의 인구 통계학적 정보와 거래 행위를 바탕으로 한 추천 모형이 필요하다.

추천 시스템 중에서 연관성 기반의 추천 방식은 내용

기반(content-based) 방식, 연관성(association) 방식, 인구 통계학적(demographic) 방식, 협동추천(collaborative) 방식으로 분류할 수 있다. 연관성 기반의 추천 시스템의 주된 알고리즘은 피어슨의 상관 계수를 이용한다.

피어슨의 상관계수를 이용하는 방법은 아이템간의 유사성을 측정하여 사용자가 선택한 아이템에 대하여 유사도가 높은 아이템을 사용자에게 추천하는 방식으로 대부분의 웹 문서 추천 시스템에서 사용된다. 이 방법의 문제점으로는 높은 예측의 정확도를 보임에도 불구하고 학습시간이 너무 오래 걸리는 단점을 들 수 있다.

2.2 SOM을 이용한 추천 시스템

SOM은 인간 두뇌의 경쟁적 시스템으로부터 모형화 되었기 때문에 다른 신경망 모형에 비해 인간의 두뇌에 가까우며 자료에 대한 학습 시간이 상당히 빠른 모형이다. 이러한 SOM의 빠른 학습 능력은 실시간으로 빠르게 학습이 진행되어야 하는 인터넷 정보검색 에이전트의 웹문서 그룹화와 분류 문제 등에 적용할 수 있다.

SOM은 실세계의 여러 문제에 광범위하게 적용되었고 많은 중요한 연구들이 진행되어 왔지만 다음과 같은 해결되지 않은 문제점들이 있다[2]. 첫째로, SOM은 자료 공간(data space) 내에서 밀도 함수(density function)가 정의되어 있지 않다. 이는 학습 데이터의 분포와 가중치 분포간의 관계가 정형화되는 방법이 존재하지 않음을 의미한다. 둘째로 SOM의 학습 알고리즘에는 오차(error)를 관리하는 목적 함수(objective function)를 최적화하는 과정이 없다. 따라서 이러한 목적 함수의 역할을 담당하는 확률적 분포를 이용하여 목적 함수를 구축하고 최적화 하는 작업이 필요하다. 셋째로 SOM의 학습 알고리즘이 항상 수렴한다는 일반적인 보장이 없다. 특히, 학습이 끝난 최종 모형이 국지적 최적값(local optimal value)으로 빠지는 경우가 종종 있게 된다. 따라서 학습 데이터를 이용한 모델이 허용 한계 내에서 전역적 최적값(global optimal value)에 수렴하게 하는 방안이 필요하다. 넷째로 SOM은 모형의 모수로 선택되는 최적값을 결정하는 이론적인 구조가 없다. 즉, SOM은 입력층으로부터 출력층으로의 연결을 통한 가중치의 갱신을 통한 모형이지만 새로운 입력에 대한 예측값을 계산할 때 예측값이 나오게 된 과정보다는 예측값 자체에 의미를 두기 때문에 예측 모형에 대한 설명이 절대적으로 부족하다. 예를 들어 학습률(learning rate)에 대한 초기값의 결정과 변화규칙, 이웃 함수(neighborhood function)의 감소율 등에 대한 함수, 가중치의 갱신 분포 등에 대한 이론적 구조가 전혀 없다. 이렇게 이론적 바탕이 부족한 신경망 모형에서 예측 결과값에 대한 모형적 해석을 가능하게 하는 방안이 필요하다. 다섯번째로 SOM은 다른 SOM 모형이나 이질적인 구조(architectures)의 학습 모형들과의 비교가 불가능하다. 이는 모형에 대한 명확한 수리적 규칙이 정의되지 않기 때문이다. 따라서 일반적인 모형 평가의 수리적 측도를 SOM 모형에도 적용시켜야 한다. 이러한 SOM의 해결할 수 없는 모형적 한계를 극복하는 방안이 하이브리드 SOM을 이용한 방법이다. 본 논문에서 제안하는 하이브리드 SOM은 SOM의 빠른 학습 능력을 유지하면서 군집 결과에 대한 수렴성을 보장하는 모형이다.

