엔트로피 가중치 및 SVD를 이용한 군집 특징선택

(Cluster Feature Selection using Entropy Weighting and SVD)

이 영 석 † 이 수 원 ††

(Young Seok Lee) (Soowon Lee)

요 약 군집화는 객체들의 특성을 분석하여 유사한 성질을 갖고 있는 객체들을 동일한 집단으로 분류하는 방법이다. 전자 상거래 자료처럼 차원 수가 많고 누락 값이 많은 자료의 경우 입력 자료의 차원축약, 잡음제거를 목적으로 SVD를 사용하여 군집화를 수행하는 것이 효과적이지만, SVD를 통해 변환된 자료는 원래의 속성 정보를 상실하기 때문에 군집 결과분석에서 원본 속성의 가치 해석이 어렵다. 따라서 본 연구는 군집화 수행 후 엔트로피 가중치 및 SVD를 이용하여 군집의 중요한 속성을 발견하기 위한 군집 특징 선택 기법 ENTROPY-SVD를 제안한다. ENTROPY-SVD는 자료의 속성들과 유사객체 군과의 묵시적인 은닉 구조를 활용하기 위하여 SVD를 이용하고 유사객체 군에 포함된 응집도가 높은 속성들을 발견하기 위하여 엔트로피 가중치를 사용한다. 또한 ENTROPY-SVD를 적용한 모델 기반의 협력적 여과 기법의 추천 시스템 CFS-CF를 제안하고 그 효용성 및 효과를 평가한다.

키워드 : 특징선택, 클러스터링, 협력적 여과, SVD, 엔트로피 가중치

Abstract Clustering is a method for grouping objects with similar properties into a same cluster. SVD(Singular Value Decomposition) is known as an efficient preprocessing method for clustering because of dimension reduction and noise elimination for a high dimensional and sparse data set like E-Commerce data set. However, it is hard to evaluate the worth of original attributes because of information loss of a converted data set by SVD. This research proposes a cluster feature selection method, called ENTROPY-SVD, to find important attributes for each cluster based on entropy weighting and SVD. Using SVD, one can take advantage of the latent structures in the association of attributes with similar objects and, using entropy weighting one can find highly dense attributes for each cluster. This paper also proposes a model-based collaborative filtering recommendation system with ENTROPY-SVD, called CFS-CF and evaluates its efficiency and utilization.

Key word : Feature Selection, Clustering, Singular Value Decomposition, Entropy Weighting

1. 서 론

인터넷 및 데이터베이스의 발달로 상품 및 서비스를 제공하는 업체들은 방대한 자료들을 보유하게 되었으며 축적된 자료들의 가치를 높이기 위한 기술로 데이터마이닝 연구가 활성화되었다. 데이터마이닝 연구분야에는 분류(Classification), 군집화(Clustering), 연관규칙 탐사(Association Rule Discovery), 예측(Prediction) 등과 같은 다양한 연구분야가 있다. 이러한 연구 분야 중에서 군집화는 객체들의 특성을 분석하여 유사한 성질을 갖고 있는 객체들을 동일한 집단으로 분류하는 방법이다. 군집화 알고리즘에서의 입력자료 구성은 객체-속성집합으로 이루어져 있으며, 객체는 군집화 대상이며 속성집합은 객체를 구분하는 기준으로 사용된다. 군집화의 응용 분야 중 전자상거래 고객 상품 구입 목록을 기준으로 유사 고객군을 형성하거나 문서 내 포함 단어를 기준으로 문서를 군집화하는 문제에서는 속성집합의 수가 너무 많기 때문에 알고리즘의 수행 성능에 속성집합의 수가 영향을 미치게 된다. 따라서 이러한 문제를 해결

† 학생회원 : 숭실대학교 컴퓨터학과
pado1004@valentine.ssu.ac.kr
†† 종신회원 : 숭실대학교 컴퓨터학부 교수
swlee@computing.soongsil.ac.kr
논문접수 : 2001년 8월 3일
심사완료 : 2002년 1월 10일

