개인화된 추천시스템을 위한 사용자-상품 매트릭스 축약기법

김 경 재* · 안 현 철**

User-Item Matrix Reduction Technique for Personalized Recommender Systems

Kyoung-jae Kim* · Hyunchul Ahn**

Abstract

Collaborative filtering(CF) has been a very successful approach for building recommender system, but its widespread use has exposed to some well-known problems including sparsity and scalability problems. In order to mitigate these problems, we propose two novel models for improving the typical CF algorithm, whose names are ISCF(Item-Selected CF) and USCF(User-Selected CF). The modified models of the conventional CF method that condense the original dataset by reducing a dimension of items or users in the user-item matrix may improve the prediction accuracy as well as the efficiency of the conventional CF algorithm. As a tool to optimize the reduction of a user-item matrix, our study proposes genetic algorithms. We believe that our approach may relieve the sparsity and scalability problems. To validate the applicability of ISCF and USCF, we applied them to the MovieLens dataset. Experimental results showed that both the efficiency and the accuracy were enhanced in our proposed models.

Keywords : Collaborative Filtering, Reduction Technique, Product Recommender System, User-Item Matrix

논문접수일 : 2009년 02월 20일 논문게재확정일 : 2009년 03월 13일

※ 이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2006-332-B00100).

* 교신저자, 동국대학교 경영대학 경영정보학과 부교수, Tel : 02-2260-3324, e-mail : kjkim@dongguk.edu

** 국민대학교 경상대학 비즈니스IT학부 전임강사, e-mail : hcahn@kookmin.ac.kr

1. 서 론

인터넷의 출현과 확산은 마케팅 분야에서도 여러 가지 변화를 가져 왔다. 소비자들은 인터넷에 산적해 있는 많은 양의 자료와 정보를 활용하여 보다 저렴하고 좋은 품질의 상품을 구매할 가능성이 높아졌다. 그러나 최근에는 이러한 자료와 정보의 양이 너무 과다해져서 소비자들이 이를 활용하여 구매의사결정을 내리는 것이 어려워졌다. 즉, 제공되는 자료와 정보의 양이 너무 많아서 소비자들이 이를 적절하게 선택하여 활용하지 않으면 구매의사결정에 너무 많은 시간을 허비하게 되고, 한편으로는 선택적인 정보 활용에 따른 왜곡되거나 편향된 의사결정을 하게 될 가능성도 있다.

최근에는 이러한 인터넷 상의 자료와 정보 과다현상을 어떻게 해결할 것인지가 매우 중요한 연구주제로 대두되고 있다. 특히, 마케팅 분야에서는 소비자가 필요로 하는 자료와 정보를 선택적으로 제공함으로써 소비자의 구매의사결정을 지원하는 것이 무엇보다 중요한 과제라고 할 수 있다. 이러한 필요성에 의해서 상품추천시스템이 소비자의 구매의사결정을 지원하기 위한 인터넷 상의 의사결정지원시스템의 역할을 하고 있다.

상품추천시스템은 기본적으로 고객의 성향과 기호를 파악하여 고객이 원하는 상품에 대한 자료와 정보를 제공함으로써 고객의 구매의사결정을 지원하고자 하는 시스템이라고 할 수 있다. 따라서 고객의 성향과 기호를 잘 파악하고, 그 성향과 기호에 적합한 상품을 잘 선택하고, 이에 대한 자료와 정보를 잘 수집하여, 이를 고객에게 적절한 시간 내에 적절한 양으로 제공하는 것이 좋은 상품추천시스템의 필수요건이라고 할 수 있다. 즉, 정확한 고객성향파악을 통한 적절한 상품선택 추론기법의 개발 및 시스템의 효율성 개선 등이 상품추천시스템 분야의 연구에 있어 중요한 요소라고 할 수 있다.

추천시스템에서 고객성향파악에 효과적이라고 알려져 있으며, 가장 활발하게 이용되는 추천기법 중 하나가 협동필터링(collaborative filtering : CF)이다. 협동필터링은 소비자(사용자)의 성향과 기호를 그의 구매선호점수를 통하여 파악한 후 유사한 구매선호점수를 가진 다른 소비자가 구매한 상품이나 구매선호점수가 높은 다른 상품을 추천하는 기법이다. 즉, 사용자와 상품 간의 관계를 구매내역자료를 바탕으로 분석하여 추출한 후 활용하는 방법이다. 협동필터링은 여러 연구결과를 통해 그 유용성이 확인되었으며 실제 많은 추천시스템의 추천기법으로 활용되고 있다.

그러나 협동필터링은 구매내역자료를 분석하는 방법이기에 구매내역자료가 너무 많은 경우에는 추천의 정확도나 효율성이 떨어질 가능성이 있다. 협동필터링을 수행하는 데에는 많은 연산과정이 필요하기에 연산이 복잡한 경우에는 추천에 많이 시간이 필요하게 되고 추천의 정확도가 떨어질 수 있기 때문이다. 이러한 문제점들은 흔히 추천시스템 연구들에서는 희박성과 확장성의 문제로 다루어 왔다[김재경 등, 2002; Cho et al., 2002; Kim et al., 2002; 김재경 등, 2003; 김종우 등, 2004; 조윤호 등, 2004; Cho and Kim, 2004; 김재경 등, 2005; 김경재와 안현철, 2006].

본 연구에서는 추천시스템에 가장 많이 활용되고 있는 추천기법 중 하나인 협동필터링이 가질 수 있는 희박성과 확장성의 문제를 완화하면서도 추천의 성과를 유지 또는 더 높일 수 있는 보완된 협동필터링 기법을 제안하고자 한다.

본 연구에서 제안하는 보완된 협동필터링 기법은 연산을 수행하기 이전에 연산의 복잡성을 줄이면서도 추천의 정확성을 유지할 수 있도록

연산에 사용될 구매내역자료의 최적화를 수행하게 하는 방법이다. 구체적으로는 협동필터링에서 사용되는 사용자-상품 매트릭스의 사용자와 상품의 데이터를 축약하는 과정을 거치게 된다. 데이터 축약은 데이터마이닝에서 매우 중요한 과정이지만 축약에 따른 정보의 손실을 야기할 수 있다. 따라서 최소한의 정보손실을 통해 최대한의 효과를 확보할 수 있는 방안이 필요하며, 이를 위하여 본 연구에서는 최적화 알고리즘을 활용하여 데이터 축약을 수행한다. 본 연구에서는 전역 최적화 알고리즘인 유전자 알고리즘을 활용한다.

본 연구는 다음과 같이 구성된다. 다음 장에서는 전통적인 추천시스템에 대한 선행연구를 살펴 보고, 그 한계점을 확인한다. 제 3장에서는 본 연구에서 제안하는 추천모형을 설명하고, 제 4장과 제 5장에서는 본 연구에서 제안하는 추천모형을 검증하기 위한 실험설계와 그 결과를 각각 제시한다. 끝으로 마지막 장에서는 본 연구의 결론과 한계점 등에 대해 논의한다.

2. 연구 배경

2.1 기존 추천 기법의 한계점

인터넷 쇼핑몰에서의 상품추천시스템은 실생활에서도 활발하게 이용되고 있다. 그 대표적인 예로는 아마존(Amazon), 무비랜즈(MovieLens), 시디나우(CD Now), 제이씨 페니(JC Penny) 등이 있다. 이들 상품추천시스템의 가장 핵심적인 부분은 추천기법이라 할 수 있으며, 대표적인 추천알고리즘으로는 협동필터링과 내용기반필터링 등의 방법이 있다. 두 방법은 각각 장단점을 가지고 있으나 현실적으로는 협동필터링이 선행연구에서 더 활발하게 이용되고 있다[김재경 등, 2002; 김종우 등, 2004; 조윤호 등, 2004; 김재경 등, 2005 참고].