2.3 Support Vector Machine을 통한 추천 시스템

1970년대 후반 Vapnik은 주어진 데이터, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$을 이분법적으로 가장 잘 나눌 수 있는 선형평면을 구하는 방법인 Support Vector Machine(SVM)을 제안하였다[4, 5].

임의의 평면방정식이 주어졌을 때 분류 문제를 해결하는 함수식은 식 (1)과 같이 정의된다.

$$f(x) = sign(\omega \cdot x + b) \tag{1}$$

식 (1)에서 x는 입력변수, y는 출력변수를 나타낸다. y는 +1과 -1의 두 값을 갖으며, 이 값에 의해 주어진 개체들을 분류한다. 커널(kernel) 함수인 $sign(\cdot)$의 역할은 주어진 학습 데이터로부터 분류를 가장 잘 수행하는 모형 형태의 결정으로, 최종적인 분류는 $f(x)$의 부호값에 의해 결정된다.

최적 평면을 구하는 과정은 식 (2)와 같은 조건을 만족하는 함수의 해를 구하는 문제로 볼 수 있다.

$$\min_{w,b} \langle w \cdot w \rangle, \ (\ s.t. \ y_i(wX_i + b) \geq 1 \ , \quad i = 1, \ldots, n \ ) \tag{2}$$

SVM은 빠른 학습 시간 때문에 동적인 예측 모형에서 사용되지만 예측의 정확성에 있어 기존의 웹 예측 시스템에서 주로 사용하고 있는 피어슨의 상관 계수 알고리즘에 비해 향상된 성능을 보이지는 못한다.

본 논문에서는 이러한 문제의 해결 방안으로 7개의 커널 함수들을 사용하여 주어진 학습 데이터를 가장 잘 모형화하는 한 개의 커널함수를 찾아내는 앙상블 전략을 사용한다.

3. 앙상블 Support Vector Machine과 하이브리드 SOM 알고리즘

3.1 앙상블 Support Vector Machine

기존의 SVM 모형은 웹 데이터 분석을 위해 한 개의 커널 함수만을 사용하지만, 앙상블 SVM 모형에서는 다수의 커널 함수(*spline*($\cdot$), *sign*($\cdot$), *smother*($\cdot$) 등)를 사용하여 가장 좋은 성능의 함수를 결정하여 최적의 모형을 구축하게 된다.

본 논문에서는 서로 다른 커널함수를 갖는 SVM 학습모형의 군집들 중에서 학습 과정을 통해 가장 작은 평균제곱오차(MSE: Mean Square Error)를 갖는 모형을 결정한다. 이 방법에 의해 기존의 SVM보다 더 정확한 예측력을 갖는 모형을 얻게 된다.

앙상블 모형을 수행하기 위해 소요되는 시간의 문제점을 해결하기 위해 본 논문에서는 붓스트래핑(Bootstrapping)기법을 사용한다. 이 방법은 모형의 학습을 위해 전체 데이터를 사용하지 않고 반복을 허용하는 임의 추출 방법인 재표본 기법을 사용한다. 이 경우, 데이터의 크기가 N인 전체 데이터의 각 개체는 모두 같은 확률 $1/N$로 추출되어 M개 크기의 표본이 이루어지고, 이 표본을 통해 앙상블 학습이 이루어진다. 이때, 전체 데이터 크기인 N에 비해서 앙상블 학습에 사용되는 표본의 크기인 M은 매우 작도록 결정된다[6, 7].

앙상블 SVM의 알고리즘은 알고리즘 1과 알고리즘 2의 두 단계로 구성된다. 알고리즘 1은 앙상블 SVM의 초기화 단계로서, 커널 함수를 포함한 앙상블 SVM 모형에서 필요로 하는 모든 모수를 결정하는 과정이다.