하기 위한 방법으로 특징 가중치(Feature Weighting)와 같은 기법이 군집화의 전처리로 사용되거나 입력 자료를 분석하여 다른 형태로 변환하는 SVD(Singular Value Decomposition), 주성분분석(Principal Component Analysis)등과 같은 방법을 사용하여 자료를 재구성하기도 한다. 그러나 자료를 재구성하는 방법은 원본 속성 자료를 축약하는 특성이 있어 결과분석 단계에서 원본 자료와 군집화 결과의 객체 소속 정보를 바탕으로 군집을 형성하는 기준 속성들을 재분석해야한다. 본 연구는 군집 결과의 분석 및 응용을 위한 중간 단계의 자동화 과정으로 SVD와 같은 기법을 사용하는 군집화의 후처리 결과를 재분석하여 군집을 형성하는데 기준이 된 대표 속성 집합을 발견하고 군집 내에서의 속성의 가치와 중요도를 가중치로 설정하는 기법의 연구이다.

본 연구에서는 유사한 성질을 가지는 군집의 속성(Attribute)들 중에서 군집을 대표하는 속성들의 집합을 군집의 특징(Feature)으로 보고 특징을 구성하는 속성들의 중요도(Importance) 및 가치(Value)를 위한 속성 가중치(Attribute Weighting)를 설정한다. 군집 대표 속성들의 가중치를 부여하는 기법을 군집 특징 선택이라 한다. 본 연구에서는 군집 속성들의 가중치 재설정을 위하여 속성의 엔트로피 가중치(Entropy Weighting) 및 SVD를 이용한 군집 특징 선택 알고리즘 ENTROPY-SVD를 구현하고, 이 선택 기법에 의해 발견된 대표 속성들의 효용성을 평가하기 위한 방안으로 군집 대표 속성 집합을 사용하는 추천시스템 CFS-CF를 구현한다.

본 연구의 2장에서는 제안하고자 하는 기법의 기반을 이루는 특징선택 기법, 엔트로피 가중치 기법, SVD 기법과 입력 자료의 형성을 위해서 사용된 군집화 알고리즘, 제안 기법의 평가를 위한 추천 시스템에 대해서 언급한다. 3장에서는 본 연구에서 제안하는 군집 특징선택 기법을 단계별로 자세히 설명하며, 4장에서는 구축된 추천 시스템과 평가를 위한 실험방법 및 결과 분석 내용을 기술한다. 5장에서는 결론 및 향후과제에 대하여 언급한다.

2. 관련 연구

2.1 특징 선택 기법

특징 선택(Feature Selection)은 분류 및 군집화 알고리즘의 성능 향상을 목적으로 전처리 단계에서 사용되는 기법이다. 즉, 알고리즘의 정확도가 만족되는 범위의 속성 N개를 취하여 입력자료를 형성하는데 사용된다. 따라서 각 알고리즘은 전체 속성을 사용하는 것만큼의 정확도를 만족시키면서 입력 자료의 차원 축소에 의해 알고리즘의 수행 속도가 향상된다[1][2][3].

2.1.1 Document Frequency(DF)

문서 빈도수(DF)는 어떤 한 속성이 발생하는 전체 문서의 개수를 말한다. 이 기법의 기본 가정은 빈도수가 낮은 속성은 정보가치가 떨어지며 전체 수행 성능에 큰 영향을 미치지 않는다고 보는데 있다. DF 임계치 방법은 주로 문서 분류 혹은 군집화 방법에서 사용되어 DF가 낮은 단어를 제거하여 입력 자료의 차원을 축소하는데 사용된다. 그러나 빈도수가 낮은 속성이지만 속성의 정보가치가 높은 경우도 있기 때문에 임계치를 설정하는 방법에 있어서 실험적인 조정 작업이 수행되어야 한다.