협동필터링은 사용자 사이의 연관성을 기반으로, 선호도 또는 구매 패턴이 유사한 고객군을 분류하고, 유사한 고객에 속하는 다른 사람들이 선호하는 상품을 추천하는 방법이다[Funakoshi and Ohguro, 2000]. 협동필터링에 대한 초기 연구로는 Tapestry[Goldberg et al., 1992], GroupLens[Resnick et al., 1994] 등의 사례가 대표적이며, Ringo와 Video Recommender 등과 같은 이메일과 웹 기반의 협동필터링 기법에 의한 추천시스템 등이 있다[Sarwar et al., 2000]. 협동필터링은 일반적으로 고객들이 동질적인 평가결과를 보이는 상품군에 대해 상대적으로 높은 예측력을 보이며, 데이터가 충분한 경우에는 다른 기법에 비해 상대적으로 높은 예측력을 보이는 장점을 가지고 있다[Konstan et al., 1997, Pazzani, 1999]. 이에 따라 협동필터링은 상품추천시스템 관련 연구에서 활발하게 이용되고 있으나 아래와 같은 한계점도 가지고 있다.

협동필터링은 기본적으로 상품에 대한 고객의 선호도 또는 구매이력자료를 바탕으로 추천을 하게 되므로 구매이력을 많이 보유하고 있는 대형 인터넷 쇼핑몰에서는 유용하지만, 구매이력이 상대적으로 부족한 중소 인터넷 쇼핑몰이나 사업 초기단계의 인터넷 쇼핑몰의 경우에서는 적용 가능성이 떨어진다고 할 수 있다[안현철 등, 2006]. 즉, 협동필터링의 속성상 구매이력이 부족한 경우에는 추천의 성과가 떨어질 수밖에 없으며 이런 점은 이미 선행연구에 의해 협동필터링의 가장 중요한 문제점 중 하나로 지적되고 있다[김재경 등, 2002; Cho et al., 2002; Kim et al., 2002; 김재경 등, 2003; 김종우 등, 2004; 조윤호 등, 2004; Cho and Kim, 2004; 김재경 등, 2005 참고]. 이런 문제점을 흔히 희박성(sparsity) 문제라고 하며, 희박성 문제를 완화하기 위해 선행연구에서는 웹 로그 정보를 활

용하여 간접적으로 선호도 데이터를 보충하고자 하였다[Cho et al., 2002; Kim et al., 2002; 김재경 등, 2003; Cho and Kim, 2004; 김재경 등, 2005]. 그러나 웹 로그 정보는 일반적으로 대용량이며 정제되지 않은 형태이어서 고객의 선호도 점수를 직접 취득하는 것만큼 전처리에 많은 시간과 비용이 소요된다는 단점이 있다.

협동필터링의 중요한 한계점 중 다른 하나는 고객과 거래 데이터가 증가함에 따라 유사한 고객군을 찾기 위한 연산량이 기하급수적으로 증가하는 현상이 발생할 수 있다는 것이다[김재경 등, 2002; Cho et al., 2002; Kim et al., 2002; 김재경 등, 2003; 조윤호 등, 2004; Cho and Kim, 2004; 김재경 등, 2005 참고]. 선행연구에서는 이런 문제점을 확장성(scalability) 문제라 하였는데, 이는 해결해야 하는 문제가 제시된 이후에야 추론을 시작하는 '게으른 학습방법(lazy learning technique)'의 일반적인 특징으로, 신속한 응답을 원하는 인터넷 사용자의 특성을 감안할 때 고객의 이탈을 유도할 수 있는 치명적인 한계점이다. 선행연구에서는 이 문제점을 보완하기 위해서 아래와 같은 여러 방법들을 제안하였다.

김재경 등[2002], Cho et al.[2002], Kim et al.[2002], 조윤호 등[2004], Cho and Kim[2004], 김재경 등[2005]은 확장성과 희박성의 문제를 보완하기 위하여 상품계층도(product taxonomy)를 활용하는 방법을 제안하였으나, 여전히 하나의 상품계층도의 각 상품계층군 안에서는 각 고객의 선호도가 제대로 반영되지 않아서 추천의 성과가 떨어지는 경우가 발생할 수 있다. 또, 상품계층도의 작성이 상품추천의 성과에 큰 영향을 미칠 수 있는데 선행연구에서는 전문가의 주관적인 판단을 참고하는 방식으로 연구를 진행하였으나 이 점 역시 한계점이 될 수 있다.

확장성 문제를 완화하기 위한 또 다른 노력으로는 군집분석을 협동필터링의 사전과정으로 수행하여 탐색공간을 줄이는 방법이 있다. 김재경 등[2003]은 K-Means 군집분석을 협동필터링 사전단계로 활용하여 탐색공간을 축소하였고, Roh et al.[2003]과 강부식[2003]은 군집분석 기법의 하나인 자기조직화지도를 활용하여 사례탐색공간을 축소하였다. Kim and Han[2001]은 협동필터링과 함께 '게으른 학습방법'의 하나인 사례기반추론에서 자기조직화지도 분석을 추론 이전 단계에 활용하여 분석 데이터의 양을 줄이고자 하였다. 그러나 이 같은 선행연구들은 '게으른 학습방법'에서 확장성 문제의 심각성을 인지하고 이에 대한 보완방법으로 군집분석의 방법을 이용하는 것이 유용하다는 주장만 하였을 뿐, 제안되었던 상품추천시스템의 성능을 향상시키는데 결정적인 요인이 될 수 있는 군집분석의 방법론적 개선에는 큰 관심을 두지 않았다.

이상의 선행연구들은 전통적인 협동필터링의 한계점을 보완하고자 노력하였으나 여전히 몇 가지 한계점을 남겨 두었다. 첫째, 몇몇 연구들은 웹로그나 상품계층도를 활용하여 희박성 문제를 보완할 수 있는 새로운 정보를 추가하고자 하였으나 이러한 시도는 추가적인 비용과 노력을 필요로 한다. 둘째, 확장성 문제를 완화하기 위해 협동필터링의 사용자-상품 매트릭스 상의 사용자 차원 또는 상품 차원의 축소를 시도하였으나, 두 개의 차원 중 하나의 차원의 축소를 통해서는 확장성 문제를 해소하는 데에 한계가 있다는 점이다. 즉, 사용자-상품 매트릭스의 두 개의 차원들을 동시에 축소하는 방식이 한 차원만을 축소하는 방법보다 더 확실한 확장성 문제 완화효과를 기대할 수 있음은 직관적으로 이해할 수 있다. 본 연구에서는 사용자-상품 매트릭스의 사용자 차원과 상품 차원을 동시에 축소하는 방안을 활용함으로써 확장성과 희박성의 문제를 효과적으로 완화할 수 있을 것으로 기대된다.