알고리즘 1에 의해 필요한 모수가 결정된 후에는, 알고리즘 2와 같이 최적의 커널 함수를 결정하는 앙상블 단계가 필요하며, 이 과정을 통하여 SVM의 정확도에 대한 성능 향상이 이루어진다.

알고리즘 1. 앙상블 SVM의 초기화 단계
Algorithm 1. Initialization Phase of Ensemble Support Vector

```
Algorithm: Initialize_앙상블 SVM (Parameter[j])
          // 앙상블 SVM의 초기화
Set parameters; nsv, beta, bias, X, Y, ker, C, loss, e
//X-training inputs, Y-training targets, ker-kernel function
//C-upper bound, loss-loss function, e-insensitivity, bias-bias term
//nsv-number of support vectors, beta-difference of Lagrange multipliers
// 매개변수의 초기값 조사
 if ( nargin < 3 | nargin > 6)
 // 매개 변수의 정확한 개수 조사
 else n = size (X ,1)  // 입력 데이터 확인
 if ( nargin <5) loss = eInsensitive
  //초기 손실함수를 ε-Insensitive로 결정
 if ( nargin <4) C= Inf
  // 학습 모형의 모수 추정의 상한 결정
 if ( nargin <3) ker = linear
  // 초기 커널함수를 linear로 결정
 end
// 커널 함수의 결정
Set H = zeros (n,n)
    // 커널함수 배열의 초기화
Repeat from i=1 to n
    repeat from j =1 to n
     H(i, j) = kernel (ker, X(i), X(j))
       // 커널함수의 적용
end
```

알고리즘 2. SVM의 성능 향상을 위한 앙상블 단계
Algorithm 2. Ensemble Phase for Performance improvement of SVM

```
Algorithm: Voting_Kernel(Sel_ker[k])
// 다수의 커널 함수를 사용하여 최적의 모수를 결정
Choose optimal kernel K* such that min||w||
// 1-scatter smoothing, 2-bin, 3-running mean, 4-kernel smoother
// 5-equivalent kernel, 6-regression spline, 7-cubic smoothing spline
// 붓스트랩 샘플링
Repeat from i=1 to 13109
        if random_number < 0.1
                re_sampling
      // Bootstrapping의 재표본 기법 적용
end
// 최적 커널 함수의 결정
Repeat from k=1 to 7
        MSE[k]=risk(Sel_ker[k])
       // 위험함수 값의 최소 제곱 오차 계산
        if mse[k]=min;
        voting
  // 최소 제곱 오차 값이 가장 작은 커널함수를 선택
end
```

3.2 하이브리드 SOM

학습 데이터의 군집화에 있어서 SOM은 신경망이 스스로 학습하여 최적의 군집을 형성하게 되며 이러한 경우 데이터의 특성을 잘못 이해하여 그릇된 결론에 도달할 위험이 있다. 또한, SOM에 의해서 최종적으로 구축된 모형은 한 개의 가중치 값으로 고정된 모수를 갖는 모형이 된다. 이 모형은 일반적으로 가장 좋은 모형이라는 보장은 없다. 최종 구축된 모형에 대한 최적값의 보장을 위해서 한 개의 모형으로 고정시키지 않고 여러 가능성을 포함하는 모수에 대한 분포의 개념이 도입되어야 한다.