2.1.2 Information Gain(IG)

$$IG(t) = -\sum_{i=1}^{m} \Pr(c_i)\log\Pr(c_i) + \Pr(t)\sum_{i=1}^{m} \Pr(c_i \mid t)\log\Pr(c_i \mid t) + \Pr(\bar{t})\sum_{i=1}^{m} \Pr(c_i \mid \bar{t})\log\Pr(c_i \mid \bar{t}) \quad (1)$$

IG는 주로 기계학습 연구 분야에서 사용되는 기법으로 클래스 예측을 위하여 자료 내 속성의 존재 여부를 바탕으로 속성의 정보량을 측정한다. 예를 들면 입력 자료가 문서가 되고 속성은 단어일 경우 각 단어의 정보량 IG(t)는 식 (1)에 의해서 측정된다. m개의 문서를 대상으로 Pr(t)는 전체 m개의 문서에서 단어 t를 포함하는 문서의 비율, Pr($\bar{t}$)는 전체 m개의 문서에서 단어 t가 없는 문서의 비율, Pr(ci)는 전체 m개의 문서에서 i번째 클래스 c_i에 포함되어 있는 문서의 비율, Pr($ci \mid t$)는 단어 t를 포함하는 전체 m개의 문서들 중에서 클래스 ci안의 단어 t를 포함하는 문서들과의 비율, Pr($ci \mid \bar{t}$)는 단어 t를 포함하지 않는 전체 m개의 문서들 중에서 클래스 c_i 안의 단어 t를 포함하지 않는 문서들과의 비율을 의미한다.

2.1.3 Mutual Information(MI)

		클래스 c	클래스 ~c
단어	t	A	B
	~t	C	D

그림 1 발생 빈도표와 예

MI는 클래스와 속성간의 연관성을 측정하기 위한 기법으로 사용된다. 그림 1에서처럼 단어 t와 클래스 c의 발생 빈도표(contingency table)를 사용하는데, A는 클래스 c에 포함되는 문서들 중에서 단어 t를 가지고 있

는 문서들의 수, B는 클래스 c를 제외한 다른 클래스들에서 단어 t를 가지고 있는 문서들의 수, C는 클래스 c에 포함되는 문서들 중에서 단어 t를 가지고 있지 않는 문서들의 수, 그리고 N은 전체 문서의 수를 의미한다. MI에서는 A, B, C의 빈도 정보를 통하여 단어 t와 클래스 c와의 의존성을 수치적으로 산출해 준다.

$$I(t,c) = \log \frac{\Pr(t \wedge c)}{\Pr(t) \times \Pr(c)} \quad (2)$$

$$I(t,c) \approx \log \frac{(A \times N)}{(A+C) \times (A+B)} \quad (3)$$

식 (2)와 식 (3)의 I(t,c) 값이 "0"일 경우 두 관계는 독립적이다. 각 클래스의 전체 단어에 대한 I(t,c)를 계산한 후 각각의 t에 대한 최종 값을 산출하기 위해서 식 (4),(5)를 사용하여 평균 혹은 최대 값을 구한 후 단어 최종 t의 값을 결정한다. 전체 단어들에 대한 MI값이 결정되면 K개의 속성을 선택하여 특징 선택을 한다.

$$I_{avg}(t) \approx \sum_{i=1}^{m} \Pr(c_i) I(t,c_i) \quad (4)$$

$$I_{\max}(t) = \max_{i=1}^{m} \{I(t,c_i)\} \quad (5)$$

2.1.4 x^2 Statistic(CHI)

$$\chi^2(t,c) = \frac{N \times (AD-CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (6)$$

CHI는 MI와 같이 발생 빈도표를 사용하여 A, B, C를 계산하고 D는 클래스 c에 포함되지 않는 문서들 중에서 단어 t를 포함하지 않는 문서들의 수가 된다(그림 1). 식 (7), (8)을 사용하여 MI와 같은 방식으로 클래스 c와 단어 t와의 CHI 값들 중 특정 단어 t에 대한 평균 혹은 최고 값을 구한 후 K개의 속성을 선택하여 특징선택을 한다.

