The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 21, No. 1, pp.87-92, Feb. 28, 2021. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2021.21.1.87
JIIBC 2021-1-12

빅데이터 기반 추천시스템을 위한 협업필터링의 최적화 규제

Regularized Optimization of Collaborative Filtering for Recommander System based on Big Data

박인규*, 최규석**

In-Kyu Park*, Gyoo-Seok Choi**

요 약 빅데이터 기반의 추천시스템 모델링에서 바이어스, 분산, 오류 및 학습은 성능에 중요한 요소이다. 이러한 시스템에서는 추천 모델이 설명도를 유지하면서 복잡도를 줄여야 한다. 또한 데이터의 희소성과 시스템의 예측은 서로 반비례의 속성을 가지기 마련이다. 따라서 희소성의 데이터를 인수분해 방법을 활용하여 상품간의 유사성을 학습을 통한 상품 추천모델이 제안되어 왔다. 본 논문에서는 이 모델의 손실함수에 대한 최적화 방안으로 max-norm 규제를 적용하여 모델의 일반화 능력을 향상시키고자 한다. 해결방안은 기울기를 투영하는 확률적 투영 기울기 강하법을 적용하는 것이다. 많은 실험을 통하여 데이터가 희박해질수록 기존의 방법에 비해 제안된 규제 방법이 상대적으로 효과가 있다는 것을 확인하였다.

Abstract Bias, variance, error and learning are important factors for performance in modeling a big data based recommendation system. The recommendation model in this system must reduce complexity while maintaining the explanatory diagram. In addition, the sparsity of the dataset and the prediction of the system are more likely to be inversely proportional to each other. Therefore, a product recommendation model has been proposed through learning the similarity between products by using a factorization method of the sparsity of the dataset. In this paper, the generalization ability of the model is improved by applying the max-norm regularization as an optimization method for the loss function of this model. The solution is to apply a stochastic projection gradient descent method that projects a gradient. The sparser data became, it was confirmed that the propsed regularization method was relatively effective compared to the existing method through lots of experiment.

Key Words : factored recommender systems, stochastic gradient descent, item similarity, projected gradient

Ⅰ. 서 론

전자 상거래에서 사용자가 자신의 개인 취향에 가장 잘 맞는 상품을 선택하는 데 추천 시스템을 활용하고 있다[1][2]. 이러한 시스템은 구매(purchase), 선호도(rating) 또는 리뷰(review)와 같은 사용자가 만들어 낸 과거의

*정회원, 중부대학교 소프트웨어공학부
**종신회원, 청운대학교 컴퓨터공학과 (교신저자)
접수일자 2021년 1월 20일, 수정완료 2021년 2월 3일
게재확정일자 2021년 2월 5일

Received: 20 January, 2020 / Revised: 3 February, 2020 / Accepted: 5 February, 2021
*Corresponding Author: lionel@chungwoon.ac.kr
Dept. of Computer Science, Chungwoon University, Korea

자료(feedback)를 토대로 사용자의 선호도를 예측한다. 일반적으로 이와 같은 사용자의 피드백 정보를 이용하는 시스템은 크게 협업 필터링 기반 방법(collaborative filtering based model)과 콘텐츠 기반 방법(content based model)으로 분류 할 수 있다.

사용자의 상품에 대한 선호도정보를 활용하여 하나의 그룹내의 상품들은 유사성이 동일하기 때문에 새로운 다른 상품을 추천하는 근접이웃 모델(neighborhood based model)이 있다. 또한 기계 학습 알고리즘을 사용하여 유사성 또는 잠재 요인의 모델링을 통하여 추천 작업을 수행한다. 또한 SLIM(sparse linear method)이라는 추천 시스템은 사용자가 상품에 대해 선호도한 매우 적은 데이터(sparse data)에 대한 회귀 모델링을 기반으로 상품간의 유사한 유사도를 측정하여 사용자의 선호도에 맞는 상품을 추천하는 방식으로 우수한 성능을 보이고 있다. 그러나 이 방법은 사용자가 구매하거나 선호한 상품 간의 관계만 모델링 할 수 있는 방식이기에 상품간의 전이 관계(transitive relation)는 추천할 수 없다. 이러한 단점을 해결하기 위하여 인수분해 행렬(factorization matrix)을 구성하여 상품간의 유사성 행렬을 학습하여 사용자의 선호도를 추천하는 상품분해 유사도(FISM: factored item based similarity) 모델이 제안되었다[3][4][5].