2.2 차원축소 관련연구

전술한 바와 같이 협동필터링의 두 가지 차원을 축소하고자 하는 노력은 있어 왔지만 많은 연구가 이루어지지 못한 이유는 협동필터링에 대한 연구가 최근에 이루어졌기 때문이다. 그러나 협동필터링과 유사한 알고리즘인 사례기반추론 등의 인공지능기법에 관한 연구들을 살펴보면 차원축소에 관한 선행연구들을 충분히 발견할 수 있다. 본 장에서는 사례기반추론의 차원축소를 시도한 선행연구를 검토함으로써 협동필터링에 이러한 차원축소방법이 어떻게 적용될 수 있을지 생각해 보기로 한다. 사례기반추론의 차원축소기법에 관한 연구는 크게 특징선택(feature selection), 사례선택(instance selection)의 두 가지 연구방향으로 나누어 진다. 여기서 사례기반추론의 '특징'은 협동필터링의 '상품'에, 사례기반추론의 '사례'는 협동필터링의 '사용자'에 대응되는 개념으로 생각할 수 있다.

협동필터링의 '상품'에 해당하는 사례기반추론의 '특징'은 사례기반추론 시스템에서 적절한 인덱싱을 위해 적절히 축소하고 선택하는 과정이 필수적인 부분이다. 특징을 미리 선택함으로써 시스템은 보다 빠른 시간에 유의한 특징공간만을 탐색함으로써 적절한 사례를 추천할 수 있다. 사례기반추론의 '특징'축소 선행연구로는 다음과 같은 것들이 있다. Stearns[1976]는 최적의 특징집단을 찾기 위하여 특징의 숫자를 바꾸어 가면서 가장 높은 적중율을 보이는 특징집단을 찾아 가는 순차전진선택방법을 제안하였다. 이는 시행착오법에 의한 단순한 방법으로 최적의 특징집단을 찾지 못할 가능성이 있다. 한편, Skalak[1994]은 최적화 기법 중 하나인 Hill climbing 알고리즘에 의한 방법을 제안하였고 Domingos[1997]는 군집분석을 활용하여 최적 특징집단을 찾고자 하였다. 또한, Cardie[1993], Cardie and Howe[1997], Jarmulak et al.[2000]은 의사결정나무에 기반한 방법을 제안하였다. 그러나 상기의 방법들은 대부분 국부 최적해에 수렴할 수 있는 특징을 가진 방법들이므로 전역 최적해를 찾는 데에는 한계가 있다. Siedlecki and Skalanski [1989]는 유전자알고리즘에 의한 특징선택방법을 제안하였는데, 유전자 알고리즘은 전역 최적해를 찾는데 유용한 방법이므로 이전 연구의 한계점을 보완할 수 있었다.

협동필터링의 '사용자'에 해당하는 사례기반추론의 '사례'는 협동필터링의 확장성 문제와 같이 사례기반추론 시스템의 중요한 단점이 될 수 있는 요소이다. 사례기반추론은 사례기반으로부터 유사한 사례를 추출한 후 사례추천을 하고, 그 결과에 따라 새로운 사례가 발생하면 이를 사례기반에 추가하는 작업을 반복한다. 이러한 과정을 통해 사례기반추론 시스템도 탐색해야 할 사례기반이 기하급수적으로 증가하는 문제를 겪을 수 있으며, 이는 협동필터링의 확장성 문제와 동일한 문제를 야기시킨다. 따라서 사례기반추론 연구에서도 사례증가 문제를 해결하기 위하여 사례공간을 축소하는 연구를 진행해 왔다.

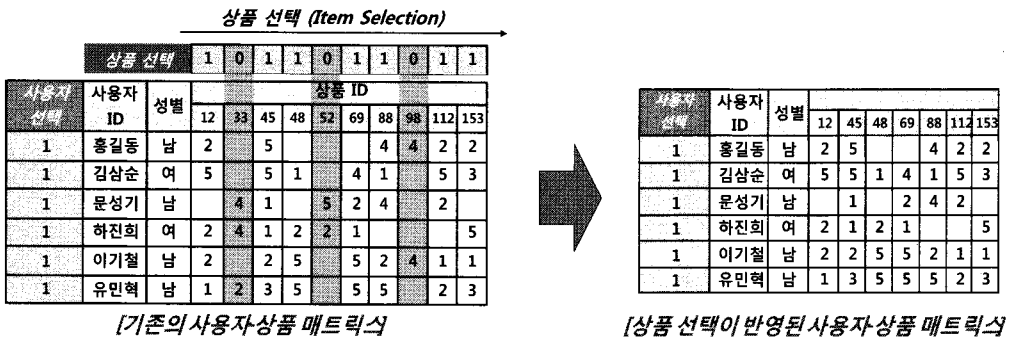
사례기반추론의 '사례'축소와 관련된 선행연구는 다음과 같다. Hart[1968]는 농축된 유사 사례 결합 알고리즘(condensed nearest neighbor algorithm)을, 그리고 Wilson[1972]이 윌슨의 방법(Wilson's method)을 제안하였다. 이들 방법들은 간단한 정보이득 개념에 기반하여 구성되었기 때문에, 적용이 쉽고 간단하다는 장점이 있다. 참조사례 선정과 관련한 최근 연구들은 예측 성과를 높이기 위해서, 보다 고급화된 수학적 기법이나 인공지능기법을 방법론으로 활용하고 있다. 예를 들어, Sanchez et al.[1997]은 근접 그래프 접근법(proximity graph approach)을 제안하였으며, Lipowezky[1998]는 선

형계획법(linear programming)에 기반한 참조사례 선정기법을 제안하였다. 아울러, Yan[1993]과 Huang et al.[2002]은 인공신경망을 활용한 참조사례 선정기법을 제안하였으며, Babu and Murty [2001]는 유전자 알고리즘에 기반한 기법을 제안하였다.

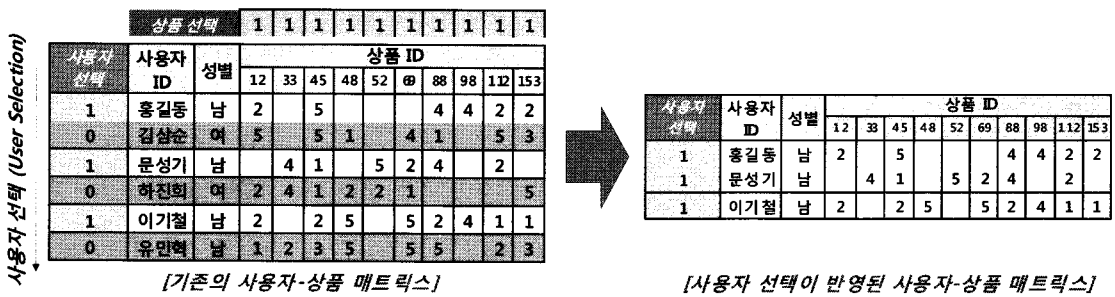
3. 유전자 알고리즘 기반의 협동필터링 차원축소 모형

전통적인 협동필터링 시스템의 예측성과와 효율성을 동시에 개선하기 위한 방법론으로, 본 연구에서는 유전자 알고리즘을 활용해 상품 선택(사례기반추론에서의 '특징 선택') 또는 사용자 선택(사례기반추론에서의 '사용자 선택')을 최적화하는 새로운 협동필터링 모형을 제안한다. 전자의 상품 선택 협동필터링 모형을 편의상 'ISCF(Item-Selected CF)'라 호칭하며, 후자의 사용자 선택 협동필터링 모형을 'USCF(User-Selected CF)'라 호칭하도록 한다. 본 제안모형들에서 나타나는 사용자-상품 매트릭스의 차원축소를 다음의 <그림 1>에서 시각적으로 묘사하고 있다. 이 그림에서 볼 수 있듯이, ISCF나 USCF는 각각 사용자-상품 매트릭스의 열과 행의 차원을 축소시켜줌으로써 협동필터링의 연산을 보다 가볍게 만들어 주는 효과를 제공한다. 또한, 대표성을 지니면서 설명력이 높은 사용자 혹은 상품만 고려하여 추천 결과를 생성하기 때문에, 보다 높은 정확도를 기대할 수 있다는 장점도 함께 갖게 된다.