하이브리드 SOM 알고리즘은 SOM의 단점인 모형에 대한 설명 부족을 베이지안 추론 기법으로 해결하였고, 사전에 군집수를 결정해야 하는 계층적 군집화의 문제점에 대해서는 사전에 군집의 개수를 정하지 않는 SOM의 유연성을 따른다. 그리고 SOM에 의해 최종적으로 구축된 모형에서 출력층의 각 노드가 한 개의 가중치 값들로 고정되는데 비해 하이브리드 SOM은 고정된 가중치 값이 아닌 가중치가 속하게 되는 분포가 결정된다. 하이브리드 SOM은 가중치가 속하는 확률분포(probability distribution)의 모수가 학습 데이터에 의해 갱신된다. 이러한 갱신은 동일한 입력값에 대해서 항상 동일한 결과값만을 계산하지 않는다. 따라서 국지적 최적값에 빠진 경우에도 얼마든지 전역 최적값으로 빠져 나올 수 있다. 모형의 가중치 갱신은 주어진 학습 데이터를 학습하기 이전의 모형에 대한 믿음인 사전 확률 분포가 학습 데이터에 의한 우도 함수와의 결합을 통하여 사후 확률 분포가 되고 이 분포는 한 개의 학습 데이터에 의해 학습이 끝난 새로운 모형이 된다. 이러한 확률적 분포 갱신 학습이 SOM의 신경망 내에서 이루어진다. 알고리즘 3은 하이브리드 SOM의 초기화 과정을 알고리즘 4는 실행과정을 나타낸다.

알고리즘 3. 하이브리드 SOM의 초기화 단계
Algorithm 3. Initialization Phase of HSOM

```
Algorithm: Initialize_HSOM (Input[i])
// HSOM의 초기화 : 네트워크 상수의 초기화
Initialize network parameter; Input_layer(int i, int o, int init_neigh_size)
num_inputs=i, num_outputs=o
  // 입력 벡터와 출력 벡터의 초기화
neighborhood_size=init_neigh_size
  // 형상 지도의 이웃반경의 초기화
weights = new float[num_inputs*num_outputs]
  // 가중치 지정
outputs = new float[num_outputs]
  // 출력값 지정
// HSOM의 초기화 : 확률 분포의 초기화
Initialize bayesian distributions // 베이지안 확률 분포 결정
// 초기 가중치의 분포를 평균과 분산이 각각 0과 1인 가우시안 분포로 결정
  Initialization of the weight vector, w_j(0) to have
        probabilistic distribution, N(0,1).
          // 학습율 함수의 결정
  Initialization of the learning rate α(0), α(t)∝t^-α ;
0<α<1.
          // 이웃 반경 함수의 결정
  Initialization of the neighborhood function K(j, j*),
        K decreases as to increase |j-j*|
```

알고리즘 4. 하이브리드 SOM
Algorithm 4. Hybrid SOM

Algorithm: Train_HSOM (Train[j])
// 하이브리드 SOM의 베이지안 확률 분포의 갱신
Determine the winner node
// 승리 노드의 결정
// 가우시안 분포를 따르는 입력 벡터의 정규화
Normalization of input vector, Gaussian distribution with mean 0, variance 1
// 가중치의 hyper-parameter 결정
Choose the distribution of weights, $w \sim f(\theta)$
// 최소 유클리디안 거리를 갖는 노드 : 승자 결정
Choose the winner node $j^* = \arg\max y_j$ using Euclidean criteria.
Update of parameters
$w_j^{New} = w_j^{Old} + \alpha(j)K(j,j^*)(X - w_j^{Old})$
,where, $K(j,j^*)$; Neighborhood function
end
// 가중치 분포의 모수 갱신 작업의 반복
Replace old distribution by current.
// 하이브리드 SOM의 승리 노드 결정
Set Winner_index=0,
maxval=-1000000
// 최대 반복 회수 지정
Find the winner neuron// 승리 노드 결정
Repeat from j=0 to num_outputs
repeat from i=0 to num_inputs
// 최소 유클리디안 거리를 갖는 노드가 승리 노드가 됨
$winner[i,j] = argmin\|x(k) - w_j\|$
end
Set m=(int)alpha, delta=alpha-m //가중치분포갱신
// 가우시안 분포로부터 모수 생성
while(count<data_size)
v=rand()/32767.0,
w=pow(y/m,delta)/(1+(y/m-1)*delta)
end
// 수렴될 때까지 반복 학습
Repeat Until given criteria satisfaction.