이러한 일련의 기계학습의 목적은 모델의 복잡성을 균형있게 조정하여 데이터에 대한 최적의 모델을 구축하는 것이 관건이다. 보통 편차(bias)가 낮은 모델은 학습오류가 낮고 과적합(overfitting)일 가능성이 높고, 역으로 분산(variance)이 낮은 모델은 학습오류는 다소 높더라도 테스트 오류가 낮고 약간의 과적합인 경우이다. 과소적합(underfitting)인 경우에는 학습오류와 테스트오류가 비슷하겠지만 위의 두 경우가 일정 수준 이상으로 높을 가능성이 많고 특별한 경우가 아니면 평균과 분산 모두 높을 가능성이 있다. 따라서 편차와 분산, 학습과 테스트 오류, 학습의 정도는 너무 복잡하지 않으면서 낮은 평균과 분산, 적은 학습 및 테스트 오류를 가지는 모델을 지향하기 위함이다. 결론적으로 비슷한 성능을 보이는 모델이라면 구조가 간단하여 일반적으로 일반화 오류(generalization error)가 더 적을 가능성이 크다.

본 논문에서는 모델의 복잡성을 줄이기 위하여 낮은 순위 분해(low rank decomposition)를 유도하기 위하여 max-norm규제를 상품기반 유사도 추천 시스템에 적용하여 기존의 모델과의 비교우위를 논하고자 한다.

II. 관련연구

추천시스템의 모델에서 상품에 대한 사용자의 선호도(preference)의 선호도는 식(1)과 같이 정의된다.

$$\tilde{r}_u = r_u S \tag{1}$$

여기서 r_u 는 모든 상품에 대한 사용자 u ($|C| = n$)의 선호도이고 S는 상품 i ($|D| = m$)간의 유사도를 나타내는 $m \times m$ 의 희소행렬이다. 따라서 상품 기반의 근접이웃 추천 모델과 유사하고 외생 변수가 없고 회귀 선형(sparse linear) 모델에서는 식(2)에 의하여 데이터에서 직접적으로 상품의 선호도를 추정한다[6][7].

$$\begin{aligned} minimize_S \ & \frac{1}{2}\|R - RS\|_F^2 + \frac{\beta}{2}\|S\|_F^2 + \lambda\|S\|_1 \\ & subject\ to\ S \geq 0,\ diag(S) = 0 \end{aligned} \tag{2}$$

$\|\cdot\|_F^2$는 *Frobenius norm*이고 ℓ_1-*norm* 규제, RS는 추천점수($\tilde{R}$)의 평균제곱 오차함수(root mean square error)를 사용한다. 사용자가 해당상품에 대한 선호도 r_{ui} 를 계산할 경우에는 해당상품의 r_{ui} 는 계산에 사용되지 않도록 $diag(S) = 0$을 적용된다.

상기의 사용자의 선호도는 유동적인데 반하여 상품간의 선호도는 유동성이 적기 때문에 정규화 특이값 분해(normalized singular value decomposition)방법에서는 선호도 예측을 위하여 상품간의 인수분해를 활용하여 유사도를 측정한다. 상품간의 유사성은 두 상품의 행렬 P와 Q의 곱으로 학습된다. 두 상품 i 와 j 의 유사도 $sim(i, j)$은 P와 Q간의 내적 $p_i \cdot q_j^T$로 상품 i 에 대해 사용자 u 에 대한 유사도는 식(3)과 같다.

$$\tilde{r}_{ui} = \hat{r}_{ui} = b_u + b_i + \sum_{j \in R_u^+} p_j q_i^T \tag{3}$$

여기서 b_u 및 b_i는 사용자 및 상품의 편차이고 R_u^+는 사용자가 선호도한 상품의 집합이다. 이 모델의 최적화를 위한 손실함수는 식(4)와 같이 정의된다[8][9][10].

$$\begin{aligned} minimize_{p,q} \ & \frac{1}{2}\sum_{u \in C}\sum_{i \in R_u^+} \|r_{ui} - \hat{r}_{ui}\|_F^2 \\ & + \frac{\beta}{2}(\|P\|_F^2 + \|Q\|_F^2) \end{aligned} \tag{4}$$