이처럼 유전자 알고리즘을 이용해 차원을 축소하는 연구는 협동필터링 보다 상대적으로 더



(a) ISCF

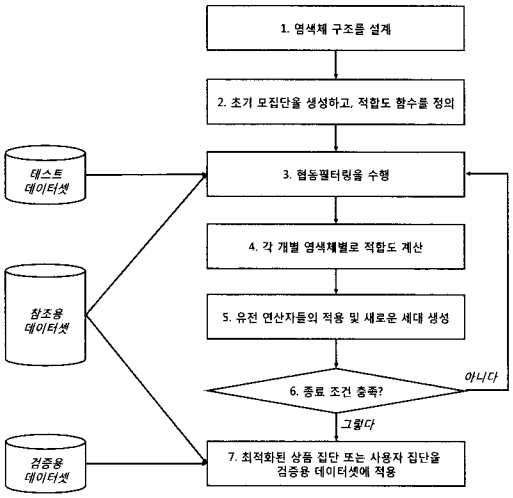


(b) USCF

〈그림 1〉 ISCF와 USCF의 차원축소 예시

오랜 역사를 갖고 있는 사례기반추론에서는 지금까지 종종 발표되어 왔다. 예를 들어, ISCF처럼 사례기반추론의 특징변수(feature)를 유전자 알고리즘을 이용해 최적화하고자 한 연구로는 Siedlecki and Sklanski[1989]와 Kim [2004] 등이 있었으며, USCF처럼 사례(instance)를 유전자 알고리즘을 이용해 최적화하고자 한 연구에는 Babu and Murty[2001] 등이 있었다. 이들 연구에서는, 적절한 특징변수의 선택 혹은 참조사례의 선택의 사례기반추론의 효율성은 물론 성과의 향상도 가져옴을 제시하고 있다. 때문에, 사실상 협동필터링과 사례기반추론은 그 근본원리가 유사하다는 점을 고려할 때, 본 연구의 제안모형은 ISCF나 USCF 역시 성과의 개선을 가져올 수 있을 것으로 기대할 수 있다.

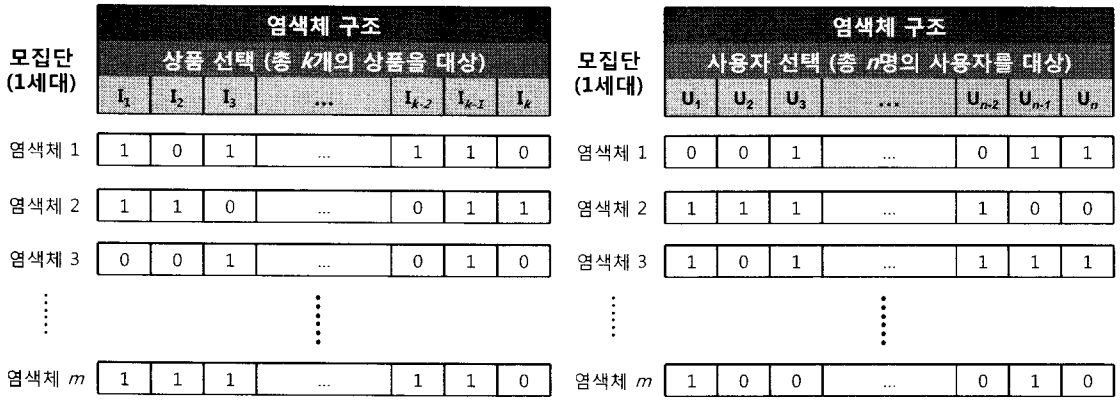
다음의 <그림 2>는 ISCF 및 USCF의 학습원리를 순서도(flowchart) 형태로 나타내고 있다. 이 그림에 제시된 것과 같이 제안 모형은 크게 7단계의 과정을 거쳐 진행되는데, 지금부터 각 단계별로 어떤 처리가 이루어지는지에 대해 보다 상세히 살펴보기로 한다.



〈그림 2〉 제안 모형의 학습원리

1단계 : 염색체 구조를 설계

유전자 알고리즘을 적용하기 위해서는 우선 '2진 문자열의 형태'로 구성되는 염색체(chromosome)의 구조를 설계해야 한다. 본 연구에서 제안하는 ISCF나 USCF의 경우, 하나의 비트로 '선택 여부(selection)'를 표현(예를 들어, 1은 선택, 0는 제외)할 수 있으므로, ISCF의 경우에는 전체 상품의 개수만큼의 길이를 갖는 염색체를 활용하면 되고, USCF의 경우는 전체 사용자의 수만큼 길이를 갖는 염색체를 활용하면 된다. 다음의 <그림 3>은 대상이 되는 전체 상품의 수가 k이고, 전체 사용자의 수가 n일 때, 즉 사용자-



(a) ISCF (b) USCF

〈그림 3〉 제안모형의 염색체 구조

상품 매트릭스가 $n \times k$의 크기를 갖는 행렬일 때, ISCF와 USCF의 염색체 구조를 나타내고 있다.

2단계 : 초기 모집단을 생성하고, 적합도 함수를 정의

본격적인 탐색에 들어가기 앞서, 초기 모집단(최적 사용자 혹은 상품의 집단을 찾기 위한 염색체들의 집합)을 초기화해야 하는데, 이 때 초기화는 각 염색체의 비트값들을 이진 무작위값으로 부여하는 방법을 통해 이루어지게 된다. 그런 다음, 이 모집단을 진화(evolution)시키기 위해서는, 모집단을 구성하는 각 염색체들을 평가할 어떤 기준이 필요하게 되는데, 이러한 기준을 '적합도 함수(fitness function)'이라고 부른다. 본 연구에서는 '예측의 대상이 되는 사례들에 대해 가장 정확한 예측력을 보이는 최적의 상품 집합 또는 사용자 집합을 결정하는 것'이 목적이므로, 우리는 예측의 정확도를 나타내는 지표 중 하나인 평균 MAE(mean absolute error)를 유전자 알고리즘의 적합도 함수로 사용하였다. MAE는 협동필터링과 관련한 기존 문헌에서 매우 빈번하게 사용되는 성과측정 지표로서, 사용자의 예측된 만족도와 실제 입력된 만족도 사이의 차이를 나타내는 지표이다[Breese et al., 1998; Sarwar et al., 1998; Goldberg et al., 2001; Roh et al., 2003; Kim and Yum, 2005]. 본 연구에서 MAE는 테스트 데이터셋(test dataset)에 대한 예측 오류를 의미하는데, 유전자 알고리즘은 이러한 오류를 최소화하는 방향으로 탐색을 진행하게 된다. 테스트 데이터셋 T에 대한 적합도 함수(f_T)를 수식으로 정리하면, 다음의 식과 같이 정리할 수 있다.

$$\text{Minimize } f_T = \frac{\sum_{k=1}^{N}\left(\sum_{i=1}^{n}|p_{k,i}-a_{k,i}|/n\right)}{N}$$

(N : 테스트 데이터셋 T에 포함되어 있는 사용자의 수, n : 테스트 데이터셋 T에 포함되어 있는 상품의 수, $p_{k,i}$: 사용자 k의 상품 i에 대한 예측된 만족도, $a_{k,i}$: 사용자 k의 상품 i에 대한 실제 만족도)

3단계 : 협동필터링을 수행

이 단계에서는 앞의 단계에서 선택된 결과에 의해 축소된 '사용자-상품 매트릭스'를 활용하여, 전통적인 협동필터링 방법을 적용하게 된다. 일반적인 협동필터링 방법은 크게 3단계로 이루어지게 되는데, 구체적으로 (1) 사용자간 유사도 측정, (2) 가장 유사도가 높은 이웃(neighbor) 선택, (3) 각 상품별 예측 만족도 산출의 세 단계로 이루어진다. 이 때, 사용자간 유사도는 보통 피어슨 상관계수(Pearson correlation)에 의해 산출되며, 가장 유사도가 높은 이웃을 선택할 때에는 1명이 아닌 특정 기준으로 만족하는 n명을 선택해, 이들의 만족도를 종합하여 예측결과를 산정하게 된다. 전통적인 협동필터링 방법에 대해서는 Breese et al.[1998]과 Sarwar et al.[1998] 등 협동필터링의 초기 연구에 상세하게 소개되어 있다.