4. 전체 시스템 개요

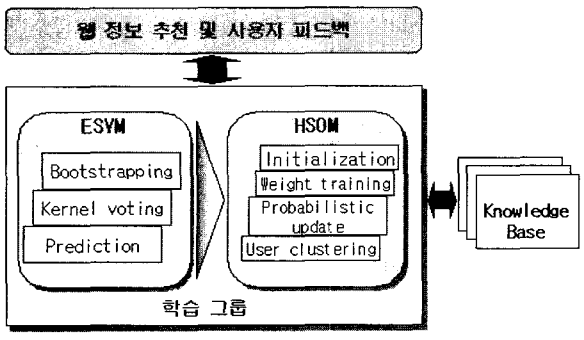

그림 1. 시스템의 전체 구성도
Fig 1. System Architecture

그림 1은 동적 웹 정보 추천을 위한 제안 시스템의 전체 개요도로서, 전체 시스템은 예측과 군집분석 도구를 갖는 학습 그룹과 웹 정보의 동적 추천을 위한 지식베이스, 그리고 웹 정보의 추천을 받아 이를 사용자에게 보여주는 사용자 그룹으로 구성된다.

학습 그룹은 다시 앙상블 SVM(ESVM)과 하이브리드 SOM(HSOM) 두 개의 학습 알고리즘 모듈을 포함하고 있다. 앙상블 SVM 모듈은 정제된 웹 로그의 클릭스트림 데이터를 이용하여 사용자의 선호도를 예측하는 통계적 학습 엔진이다. 이 부분을 통해 웹 로그 데이터가 갖는 문제점 중의 하나인 희소성 문제를 해결하며, 빠른 학습 시간과 실시간 예측 결과를 위해 요구되는 웹 정보의 동적 추천을 위해 사용된다.

하이브리드 SOM 모듈은 서로 유사한 행동패턴을 보이는 사용자에 대한 군집화를 위해 사용되는 부분이다. 이 두개의 모듈은 앙상블 SVM에 의한 결과를 대상으로 웹 사용자에 대한 군집화를 수행하는 연계성 구조를 갖는다.

지식베이스는 기존의 데이터와 이를 통한 앙상블 SVM과 하이브리드 SOM의 학습 결과에 대한 정보를 갖고 있으며, 빠른 정보 예측을 위하여 현재의 결과를 학습 그룹에 제공한다. 새로운 사용자에 대한 학습 결과에 의해 지식베이스의 내용은 지속적으로 갱신된다. 사용자 그룹은 제안 시스템의 적용 대상이 되며, 최종적으로 사용자 그룹의 피드백 정보를 통한 성능향상을 기대할 수 있다.

5. 실험 및 결과

본 논문의 실험 데이터로는 2000년도 KDD Cup 대회에서 사용되었던 웹 로그의 클릭스트림 데이터를 사용하여[8, 9], Visual C++ 6.0으로 구현하였다. 제안 알고리즘에 의한 시스템의 성능평가를 위해 피어슨의 상관계수 알고리즘을 기준으로 예측의 정확도와 학습시간의 비용에 대한 비교를 수행하였다. SVM의 문제점을 개선한 부분에 대한 성능평가는 피어슨의 상관계수 알고리즘과 제안하는 앙상블 SVM과의 비교를 수행하였다. 하이브리드 SOM에 대한 성능평가는 기존의 SOM의 비수렴성 문제를 해결하는 면에서 SOM과 하이브리드 SOM 모형 사이의 성능평가를 수행하였다.

5.1 모형의 정확도와 예측 시간 평가

이 절의 실험에서는 앙상블 SVM 모형에 의한 사용자의 웹 페이지 방문 시간에 대한 예측값을 기존 피어슨의 상관계수 알고리즘과 SVM과 비교하였다. 이때 성능 비교의 측도로는 표 1과 같은 평균제곱오차 기법을 이용하였다.

표 1의 결과, 평균제곱오차는 앙상블 SVM에 의해 예측된 값이 전체 모형 타당성 평가 결과 0.89로, 상위 50%의 데이터에 대한 결과는 0.64가 나왔다. 이는 본 모델이 실제값을 정확히 예측하고 있음을 나타낸다. 여기서, 정확성 성능평가에 대한 값은 이와 같이 전체 데이터에 대해 수행한 값과 상위 50%에 대한 값으로 나누어 계산하였다.