Ⅲ. 협업필터링의 규제기반 최적화

상품기반의 선호도는 유동정이 적지만 확장성이 취약하다. 유저와 아이템이 증가하면서 복잡도는 $O(mn)$으로 커지게 되고 상품의 피드백 정보의 희소성(sparsity)도 큰 문제가 된다. 예를 들어, 매트릭스와 타이타닉을 전부 평가한 사용자는 한명이고, 두 영화에 대하여 하나씩 평가한 사용자가 각기 2, 3명이라고 가정해보자. 두 영화에 대한 사용자들의 유사도가 동일하게 계산되었을 경우에는 극단적으로 수십만의 사용자가 있지만 두 사용자만 매우 다른 영화에 비슷한 평가를 하여 매우 다른 영화임에도 불구하고 유사도가 비슷한 경우가 발생할 수 있다. 이에 대한 해결방안으로 latent factor model 을 사용해 사용자와 상품간의 유사도를 찾는 방법이다. 결국 추천 문제는 최적화 문제가 되므로 한 사용자에게 어떤 상품에 대한 평가의 예측을 측정하기 위하여 식(5)와 같이 정의된 제곱오차 손실함수(root mean square error)를 사용하며 최소의 RMSE를 얻기 위해서 특이값 분해행렬을 사용한다.

$$\pounds(\cdot)=\sum_{i\in D}\sum_{u\in C}\|r_{ui}-\hat{r}_{ui}\|^2 \tag{5}$$

여기서 r_{ui}는 실측값이고 $\hat{r}_{ui}$는 사용자 u 가 상품 i 에 대한 선호도로 식(6)과 같이 계산된다.

$$\tilde{r}_{ui}=\hat{r}_{ui}=b_u+b_i+(n_u^+-1)^{-\alpha}\sum_{j\in R_u^+\setminus\{i\}}p_jq_i^T \tag{6}$$

여기서 R_u^+는 값이 추정되는 현재 상품 i 를 제외한 사용자 u 가 선호도한 상품의 집합이다. 이 제외가 다른 유사도 모델인 NSVD 및 SVD++와 다르다. 결국 모델의 복잡성을 줄이기 위하여 *max-norm* 규제를 추가하면 행렬 P와 Q는 식(7)에 의하여 학습된다.

$$\begin{aligned}minimize\ _{p,q}\ &\frac{1}{|S|}\sum_{(u,i)\in S}(R_{ui}-P_u^{'}Q_i)^2\\&+\mu\,max\{(\|P\|_{2,\infty}^2,\|Q\|_{2,\infty}^2\}\end{aligned} \tag{7}$$

여기서μ 는 양의 패널티이고 $max\{(\|P\|_{2,\infty}^2+\|Q\|_{2,\infty}^2\}$는 행렬 인수분해와 양의 반부호 행렬(positive semi definite) 형태를 고려하면 $max_j(\sum_k P_{jk}^{2\ 1/2},\sum_k Q_{jk}^{2\ 1/2})$으로 간략화 된 *max-norm* 규제가 될 수 있다. 또한 차원의 크기가 $m*n$ 에서 $r(m+n)$으로 축소된다. S는 평가된 사용자와 상품의 집합을 나타내고 R_{ij}는 사용자의 피드백 정보를 나타내고 P와 Q는 학습행렬이다. 결국 규제를 가지는 손실 함수는 R의 사용자의 피드백정보의 유무와 관계없이 모든 상품에 대하여 적용된다.

따라서 상품간의 유사도 행렬 S의 방대한 크기에 대하여 RMSE의 최소화에 효율적인 특이값 분해 방법을 이용하여 확률적인 기울기강하 알고리즘이 적용된다. 이는 기울기 강하 알고리즘을 확률적으로 변형하여 손실함수를 최소화하는 기울기강하 탐색법은 매개변수 P, Q가 안정된 상태에 도달할 때 까지 식(8)에 의해서 갱신되어진다. 그림 1에 전체적인 알고리즘이 나타나 있다.