4단계 : 각 개별 염색체 별로 적합도 계산

특정 염색체가 제공한 조건에 따라 모든 테스트 사례들에 대한 협동필터링 적용이 마무리되면, 그 다음에는 테스트 데이터셋 T에 대한 전체적인 적합도 함수(f_T)의 값을 계산할 수 있다. 이렇게 계산된 적합도는 이후 단계에서 각 개별 염색체의 우수성을 평가하는 지표(indicator)로 활용된다.

5단계 : 유전 연산자들의 적용 및 새로운 세대의 생성

이 단계에서는 앞의 4단계에서 도출된 각 개

별 염색체 별 적합도를 기준으로 하여, 적합도를 최소화하는 방향으로 유전자 알고리즘의 진화가 이루어질 수 있도록 유전 연산자들이 적용되는 과정이 수행된다. 유전 연산자에는 적자(the fittest)에 대한 선택, 교배, 그리고 돌연변이 등이 모두 포함되며, 이들 연산자들을 적용하여 모집단의 새로운 세대(generation)이 생성된다.

6단계 : 종료조건을 만족할 때까지, 3단계~5단계를 반복

유전자 알고리즘은 최적 혹은 유사최적해를 찾을 때까지 진화를 계속 반복하도록 설계되어 있다. 이에 본 연구의 모형에서도 앞의 5단계를 통해, 새로운 모집단이 생성되면, 그 모집단을 기준으로 다시 3단계~5단계의 활동을 반복수행하게 된다. 이는 사전에 정한 종료조건에 도달할 때까지 계속해서 반복수행되는데, 일반적으로 종료조건은 '최대 진화 세대수(the number of generations)'로 설정된다. 종료조건에 따라 진화과정이 마무리되면, 최적화된 상품 집단(ISCF의 경우) 혹은 최적화된 사용자 집단(USCF의 경우)이 메모리에 임시로 저장된다.

7단계 : 최적화된 상품 집단 또는 사용자 집단을 검증용 데이터셋에 적용

마지막 단계는 앞의 단계까지의 작업을 통해 찾아낸 최적의 상품 혹은 사용자 집단이 과연 모형구축에 활용되지 않은 데이터(unknown data)들에 대해서도 유효한 예측성과를 보이는지 최종적으로 확인하는 단계가 된다. 이를 위해 본 단계에서는 최종적으로 선택된 모형을 검증용 데이터셋(hold-out dataset)에 적용해 봄으로서, 제안모형의 최종적인 예측 정확도를 평균 MAE로 측정하게 된다.

4. 실험 설계

4.1 실험데이터 : MovieLens 데이터셋

본 연구의 제안모형을 검증하기 위한 실험은 연구목적으로 대중에 개방된 자료인 MovieLens 데이터셋[1]을 활용하였다. MovieLens는 본래 미국 미네소타 대학(University of Minnesota)의 GroupLens 연구 프로젝트 팀에 의해 웹 기반 영화 추천시스템으로 개발되었다. 하지만, 단순히 사용자들을 위한 추천시스템으로만 사용되는 것이 아니라, 연구자들의 모형 개선 연구를 지원하기 위한 실험 데이터의 원천으로도 활용되고 있으며, 추천 시스템과 관련하여 사용자 인터페이스(user interface)를 연구하는 틀(framework)로서의 기능도 수행하고 있다. MovieLens는 현재 두 종류의 데이터셋을 제공하고 있는데, 하나는 소규모 버전이고, 다른 하나는 대규모 버전이다. 본 연구에서는 대규모 버전을 사용하였는데, 이는 6,040명의 사용자로부터 입력된 3,900개의 영화에 대한 약 100만여건 정도의 평가(rating) 결과를 포함하고 있다. 데이터셋에는 특정 영화에 대한 사용자들의 평가 결과 외에도 사용자들에 대한 기본적인 인구통계정보(연령, 직업, 우편번호)도 포함되어 있다. 사용자들의 영화에 대한 평가는 5점 척도로 측정되어 있는데, 1~2점은 부정적인 의견을, 4~5점은 긍정적인 의견을, 그리고 3점은 중립적인 의견을 나타내는 것으로 해석할 수 있다.

전술한 바와 같이 본래 MovieLens의 대규모 데이터셋은 6,000명 이상의 사용자와 3,900여 개의 영화(상품)를 포함하고 있는 방대한 규모의 사용자-상품 매트릭스를 우리에게 제공하지

1) 다음 인터넷 사이트에서 배포됨.
http://www.grouplens.org/data/million(2008년 12월 확인).

만, 본 연구에서는 실험 환경상의 제약과 연구의 편의를 위해 무작위 추출을 통한 데이터셋 축소를 우선 수행하였다. 그렇게 해서, 총 1,035명의 사용자와 207개의 상품(영화)을 포함하는 사용자-상품 매트릭스를 대상으로 하여 실험을 수행하였다. 이 데이터셋을 본 연구의 제안모형을 검증하는 실험에 적용하기 위해서는 다시 데이터셋을 구분하는 작업이 요구되었다. 이를 위해, 우선 전체 상품들을 참조에 사용할 상품들(80%, 165건)과 검증에 사용할 상품들(20%, 42건)로 구분하였다. 아울러, 사용자의 경우에는 유전자 알고리즘을 통한 학습을 위해 세 개의 집단(참조용, 테스트용, 그리고 검증용)으로 구분할 필요가 있는데, 참조용 데이터셋을 60%(621명), 테스트용 데이터셋을 20%(207명), 그리고 검증용 데이터셋을 20%(207명)로 구성하였다.

4.2 실험 설계

유전자 알고리즘의 탐색을 위한 통제 변수들의 설정과 관련해서는, 100개의 개체(organism)로 구성된 모집단을 사용하였고, 교배율(crossover rate)은 0.7, 돌연변이율(mutation rate)은 0.1로 설정하였다. 종료 조건으로는 2000회 시도(즉, 100개의 개체로 된 모집단을 활용하였으므로, 총 20세대를 진화) 후 종료하는 것으로 설정하였다. 유전 연산자와 관련해, 교배기법은 일정 교배(uniform crossover) 기법을 적용하였다. 이 기법은 구조(schema)를 유지하면서, 두 부모 개체로부터 어떤 형태의 구조든지 생성해 낼 수 있는 장점을 갖고 있기 때문에, 특징변수들의 유전자 위치가 경우에 따라 왜곡될 수도 있는 1점 교배(single-point crossover)나 2점 교배(two-point crossover)에 비교해 더 우수한 기법인 것으로 평가된다. 돌연변이 연산자와 관련해서는, 각 특징변수에 대해 0에서 1사이의 난수를 발생시킨 다음 그 값이 정해진 돌연변이율보다 낮게 나왔을 때, 변수 값의 이진코드를 반대로 바꾸는 방법을 적용하였다.