피어슨의 상관 계수 알고리즘과 기존의 SVM과의 비교 평가결과를 표 1에 나타내었다.

표 1. 앙상블 SVM을 이용한 웹 예측 성능 평가
Table 1. Performance Evaluation of Web Information Prediction by Using Ensemble SVM

평균제곱오차	피어슨 방법	SVM	앙상블 SVM
전체	1.37	1.29	0.89
상위 50%	1.01	0.97	0.64

표 1에서 보듯이 전반적인 정확도에 있어 피어슨의 상관계수 알고리즘과 SVM은 큰 차이를 보이지 않고 서로 비슷한 결과를 보인다. 하지만 앙상블 SVM에 의한 모형의 정확도는 앞의 두 알고리즘에 비해 훨씬 정확함을 알 수 있다. 이렇게 기존의 두 개의 모형에 대한 결과보다 제안하는 앙상블 SVM의 결과값이 정확하게 나올 수 있는 이유는, 데이터를 학습하는 동안 최적 커널함수를 선택하는 앙상블 구조를 가지고 있기 때문이다.

모형의 정확도와 함께 동적 웹 정보 추천 시스템은 매우 빠른 학습시간을 요구한다. 따라서 이들 3개의 학습 기법들에 대한 학습시간을 데이터 량에 따라 평가하여 표 2에 나타내었다.

표 2. 데이터 량에 따른 학습 시간(단위:초)
Table 2. Learning Time according to Data Amount

데이터 크기	피어슨	SVM	앙상블 SVM
500	21,873	2,941	2,955
1,000	48,121	8,890	8,954
1,500	72,181	20,631	22,009
2,000	96,241	35,170	36,542
2,500	120,302	54,391	57,325
3,000	144,362	77,967	81,024

표 2의 결과, SVM 모형은 데이터의 크기가 증가함에도 불구하고 빠른 학습시간을 보이고 있으며, 특히 데이터의 수가 증가하면 할수록 피어슨의 상관계수 알고리즘에 비해 훨씬 빠른 학습시간을 나타내고 있다. 따라서, 앙상블 SVM이 SVM에 비해 학습시간의 면에서도 떨어지지 않는 성능을 보이고 있음을 알 수 있다. 3,000개의 데이터 크기의 경우 SVM에 비해 피어슨의 상관계수 알고리즘은 두 배 이상의 느린 결과를 나타내지만 앙상블 SVM의 경우에는 기존의 SVM과 별 차이를 보이지 않고 있음을 알 수 있다.

5.2 하이브리드 SOM의 성능 평가

이번 절에서는 KDD Cup 2000의 정제된 로그 데이터에 대한 희소성을 앙상블 SVM에 의해 제거된 데이터를 이용하여 군집의 결과에 대한 수렴성을 평가한다. 여기서, 데이터에 대한 모형화를 수행하는 이유는 주어진 데이터를 이용하여 전체 데이터에 대한 일반성을 찾기 위함이다. 만약, 주어진 데이터로부터 학습한 결과가 매번 다르게 나온다면 그 모형은 최적의 모형이라고 할 수 없다. 특히, SOM 알고리즘이 일반적으로 수렴을 하지 못하는 것에 비해 하이브리드 SOM에 의한 군집화 모형의 수렴성을 극복하고자 한다.

표 3. SOM과 하이브리드 SOM의 수렴성 비교
Table 3. Compare of Convergency between SOM and Hybrid SOM

군집수	SOM	하이브리드 SOM
3	11	7
4	29	21
5	28	43
6	19	18
7	8	9
8	5	2

표 3은 자기 조직화 형상지도의 차원을 3×3으로 결정한 상태에서의 SOM과 하이브리드 SOM의 군집결과에 대한 비교표이다.