$$\begin{aligned}P^{new}&:=P^{old}-\tau\nabla f(P^{old}Q^{old})Q^{old}\\Q^{new}&:=Q^{old}-\tau\nabla f(P^{old}Q^{old})'P^{old}\end{aligned} \tag{8}$$

$$\mu\,max\{(\|P\|_{2,\infty}^2,\|Q\|_{2,\infty}^2\}\le B \tag{9}$$

FISMmax-norm
parameters:*gamma*(η)=0.1,*gamma_factor*=0.95,ρ=*dev.*

iter : 0
Init P and Q with random values in (-0.0001, 0.0001)
$RMSE=\frac{1}{|S|}\sum_{(u,i)\in S}(R_{ui}-P_u^{'}Q_i)^2$
while $matrix(P,Q)\in argmin_{p,q}RMSE+\mu\,max\{(\|P\|_{2,\infty}^2,\|Q\|_{2,\infty}^2\}$
  R': R ∪ *SampleZeros*(R, ρ)
  R' : *RandomShuffle*(R')
  for all $r_{ui}\in R'$ do
    $x=(n_u^+-1)^{-\alpha}\sum_{j\in R_u^+\setminus\{i\}}p_j$
    $\tilde{r}_{ui}=q_i^TX$
    $e_{ui}=r_{ui}-\tilde{r}_{ui}$
    $q_i^{pre}=q_i-\eta e_{ui}x$
    for all $j\in R_u^+\setminus\{i\}$ do
      $p_j^{pre}=p_j-\eta e_{ui}(n_u^+-1)^{-\alpha}\cdot q_i$
    end for
    $q_i=Project(q_i^{pre})$
    $p_i=Project(p_j^{pre})$
  end for
  iter =iter + 1
  gamma *= *gamma_factor*
 end while
 return P, Q
end procedure

Project(v)
  v_norm = ||v||
  if $\|v\|^2\ge B$
    return $\sqrt{B}*v$ / *v_norm*
  else
    v

그림 1. FISMmax-norm 알고리즘
Fig. 1. FISMmax-norm Algorithm

여기에서 τ 은 탐색과정에서 국부해(local minima)와 종속적이다. $\tau\nabla f(P^{old}Q^{old})$은 최소화하고자 하는 목적함수의 $P^{old}Q^{old}$ 에서의 기울기로 주어지며 $\tau>0$는 각

갱신의 크기를 제어한다. 기울기 강하기법은 임의의 최적화 탐색지점에서의 음의 기울기 값이 해당 지점에서 가장 빠르게 함수값을 감소시키는 방향인 것에 착안하여, 그 방향으로 step size γ만큼 조금씩 움직여가며 손실함수값을 최소화한다. 이 과정에서 각 반복에서 무작위로 하나의 훈련 쌍 (i, j)을 선택하고, P와 Q의 기울기와 반대 방향으로 탐색하는 과정에서 식(9)의 max-norm규제를 통하여 $max\{(\|P\|_{2,\infty}^2+\|Q\|_{2,\infty}^2\}$ 값이 임의의 B보다 크면 $max\{(\|P\|_{2,\infty}^2+\|Q\|_{2,\infty}^2\}$값을 $\sqrt{B}$가 되도록 변경하는 투영 기울기(projeced gradient)방법을 적용하여 손실함수의 오차의 최소화를 보장하는 P와 Q가 학습되어진다.

Ⅳ. 실험 및 결과

ML100K(movie lens data)와 Netflix의 희소 데이터 세트에서 추천시스템의 성능을 평가하기 위하여 선호도자료(validation data)와 모델 성능 비교에 사용할 시험자료(test data)를 구성하기 위하여 약간의 튜닝과정을 거쳤다. 표 1에서 '-1'의 데이터 세트는 동일한 사용자 및 상품 집합을 유지하고 '-2'의 데이터 세트는 첫 번째 데이터 세트의 사용자 상품 매트릭스에서 상품을 임의로 제거하여 생성된다. '-3'의 데이터 세트는 두 번째 희소 버전은 두 번째 데이터 세트의 사용자 상품 행렬에서 상품을 임의로 제거하여 유사하게 생성된다. 모든 데이터 세트에는 평가 값이 있으며 상품을 1로 설정하여 암시적 피드백으로 변환하였다[11][12][13].

표 1 : Datasets.
Table 1: Datasets

Dataset	#Users	#Items	#Ratings	Rsize	Csize	Density
ML100K-1	943	1,178	59,763	63.99	50.73	5.43%
ML100K-2	943	1,178	39,763	42.57	33.75	3.61%
ML100K-3	943	1,178	19,763	21.16	16.78	1.80%
Netflix-1	6,079	5,641	429,339	70.63	76.11	1.25%
Netflix-2	6,079	5,641	221,304	36.40	39.23	0.65%
Netflix-3	6,079	5,641	110,000	18.10	19.50	0.32%