실험은 총 3개의 모형에 대해 수행되었다. 우선 본 연구의 제안모형인 ISCF와 USCF를 기본적으로 실험하였으며, 이 두 모형의 실험결과를 비교하기 위한 비교모형으로 모든 사용자와 모든 상품을 전부 고려하는 전통적인 협동필터링 방법도 실험하였다. 편의상 이후 본 논문에서는 이 비교모형을 TCF(Typical CF)로 표기하기로 한다.

제안모형과 비교모형에 대한 실험은 Microsoft Excel 2003과 상업용 유전자 알고리즘 구현도구인 Palisade Software사의 Evolver Industrial Version 4.08을 이용해 개발된 별도의 소프트웨어에 의해 수행되었다. 협동필터링 알고리즘은 Microsoft Excel 2003의 VBA(Visual Basic for Applications)를 활용해 개발되었으며, VBA 프로그램은 Evolver의 유전자 알고리즘과 결합되어, 최적의 사용자 혹은 상품을 탐색하도록 설계되었다.

5. 실험 결과

본 연구에서는 제안모형인 ISCF와 USCF의 적용가능성을 검증하기 위해 기존 기법에 비해 얼마나 예측 정확도가 향상되는지를 살펴보고자 했는데, 이를 위해 검증용 데이터셋에 대한 평균 MAE를 측정해 보았다. 그 측정 결과가 다음의 <표 1>에 제시되어 있다.

표의 결과가 보여주고 있듯이, 제안모형인 ISCF(MAE = 1.05634)와 USCF(MAE = 1.82365) 모두 전통적인 협동필터링 기법을 적용한 TCF(MAE = 2.06860)에 비해 더 우수한 예측성과를 제공함을 알 수 있었다. 상기 두 모형은 예측결과의 안정성 측면에서도 기존 기법보다 더 우수

한 결과를 보여주었다. 예측결과의 안정성은 "입력 사례에 따라 예측오차가 얼마나 심하게 변화하는지"를 측정함으로서 그 정도를 평가해 보고자 했는데, 이를 위해 예측오차의 표준편차를 살펴보았다. <표 1>을 보면, 검증용 데이터셋에 대한 예측오차의 표준편차가 ISCF(0.37105) < USCF(0.37902) < TCF(0.41803)의 순서로 나타나고 있음을 알 수 있다. 이처럼, 제안모형인 ISCF나 USCF는 상대적으로 사례에 따라 예측오차가 갑자기 커지거나, 작아지지 않고 비교적 안정적으로 예측결과를 산출할 수 있다는 점도 함께 확인할 수 있었다.

〈표 1〉 실험결과

모형 이름	TCF	ISCF	USCF
평균 MAE (검증용 데이터셋)	2.06860	1.05364	1.82350
MAE의 표준편차 (검증용 데이터셋)	0.41803	0.37105	0.37972
선택된 상품수	165	54	165
선택된 사용자수	621	621	311
사용자-상품 매트릭스 크기	102,465	33,534	51,315

아울러, 상기 실험결과는 제안모형이 사용자-상품 매트릭스의 크기도 급격하게 축소시킴으로써 연산을 보다 효율적으로 만들어 주고 있음을 함께 보여주고 있다. TCF가 탐색해야 할 사용자-상품 매트릭스의 크기(행의 수와 열의 수를 곱해서 산출)를 100%라고 할 때, USCF는 50.08%, 그리고 ISCF는 32.73%만 필요로 하는 것으로 나타났다. 이 결과를 통해, 본 연구의 제안모형이 메모리 기반 협동필터링(memory-based CF) 방법의 효율성을 개선시켜, 확장성과 희박성 문제에 대한 해결책이 될 수 있음을 알 수 있다.

앞서 <표 1>에서 제시된 모형 간 예측 정확도의 차이가 과연 통계적으로도 유의한지를 검증하기 위해, 대응표본 t-검정(paired samples t-test)을 적용하였다. 대응표본 t-검정은 실험 전-후의 비교처럼, 동일한 표본에서 추출된 두 집단의 값들간에 서로 평균의 차이가 있는지를 검정할 때 사용된다. 이 기법은 종종 연관표본 t-검정(correlated samples t-test) 혹은 종속표본 t-검정(dependent samples t-test)이라는 이름으로 불리기도 한다[Green et al., 2000].

이 검정에서 귀무가설은 $H_0 : MAE_i - MAE_j \neq 0 (i = 1, 2, j = 2, 3)$이고, 대립가설은 $H_a : MAE_i - MAE_j \neq 0 (i = 1, 2, j = 2, 3)$이 된다. 이 때, MAE_k는 모형 k의 MAE를 의미한다. 다음의 <표 2>는 t-값과 통계적 유의수준을 포함한 대응표본 t-검정에 대한 수행 결과를 제시하고 있다.

〈표 2〉 대응표본 t-검정 결과

	USCF	ISCF
TCF	47.876*	34.784*
USCF		28.694*

주) *99% 신뢰수준 하에서 유의.

상기 표에 제시된 바와 같이, 제안모형인 ISCF와 USCF 모두 비교모형인 TCF에 비해 99% 신뢰수준 하에서 통계적으로 유의한 성과의 차이를 보이고 있음을 알 수 있다. 따라서 본 연구에서 제안한 사용자-상품 매트릭스 축약기법에 의한 새로운 추천시스템이 전통적인 추천시스템에 비해 추천의 품질이 우수함을 확인할 수 있다. 또한 제안모형 중 ISCF의 성과가 USCF의 성과와 비교해 볼 때, 역시 99% 신뢰수준 하에서 유의한 차이를 나타내고 있음을 확인할 수 있으며, 이는 상품 데이터의 축약이 사용자 데이터의 축약보다 추천의 품질 측면에서 더 유용함을 나타낸다.

6. 결 언

협동필터링은 학계와 업계 모두에서 매우 인기리에 적용되어 온 기법으로, 오늘날 많은 추천시스템의 핵심 엔진들이 협동필터링에 기반하고 있다. 그러나, 희박성과 확장성이라는 두 본질적인 문제로 인해 현실세계에서의 협동필터링 적용이 많은 제약을 받고 있는 것도 사실이다. 본 논문에서는 유전자 알고리즘을 활용한 사용자-상품 매트릭스의 두 차원, 즉 사용자 차원과 상품 차원의 효과적인 축소기법을 제시함으로써 희박성과 확장성 문제를 해결해 보고자 하였다. 아울러, 협동필터링의 효율성만 개선하는 것이 아니라, 적절한 사용자-상품만 유사도 계산에 사용함으로써 예측 정확도도 함께 개선시킬 수 있는 일석이조의 방법을 제안코자 하였다.

모형의 검증을 위한 실험 결과, 제안모형들이 예상대로 전통적인 협동필터링 기법과 비교해 보다 효율적으로 작동되면서도, 동시에 더 높은 예측 정확도를 산출함을 확인할 수 있었다. 아울러, 예측결과의 안정성 측면에서도 기존 기법에 비해 더 우수한 성과를 보임을 확인할 수 있었다.