$$\chi^2_{avg}(t) \approx \sum_{i=1}^{m} \Pr(c_i) \chi^2(t,c_i) \quad (7)$$

$$\chi^2_{\max}(t) = \max_{i=1}^{m} \{\chi^2(t,c_i)\} \quad (8)$$

2.2 SVD를 이용한 LSI 기법

LSI(Latent Semantic Indexing) 혹은 LSA(Latent Semantic Analysis)는 정보검색 연구에서 검색어의 유의어(Synonym), 다의어(Polysemy) 문제를 해결하기 위한 연구이다. LSI의 연구는 웹 문서 검색 분야에서 주로 나타나는데 입력되는 사용자의 부실한 검색정보만을 통해서 양질의 검색결과를 만들기 위한 방법으로 사용된다. LSI는 입력 검색어를 의미적으로 확장하는 과정을 수행한다. 검색어의 확장을 위해서 수작업에 의해 구축된 정보를 사용하기도 하며 시스템 내부에 축적된 문서 정보를 바탕으로 검색어 확장을 위한 정보를 구축하기도 한다. SVD는 행렬의 분해방법으로 원본 입력행렬의 근사행렬을 만들어 낸다. 따라서 검색 시스템의 단어-문서집합의 규모가 큰 행렬을 원본 그대로 사용하게되면 성능이 떨어지므로 SVD를 사용하여 근사행렬을 만들어 검색에 이용한다. SVD는 입력 자료의 잠정적인 구조(Latent Structure)를 파악할 수 있는 이점이 있어 검색어의 의미를 확장하는데 사용된다 [4][5][6][7]. 즉 자동차와 같은 검색어가 주어졌을 경우 실제 자동차 관련 문서들에는 "자동차"라는 단어와 함께 같은 의미로 사용되는 단어들도 포함되어 있다. "자동차", "차", "아반테", "티코" 등과 같은 단어들은 모두 자동차 관련 단어로서 모두 같은 의미로 해석할 수 있다. LSI방법을 적용하기 위해서는 먼저 TFIDF(Term Frequency Inverse Document Frequency) 가중치 부여 기법을 사용하여 단어-문서 행렬을 구성한다. 이 기법은 한 문서 내 발생 빈도가 높은 단어의 가중치를 낮게 책정하고 전체 문서에서 골고루 분포되어 있고 발생 빈도가 높은 단어의 가중치를 높게 책정하는 특징이 있다. TFIDF 기법을 통해서 만들어진 단어-문서 행렬(Term-Document Matrix)을 SVD에 적용하게 되면 SVD는 입력 행렬을 3개의 다른 행렬(U,Σ,V)들로 분해하는데, 이렇게 분해된 행렬 U,Σ,V를 검색에 이용한다 [5][8].

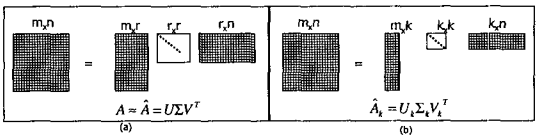

그림 2 SVD 개념

만일 입력자료가 단어-문서 행렬로 표현되었을 경우 분해된 행렬 U는 각 문서들의 관계를 토대로 구성된 단어 벡터(Term Vector) 행렬이고, V는 단어들의 관계를 토대로 구성된 문서 벡터(Document Vector) 행렬이며, Σ는 단일값(Singular Value)들로 구성된 대각행렬이다. SVD 분해 행렬들을 응용하기 위해서는 그림 2.b에서처럼 SVD 분해 행렬들에서 k 개의 속성들을 선택하여 문제에 적용하게 되는데 이와 같은 방법을 Truncated-SVD라 한다. Truncated-SVD를 사용하게 되면 객체의 속성 즉 차원을 축소하는 효과를 보인다. 실제 SVD를 통해서 얻어진 U,Σ,V를 그림 2.a처럼 m x n 행렬로 복