표 3은 총 100번의 군집화를 수행하는 동안 군집화의 결과로서 얻어진 최종 군집의 수를 나타낸다. 여기서, SOM의 군집결과를 살펴보면, 최종 군집 수가 4, 5, 6인 경우가 76%를 차지한다. 따라서, SOM의 결과만으로 최종 군집화의 정확한 군집수를 결정하기는 어렵다. 그러나 하이브리드 SOM에서는 군집수 5인 경우가 43번이나 발생하였다. 이는 기존의 SOM에 비해 본 논문에서 제안하는 군집화 기법인 하이브리드 SOM이 군집 수 5에 절대적으로 수렴하고 있음을 나타낸다. 따라서, SOM의 비수렴성 문제를 하이브리드 SOM은 어느 정도 잘 해결하고 있음을 알 수 있다.

5.3 군집분석에 의한 동적 추천

표 3의 결과 하이브리드 SOM의 최적의 군집수는 5에 수렴함을 알 수 있다. 따라서, 5개의 군집을 목표로 최종 하이브리드 SOM을 이용한 군집 수행 결과는 표 4와 같다.

표 4. 하이브리드 SOM 최종 군집 결과
Table 4. Final Result of Hybrid SOM

군집	개체 수	1순위	2순위	3순위	4순위	5순위
그룹1	2,423	page 37 (25.1초)	page 46 (21.3초)	page 8 (18.8초)	page 52 (18.3초)	page 42 (18.1초)
그룹2	1,989	page 201 (30.1초)	page 123 (26.9초)	page 169 (24.1초)	page 137 (14.1초)	page 144 (9.9초)
그룹3	1,903	page 83 (19.6초)	page 102 (13.8초)	page 116 (13.1초)	page 78 (10.2초)	page 126 (7.1초)
그룹4	1,471	page 219 (24.3초)	page 194 (21.9초)	page 83 (17.9초)	page 261 (16.4초)	page 137 (14.1초)
그룹5	866	page 19 (28.3초)	page 63 (27.8초)	page 23 (23.6초)	page 42 (21.4초)	page 62 (20.6초)

표 4는 8,652명의 Id에 대한 사용자의 군집결과를 나타낸다. 이는 각 군집에 최종적으로 할당된 사용자 개체 상위 1순위부터 5순위까지의 웹 페이지, 해당 페이지에 대한 접속가능 시간의 평균값을 나타낸다. 그 예로서, 그룹 1에 속한 사용자는 2,423명이고, 이 집단에서 가장 우선순위가 높은 웹 페이지는 페이지 37, 이 집단의 사용자들이 페이지 37에 머문 평균 시간은 25.1초임을 나타낸다.

표 4를 이용하여 8,653번째의 새로운 사용자(Id: 8653)가 접속을 시도하게 되면 각 웹 페이지에 대한 접속시간을 이용하여 유클리디안 거리를 이용한 유사도 계산을 수행한다.

표 5. 새로운 사용자(Id: 8653)와 군집간의 유사도 계산 결과
Table 5. Result of Similarity Calculation between New user and the Cluster

군 집	그룹 1	그룹 2	그룹 3	그룹 4	그룹 5
거 리	2.49	3.76	1.42	4.33	2.85

표 5의 결과, Id: 8653인 사용자는 가장 작은 유클리디안 거리가 계산된 군집 3에 할당된다. 따라서, 이 사용자에 대한

최우선 웹 페이지의 추천은 표 5의 결과를 이용하여 군집 3의 가장 상위의 대표 선호도를 갖는 웹 페이지 83을 추천하게 된다. 그리고 이때, 이 사용자가 해당 페이지에 머물 시간은 19.6초로 예측하게 된다. 실제로 Id, 8653인 이 사용자는 페이지 83을 21.3초 동안 접속하였다. 따라서, 동적 웹 정보 예측 시스템으로 추천한 값과 실제값의 차이는 1.7초가 된다.

표 6은 제안 웹 선호도 추천 방법과 피어슨 기반의 추천 시스템에 대한 평균제곱오차의 계산 결과를 나타낸다.