#Users, #Items및 #Ratings은 각 데이터 세트의 사용자 수, 상품 및 선호도이고, Rsize및 Csize열은 각 데이터 세트에서 각 사용자 및 각 상품에 대한 평균 선호도 점수이다. 추천 품질은 HR(hit ratio) 및 ARHR(average reciprocal hit ratio) 와 nDCG(number discount cumulative gain)등 여러 가지가 있는데 여기서 사용한 척도는 식(10)과 같이 HR을 사용한다.

$$HR=\frac{\#hits}{\#users} \tag{10}$$

선호도는 앞에서 제시한 바와 같이 FISM 모델의 손실함수의 최적화를 위하여 FISM에서 사용된 규제에 의한 손실함수에 의한 최적화와 본 논문에서 제시한 max-norn 규제에 의한 손실함수가 추천 성능에 미치는 영향을 알아본다. 공간 부족으로 인해 ML100K-3 (ML100K로 표시) 및 Netflix-3 (Netflix로 표시) 데이터 세트에 대해서만 실험을 수행하였다. 상품 추천시에 상품의 자체 선호도는 제외하는 FISMrmse의 접근 방식과 효과를 조사하기 위해 선호도추천시에 max-norm규제를 가지는 FISM방식 즉, FISMmax-norm의 성능을 비교한다.

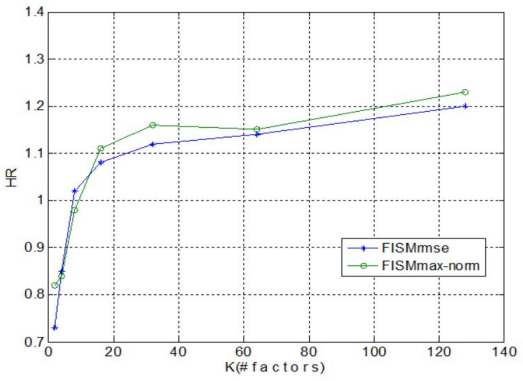

그림 2. ML100K 데이터에 대한 평가 결과
Fig. 2. Effect of estimation on ML100K

나머지 매개 변수의 값들을 일정하게 유지하면서 latent factor k의 수를 변경하면서 FISMrmse와 FISMmax-norm의 성능을 비교하였다. 그림 2, 3은 각각 ML100K MovieLens와 Netflix 데이터 세트에 대한 결과를 보여준다. k 값이 더 작은 경우 두 방식의 성능이 매우 유사하다는 것을 알 수 있다. 그러나 k 값이 증가하면 FISMmxa-norm이 FISMrmse보다 더 나은 성능을 나타내기 시작하고 k 값이 증가할수록 방법의 성능 차이가 커진다는 것을 알 수 있다. 이는 제안한 방식에서 사용하는 max-norm 규제 방식이 NSVD 및 SVD ++와 같은 접근 방식에서 사용되는 것보다 우수하고 factor 크기가 커질수록 성능이 비례한다는 것을 알 수 있다.

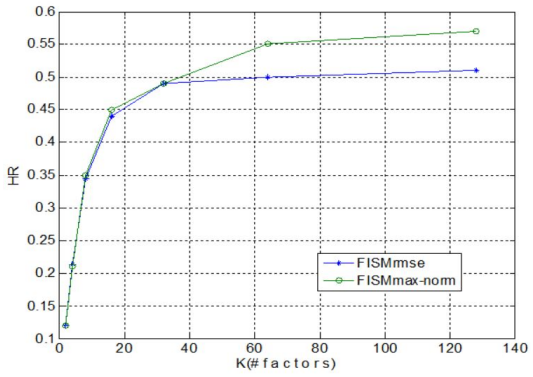

그림 3. Netflix 데이터에 대한 평가 결과
Fig. 3. Effect of estimation on Netflix

또한 Netflix 데이터의 경우 투영기울기 방법을 사용한 FISMmax-norm 경우에 $f(X)$는 0.7527, max-norm규제($\|X\|_{max}$)는 2.3446을 달성하여 k가 $5*10^{-5}$ 인 경우에 $f(X)+\mu\|X\|_{max}$는 0.7538의 손실오차가 발생하였다. FISM의 0.8235보다 약간의 손실은 보상되었지만 모델의 여러 가지 매개변수의 값들의 조정에 따라서 변화가 있음을 알 수 있었다.