원해도 원본 행렬 A로 복원되지는 않는다. 다만 A의 근사행렬로 복원되며 U,Σ,V 각각에서 k를 많이 취할수록 원본 행렬 A에 더 근접해질 뿐이다. 대부분 목적에 따라 U 혹은 V 행렬을 U_k, V_k로 k 만큼 잘라서 사용하는데 즉 k개의 속성을 가진 객체로 표현된다. SVD를 적용하기 이전 벡터의 경우에는 식 (9), 식 (10)을 사용하여 SVD 벡터 공간으로 변환해 주어야 한다. 또한 용도에 따라서 U_k,V_k 각각을 식 (11), 식 (12)처럼 사용할 수 있다. 정보검색 분야에서 SVD를 사용하는 방법으로는 다음과 같다[4][8].

2.2.1 입력 단어들의 SVD 벡터공간 변환

$$q' = q^T U\Sigma^{-1} \tag{9}$$

주어진 검색어들을 질의벡터 q로 구성하고 식 (9)를 적용하게되면 SVD 벡터공간으로 변환된 Pseudo-Document 벡터 q′를 생성할 수 있다.

2.2.2 입력 문서들의 SVD 벡터공간 변환

$$d' = d^T V\Sigma^{-1} \tag{10}$$

특정단어를 포함하는 문서벡터 d를 구성하고 식 (10)을 적용하게되면 SVD 벡터공간으로 변환된 Pseudo-Term 벡터 d′를 생성할 수 있다.

2.2.3 단어간의 비교

$$\hat{A}\hat{A}^T = U\Sigma^2 U^T \tag{11}$$

단어간의 유사도를 측정하기 위해서 단어벡터 U를 사용하게 되는데 식 (11)를 적용하게 되면 단어들의 유사도 행렬(Similarity Matrix)를 구할 수 있다.

2.2.4 문서간의 비교

$$\hat{A}^T\hat{A} = V\Sigma^2 V^T \tag{12}$$

문서간의 유사도를 측정하기 위해서 문서벡터 V를 사용하게 되는데 식 (12)를 적용하게되면 문서들의 유사도 행렬을 구할 수 있다.

2.3 군집화 알고리즘

본 연구는 군집화 알고리즘의 후처리 작업으로 군집 결과를 분석하여 군집의 대표속성을 추출하는 방법이다. 사용되는 군집화 알고리즘은 K-Means로서, K-Means 알고리즘은 각 자료와 각 군집 중심과의 거리를 고려한 유사도 측정에 기초한 목적함수의 최적화 방식을 사용한다. K-Means는 지역적 최소해에 잘 빠지며, 잡음에 민감하다는 단점이 있으나 수행속도가 빠르므로 광범위하게 이용되고 있다. 본 연구에서는 잡음제거, 차원축소의 이점을 얻기 위해서 SVD를 사용하여 단어-문서 행렬을 분해하고 얻어진 분해행렬들 중 SVD 문서 벡터를 K-Means 군집화 알고리즘으로 처리하여 문서들을 군집화한다[10].

K-Means 알고리즘의 수행절차는 다음과 같다.

단계 1. k개만큼의 군집 중심점을 입력 자료에서 임의적으로 선택한다.

단계 2. n번 반복하면서 입력 자료들과 군집 중심점과의 거리를 계산하여 가장 가까운 군집으로 입력 자료를 분류한다. 일반적으로 입력 자료 x와 군집 중심 m과의 거리는

$$dist(x,m) = \|x-m\|^2$$

로 표현된다[11][12]. 이와 같은 방법으로 입력 자료 x와 k개의 군집 중심점 m과의 거리를 계산하여 가장 가까운 군집으로 입력 자료 x를 대입한다.