표 6. 동적 추천의 최종 성능 평가 결과
Table 6. Final Result of the Recommendation System

	피어슨 기반의 추천 시스템	제안 알고리즘
평균제곱 오차	1.63	0.94

표 6의 결과 제안하는 시스템의 평균제곱 오차값이 기존 피어슨 기반의 동적 추천 시스템의 결과값 보다 훨씬 작음을 알 수 있다. 이는 본 논문에서 제안하는 동적 웹 정보 추천 시스템이 기존의 추천 시스템에 비해 더 정확한 예측 결과를 보임을 알 수 있다.

6. 결론

본 논문에서는 앙상블 SVM기법에 의해 KDD Cup 실험 데이터를 이용하여 총 269개의 페이지에 대한 예측 모형화를 통해 사용자별로 접속하지 않은 웹 페이지에 대한 접속 가능 시간을 예측하여 접속시간이 가장 길 것으로 예상되는 사용자의 선호도 웹 정보를 예측하여 사용자에게 동적 추천하는 웹 페이지에 대한 사용자 모델링 방법을 제안하였다. 또한, 사용자가 웹 사이트의 전체 페이지에 비해 매우 작은 페이지만을 방문하면서 발생하는 웹 로그 데이터의 희소성 문제도 해결하였다. 이 과정에서 앙상블 SVM에 의해 완전한 구조를 갖는 데이터들은 하이브리드 SOM 알고리즘에 의해 기존의 SOM에서 발생하는 군집결과에 대한 비수렴성 문제를 해결하여 정확히 한 개의 모델로 수렴하도록 설계되었다.

향후 연구 과제로는 다양한 데이터 마이닝 알고리즘을 개발하여 로그 데이터의 희소성 문제를 분석하여 정확한 예측을 위해 반영될 수 있도록 설계할 예정이다.

참 고 문 헌

[1] Teuvo Kohonen, Self Organizing Maps, Springer, 1997.

[2] Bishop, C. M., Svensen, M., Williams, C. K. I., "GTM: A Principled Alternative to the Self Organizing Map," Proceedings 1996 International Conference on Artificial Neural Networks, ICANN 96, Vol. II, pp. 165-170, Bochum, Germany, 1996.

[3] Bishop, C. M., Svensen, M., Williams, C. K. I., "GTM : a Principled Alternative to the Self Organizing Map," Proceedings 1996 International Conference on Artificial Neural Networks, ICANN 96, Volume1112. pp. 165-170, Bochum, Germany, 1996.

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

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

[6] Gelfand, A. E., Smith, A. F. M., "Sampling Based Approaches to Calculating Marginal Densities," Journal of the American Statistical Association, 85, 398-409, 1990.

[7] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B., "Bayesian Data Analysis," Chapman & Hall, 1995.

[8] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web," *In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence(ICTAI'97)*, pp. 1-3, November 1997.

[9]"http://www.ecn.purdue.edu/KDDCUP/", 2002.

저 자 소 개

윤경배(Kyung-bae Yoon)

1986년 : 인하대학교 수학과 (이학사)
1998년 : 서강대학원 정보기술경제학 (경제학석사)
2003년 : 인하대학원 전자계산공학과 (공학박사)
1998년~현재 김포대학 컴퓨터계열 조교수

관심분야 : 데이터마이닝, 지문 및 음성 인식, 인공지능 등
Phone : 016-314-9280
E-mail : kbyoon@kimpo.ac.kr

최준혁(Jun-Hyeog, Choi)

1990년 : 경기대학교 전자계산학과 졸업(이학사)
1995년 : 인하대학교 대학원 전자계산공학과 졸업(공학석사)
2000년 : 인하대학교 대학원 전자계산공학과 졸업(공학박사)
1997년 - 현재 김포대학 컴퓨터계열 조교수

관심분야 : 데이터마이닝, 신경망, 바이오 정보시스템 등
Phone : 016-690-5451
E-mail : jhchoi@kimpo.ac.kr