본 연구의 결과에서 한 가지 흥미로운 사실은 본 연구의 제안모형 중 ISCF와 USCF의 실험결과를 서로 비교했을 때, 예측 정확도, 안정성, 사용자-상품 매트릭스의 탐색공간 크기 등 모든 측면에서 ISCF가 더 우수한 성과를 보인다는 점이다. 차원의 관점에서만 볼 때에는 사용자 차원은 총 621개의 항목으로 구성되어 있고, 상품 차원은 단 165개로만 구성되어 있기 때문에, 직관적으로는 사용자 차원을 최적화 하는 것이 더 효과적일 것으로 추정해 볼 수 있다. 하지만, 앞서 제시한 바와 같이 결과는 반대로 나타났다. 이러한 현상이 나타난 원인으로는 물론 실험에 적용된 데이터셋의 고유한 특성도 생각해 볼 수 있겠지만, 일반적으로 협동필터링을 최적화하는데 있어서, 사용자보다는 상품이라는 차원이 더 큰 영향을 미칠 가능성이 있다는 점을 시사하고 있다고도 해석할 수 있다. 이러한 사용자 차원과 상품 차원에 대한 상대적 영향력에 대한 분석은 추후 다른 데이터셋을 적용한 연구에서 좀 더 깊이 있게 연구되고, 고찰될 필요가 있다.

이 외에 본 연구가 갖는 다른 한계점과 향후 연구방향을 살펴보면 다음과 같다. 우선 본 연구의 학습과정을 더 효율화할 수 있는 방법에 대한 연구가 요구된다. 앞의 제안모형에서 설명했듯이, 최적인 상품 또는 사용자를 선택하기 위해서는 유전자 알고리즘이 수천 번의 협동필터링 알고리즘을 반복 수행해야 하는데, 이는 현업에서 실제 적용하고자 할 때, 상당히 방대한 연산 자원을 요구할 수 있다. 때문에, 유전자 알고리즘의 탐색과정을 좀 더 효율적으로 개선시키는 방법 등을 활용하여, 제안모형의 학습과정을 보다 단순화시키고 효율화시키는 연구가 추후 수행되어야 한다.

둘째로, 제안모형의 확장에 대한 후속 연구가 요구된다. 본 연구에서는 사용자와 상품의 단일 차원 축소만 다루었지만, 협동필터링이 아닌 사례기반추론에서는 두 차원을 동시에 최적화하는 모형에 대한 연구가 요 근래 활발히 논의되고 있다[Kuncheva and Jain, 1999; Rozsypal and Kubat, 2003; Ahn et al., 2006, 2007]. 이러한 맥락에서 협동필터링에서도 사용자-상품 매트릭스의 두 차원을 동시에 최적화하는 모형에 대한 연구가 추후 이루어져야 할 것이다.

앞서 언급한 두 차원의 동시최적화 외에, 사용자간 유사도 산출 시 각 상품별 혹은 사용자별 가중치를 차별화하는 방안에 대한 연구도 향후 고려해 볼 필요가 있다. 본 연구에서는 두 차원에 대한 '선택(selection)'만 고려했을 뿐,

'가중치 부여(weighting)'는 고려하지 않았다. 하지만, 협동 필터링에서 상품에 대한 적절한 가중치의 선정이 예측 정확도의 개선을 가져올 수 있다는 기존 연구도 발표된 바 있고[Yu et al., 2003; Zeng et al., 2004], 0~1사이의 실수를 가중치로 부여하는 '가중치 부여 방식'의 경우, 0 또는 1의 정수를 가중치로 부여하는 '선택 방식'을 항상 포함하고 있음을 고려할 때, 가중치 부여를 접목한 모형의 개선에 대한 연구가 추후 이루어져야 할 것으로 보인다.

마지막으로, 본 연구의 검증결과는 실험용 데이터셋에 따라 바뀔 수 있는 가능성이 있다. 따라서, 협동필터링의 경우 본 연구에서 활용한 MovieLens 데이터셋 외에도 EachMovie 등 다른 공개된 데이터셋이 존재하므로, 다른 데이터셋에서도 제안모형의 예측결과가 유효한 지에 대한 검증이 추후 보완되어야 할 것이다. 아울러, 추천시스템의 성능 개선 연구에서는 추천결과가 실제 매출이나 기업성과에 어떻게 영향을 미치는지를 파악하는 것이 매우 중요함에도 불구하고, 본 연구에서는 실험상으로만 모형의 유용성을 확인하고 있다. 때문에, 추후 실제 인터넷 쇼핑몰 등에 제안모형을 적용해 보고, 모형의 실용적 가치를 실증적으로 확인하는 후속 연구가 수행되어야 할 것으로 생각된다.

참 고 문 헌

[1] 강부식, " 자기 조직화 신경망을 이용한 협력적 여과 기법의 웹 개인화 시스템에 대한 연구", *한국지능정보시스템학회논문지*, 제9권 제3호, 2003, pp. 117-135.

[2] 김경재, 안현철, "개선된 데이터마이닝 기술에 의한 웹 기반 지능형 추천시스템 구축", *Journal of Information Technology Applications and Management*, 제12권 제3호, 2005, pp. 42-56.

[3] 김재경, 서지혜, 안도현, 조윤호, "A personalized recommendation methodology based on collaborative filtering", *한국지능정보시스템학회 논문지*, 제8권 제2호, 2002, pp. 139-157.

[4] 김재경, 안도현, 조윤호, "Development of a personalized recommendation procedure based on data mining techniques for internet shopping malls", *한국지능정보시스템학회논문지*, 제9권 제3호, 2003, pp. 177-191.

[5] 김재경, 안도현, 조윤호, "개인별 상품추천시스템, WebCF-PT : 웹 마이닝과 상품계층도를 이용한 협업필터링", *경영정보학연구*, 제15권 제1호, 2005, pp. 63-79.

[6] 김종우, 배세진, 이홍주, "협업 필터링 기반 개인화 추천에서의 평가자료의 희소정도의 영향", *경영정보학연구*, 제14권 제2호, 2004, pp. 131-149.

[7] 조윤호, 박수경, 안도현, 김재경, "재구성된 제품 계층도를 이용한 협업 추천 방법론 및 그 평가", *한국경영과학회지*, 제29권 제2호, 2004, pp. 59-75.

[8] 안현철, 한인구, 김경재, "연관규칙기법과 분류모형을 결합한 상품 추천 시스템 : G 인터넷 쇼핑몰의 사례", *Information Systems Review*, 제8권 제1호, 2006, pp. 181-201.

[9] Ahn, H., Kim, K.-j., and Han, I., "Hybrid Genetic Algorithms and Case-based Reasoning Systems for Customer Classification", *Expert Systems*, Vol. 23, No. 3, 2006, pp. 127-144.

[10] Ahn, H., Kim, K.-j., and Han, I., "A Case-based Reasoning System with

the Two-Dimensional Reduction Technique for Customer Classification", *Expert Systems with Applications*, Vol. 32, No. 4, 2007, pp. 1011-1019.

[11] Babu, T. R. and Murty, M. N., "Comparison of Genetic Algorithm Based Prototype Selection Schemes", *Pattern Recognition*, Vol. 34, No. 2, 2001, pp. 523-525.

[12] Breese, J. S., Heckerman, D., and Kadie, C., "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence*, 1998, pp. 43-52.

[13] Cardie, C., "Using Decision Trees to Improve Case-Based Learning", *Proceedings of the 10th International Conference on Machine Learning*, 1993, pp. 25-32.