V. 결 론

본 논문에서는 max-norm 규제를 적용하여 FISM모델의 최적화를 향상을 위하여 Top-N 추천 시스템을 위한 Factored Item Similarity 모델의 목적함수에 max-norm 정칙화를 적용하여 최적화를 유도하였다. 이는 Max-Norm 정칙화가 low rank decomposition를 유도한다는 접근법에서 착안한 것으로 데이터 행렬의 순위(rank)를 낮춰 모델의 복잡성을 줄여 목적함수의 손실을 최적화하도록 한 것이다. FISM의 factor기반 회귀 모델링에 있어서 실제 데이터에 max_norm 규제를 갖는 FISMrmse 알고리즘을 적용한 결과, SLIM이나 FISM과 같은 기존 방법보다 양호하였다. 결과적으로 제안한 목적함수 최적화 기법이 다른 방법보다 성능이 뛰어나고 데이터 세트가 희소해지면 성능 격차가 더욱 증가하는 것으로 나타났다. 본 논문에서 제시한 Max-Norm규제를 활용하면 사용자의 피드백 정보의 손실을 줄여 사용자의 선호도에 맞는 상품을 추천할 수 있을 것으로 기대한다.

References

[1] X. Yan, A. C. Reynolds. "Optimization Algorithms Based on Combining FD Approximations and Stochastic Gradients Compared With Methods Based Only on a Stochastic Gradient", Journal Article published 15 Oct in SPE Journal vol. 19 no. 05 pp. 873-890, 2014.
DOI: https://doi.org/10.2118/163613

[2] P. Cremonesi, Y. Koren, and R. Turrin. "Performance of recommender algorithms on top-n recommendation tasks". In Proceedings of the fourth ACM conference on Recommender systems, pp. 39-46, 2010.
DOI: https://doi.org/10.1145/1864708.1864721

[3] M. Deshpande and G. Karypis. "Item-based top-n recommendation algorithms". ACM Transactions on Information Systems (TOIS), vol. 22, no.1 pp. 143-177, 2004.
DOI: https://doi.org/10.1145/963770.963776

[4] R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis. "Large-scale Matrix factorization with distributed stochastic gradient descent". In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 69-77. ACM, 2011.
DOI: https://doi.org/10.1145/2020408.2020426

[5] S. Kabbur, Xia Ning, G. Karypis. "FISM: Factored Item Similarity Models for Top-N Recommender Systems", 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2013.
DOI: https://doi.org/10.1145/2487575.2487589

[6] R. J. Mooney and L. Roy. "Content-based book recommending using learning for text categorization". In Proceedings of the fifth ACM conference on Digital libraries, pp. 195-204. ACM, 2000.
DOI: https://doi.org/10.1145/336597.336662

[7] X. Ning and G. Karypis. "Slim: Sparse linear methods for top-n recommender systems. In Data Mining", Proceedings of the 2011 IEEE 11th International Conference on Data Mining, December, pp. 497-506, 2011.
DOI: https://doi.org/10.1109/icdm.2011.134

[8] Y. S. Im, E. Y. Kang, "MPEG-2 Video Watermarking in Quantized DCT Domain," The Journal of The Institute of Internet, Broadcasting and Communication(JIIBC), vol. 11, no. 1, pp. 81-86, 2011.
DOI: https://doi.org/10.1109/tip.2006.873476

[9] I. Jeon, S. Kang, H. Yang, "Development of Security Quality Evaluate Basis and Measurement of Intrusion Prevention System," Journal of the Korea Academia-Industrial cooperation Society (JKAIS), vol. 11, no. 1, pp. 81-86, 2010.
DOI: https://doi.org/10.5762/kais.2010.11.4.1449

[10] J. S. Oh, B, S. Lee, "A Study for Lifespan Prediction of

Expansion by Temperature Status", The Journal of KISTI, vol. 19, no. 10, pp. 424-429, 2018.
DOI: http://dx.doi.org/10.5762/KAIS.2018.19.10.424

저 자 소 개

박 인 규(정회원)

- 제10권 5호 참조
- 현 중부대학교 소프트웨어공학부 교수
- 주관심분야 : 데이터마이닝, 퍼지집합, 러프집합

최 규 석(종신회원)

- 제9권 6호 참조
- 1991 ~ 1996 : (주)SK텔레콤 중앙연구원 책임연구원
- 1997 ~ 현재 : 청운대학교 컴퓨터공학과 교수
- 주관심분야 : 인공지능, 이동통신, 빅데이터, ITS

※ 본 논문은 2020학년도 청운대학교 학술연구조성비에 의하여 지원되었음.