[14] Cardie, C., and Howe, N., "Improving Minority Class Prediction Using Case-Specific Feature Weights", *Proceedings of the 14th International Conference on Machine Learning*, 1997, pp. 57-65.

[15] Cho, Y. H. and Kim, J. K., "Application of Web Usage Mining and Product Taxonomy to Collaborative Recommendations in E-Commerce", *Expert Systems with Applications*, Vol. 26, No. 2, 2004, pp. 233-246.

[16] Cho, Y. H., Kim, J. K., and Kim, S. H., "A Personalized Recommender System Based on Web Usage Mining and Decision Tree Induction", *Expert Systems with Applications*, Vol. 23, No. 3, 2002, pp. 329-342.

[17] Domingos, P., "Context-Sensitive Feature Selection for Lazy Learners", *Artificial Intelligence Review*, Vol. 11, Nos. 1-5, 1997, pp. 227-253.

[18] Funakoshi, K., and Ohguro, T., "A Content-Based Collaborative Recommender System with Detailed Use of Evaluations", *Proceedings of the 4th International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, 2000, pp. 253-256.

[19] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D., "Using Collaborative Filtering to Weave an Information Tapestry", *Communications of ACM*, Vol. 35, No. 12, 1992, pp. 61-70.

[20] Goldberg, K., Roeder, T., Gupta, D., and Perkins, C., "Eigentaste : A Constant Time Collaborative Filtering Algorithm", *Information Retrieval Journal*, Vol. 4, No. 2, 2001, pp. 133-151.

[21] Green, S. B., Salkind, N. J., and Akey, T. M., *Using SPSS for Windows, Second Ed.* Prentice Hall, 2000.

[22] Hart, P. E., "The Condensed Nearest Neighbor Rule", *IEEE Transactions on Information Theory*, Vol. 14, No. 3, 1968, pp. 515-516.

[23] Huang, Y. S., Chiang, C. C., Shieh, J. W., and Grimson, E., "Prototype Optimization for Nearest-Neighbor Classification", *Pattern Recognition*, Vol. 35, No. 6, 2002, pp. 1237-1245.

[24] Jarmulak, J., Craw, S., and Rowe, R., "Self-Optimizing CBR Retrieval", *Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence*, 2000, pp. 376-383.

[25] Kim, K., "Toward Global Optimization of Case-Based Reasoning Systems for Financial Forecasting", *Applied Intelligence*, Vol. 21, No. 3, 2004, pp. 239-249.

[26] Kim, J. K., Cho, Y. H., Kim, W. J., Kim, J. R., and Suh, J. H., "A personalized recommendation procedure for Internet shopping support", *Electronic Commerce Research and Applications*, Vol. 1, 2002, pp. 301-313.

[27] Kim, K.-S., and Han, I., "The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases", *Expert Systems with Applications*, Vol. 21, No. 3, 2001, pp. 147-156.

[28] Kim, D., and Yum, B.-J., "Collaborative Filtering Based on Iterative Principal Component Analysis", *Expert Systems with Applications*, Vol. 28, No. 4, 2005, pp. 823-830.

[29] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Riedl, J., "GroupLens : Applying Collaborative Filtering to Usenet News", *Communications of ACM*, Vol. 40, No. 3, 1997, pp. 77-87.

[30] Kuncheva, L. I. and Jain, L. C., "Nearest Neighbor Classifier : Simultaneous Editing and Feature Selection", *Pattern Recognition Letters*, Vol. 20, No. 11-13, 1999, pp. 1149-1156.

[31] Lipowezky, U., "Selection of the optimal prototype subset for 1-NN classification", *Pattern Recognition Letters*, Vol. 19, No. 10, 1998, pp. 907-918.

[32] Pazzani, M. J., "A framework for collaborative, content-based and demographic filtering", *Artificial Intelligence Review*, Vol.13, No. 5-6, 1999, pp. 393-408.

[33] Resnick, P., Iacovou, N., Suchak, M., and Bergstrom, P., "GroupLens : An open architecture for collaborative filtering of netnews", *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 1994, pp. 175-186.

[34] Roh, T. H., Oh, K. J., and Han, I., "The Collaborative Filtering Recommendation Based on SOM Cluster-Indexing CBR", *Expert Systems with Applications*, Vol. 25, No. 3, 2003, pp. 413-423.

[35] Rozsypal, A. and Kubat, M., "Selecting Representative Examples and Attributes by a Genetic Algorithm", *Intelligent Data Analysis*, Vol. 7, No. 4, 2003, pp. 291-304.

[36] Sanchez, J. S., Pla, F., and Ferri, F. J., "Prototype Selection for the Nearest Neighbour Rule Through Proximity Graphs", *Pattern Recognition Letters*, Vol. 18, No. 6, 1997, pp. 507-513.

[37] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J., "Application of dimensionality reduction in recommender

systems : A case study", *Proceedings of the WebKDD Workshop at the ACM SIGKKD*, 2000.

[38] Sarwar, B. M., Konstan, J. A., Borchers, A., Herlocker, J., Miller, B., and Riedl, J., "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System", *Proceedings of 1998 ACM Conference on Computer Supported Cooperative Work (CSCW)*, 1998, pp. 345-354.

[39] Siedlecki, W., and Sklanski, J., "A Note on Genetic Algorithms for Large-Scale Feature Selection", *Pattern Recognition Letters*, Vol. 10, No. 5, 1989, pp. 335-347.

[40] Skalak, D. B., "Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms", *Proceedings of the 11th International Conference on Machine Learning*, 1994, pp. 293-301.

[41] Stearns, S., "On Selecting Features for Pattern Classifiers", *Proceedings of the 3rd International Conference on Pattern Recognition*, 1976, pp. 71-75.

[42] Wilson, D. L., "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 2, No. 3, 1972, pp. 408-421.

[43] Yan, H., "Prototype Optimization for Nearest Neighbor Classifier Using a Two-Layer Perceptron", *Pattern Recognition*, Vol. 26, No. 2, 1993, pp. 317-324.

[44] Yu, K., Xu, X., Ester, M., and Kriegel, H. -P., "Feature Weighting and Instance Selection for Collaborative Filtering : An Information-Theoretic Approach", *Knowledge and Information Systems*, Vol. 5, No. 2, 2003, pp. 201-224.

[45] Zeng, C., Xing, C.-X., and Zhou, L.-Z., "Similarity Measure and Instance Selection for Collaborative Filtering", *International Journal of Electronic Commerce*, Vol. 8, No. 4, 2004, pp. 115-130.

저자소개

김 경 재

현재 동국대학교 경영대학 경영정보학과 부교수로 재직 중이다. 한국과학기술원에서 경영정보시스템을 전공으로 박사학위를 취득하였으며, 연구관심분야는 데이터마이닝, 지능형 신용평가시스템, 고객관계관리 등이다.

안 현 철

현재 국민대학교 비즈니스IT학부 전임강사로 재직 중이다. KAIST에서 산업경영학사를, 그리고 KAIST 테크노경영대학원에서 경영정보시스템 전공으로 경영공학석사와 경영공학박사를 취득하였다. 학위 취득 후, 한국국방연구원(KIDA)의 선임연구원 그리고 성신여자대학교 사회과학대학 경영학과 전임강사로 근무하였다. 주요 관심분야는 인공지능 및 데이터마이닝, 고객관계관리, 재무정보시스템, 정보시스템 및 집단지성의 수용과 관련한 행동 모형 등이다.

◆ 이 논문은 2009년 02월 20일 접수하여 1차 수정을 거쳐 2009년 03월 13일 게재확정되었습니다.