단계 3. 만일 식 (13)의 종료 조건 값이 더 이상 변하지 않을 경우 알고리즘을 종료하며 그렇지 않을 경우 단계 2로 간다. 식 (13)의 종료조건은 전체 자료를 대상으로 i번째 군집 C_i에 포함된 객체의 중심 m_i와 x의 거리를 구하고 모든 거리의 합이 최소화됨을 보이며 그 값이 수렴하는 시점이 종료조건이 된다.

$$\min \sum_{i=1}^{k} \Sigma_{x\in C_i} \|x-m_i\|^2 \tag{13}$$

2.4 협력적 여과기법

협력적 여과기법은 특정 고객의 상품에 대한 선호도를 예측하기 위하여 일반적으로 피어슨 상관 계수를 이용하여 유사한 선호도를 가지는 이웃들(Neighborhood)을 결정하고 예측(Prediction)과 상품 추천(Recommendation)을 한다[13].

협력적 여과 기법의 구분은 입력 자료 집합을 처리하는 방법에 따라 메모리 기반 협력적 여과 기법(Memory-based Collaborative Filtering Method)과 모델 기반 협력적 여과 기법(Model-based Collaborative Filtering Method)으로 구분한다[14]. 메모리 기반 협력적 여과 기법은 전체 자료 집합을 입력으로 하여 추천하는 방식을 말하며 모델 기반의 협력적 여과 기법은 전체 자료 집합에서 필요한 정보를 추출하여 사용하는 차이를 보인다. 또한 협력적 여과 기법은 사용되는 유사도 함수의 측정 대상에 따라 고객 기반 협력적 여과 기법(User-based Collaborative Filtering)과 항목 기반 협력적 여과 기법(Item-based Collaborative Filtering)으로 구분된다[14][15]. 고객 기반 협력적 여과 기법은 비교 대상이 고객이 되며 항목 기반 협력적 여과 기법은 비교 대상이 항목 즉 상품이 되며 고객이 구입한 상품과 가장 유사한 비구입 항목을 찾아 추천한다.

모델 기반 협력적 여과 기법으로 연구되는 알고리즘으로는 군집화 알고리즘과 연관규칙 탐사 알고리즘이 있다. 군집화 알고리즘을 사용하는 방법은 고객 기반 협력적 여과 기법에서 사용되는 K-Nearest Neighborhood 방법의 확장이라 할 수 있는데 먼저 유사 고객군을 형성하고 예측 및 추천에 필요한 정보를 모델로 구축한 후 입력 고객이 포함되는 군집의 모델 정보를 사용하여 추천한다. 연관 규칙 탐사 알고리즘을 사용하는 방법은 고객 기반 협력적 여과 기법의 확장으로 전체 자료를 바탕으로 상품간의 연관 규칙을 발견하여 모델을 구축하고 구축된 연관 규칙 모델을 사용하여 추천한다

3. 연구 내용

3.1 연구개요

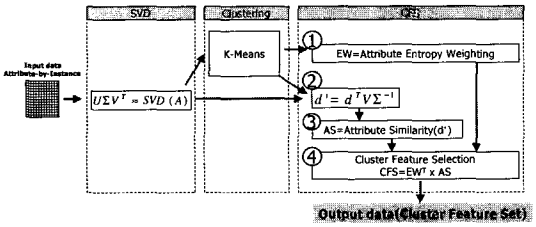

그림 3 군집 특징 선택 구조도

본 연구에서 제안하는 ENTROPY-SVD는 군집화 알고리즘의 후처리 단계에서 군집 결과를 바탕으로 군집 내 대표 속성들을 발견하는 알고리즘이다. 세부적으로 ENTROPY-SVD 과정을 기술하면 다음과 같다.

과정1 : 형성된 군집결과를 바탕으로 속성의 엔트로피 가중치를 측정한다(Attribute Entropy Weighting).
과정2 : 군집 내 유사문서 군을 Pseudo-Term 벡터로 변환한다($d' = d^T V \Sigma^{-1}$).
과정3 : 형성된 군집 내 객체들을 대상으로 개념적으로 유사한 혹은 근접한 속성을 발견한다(Attribute Similarity).
과정4 : 과정1과 3에 의해 얻어진 정보를 병합하여 군집의 속성 가중치를 결정한다(Cluster Feature Selection).

그러나 이러한 단계를 수행하는 과정 중 두 번째 단계에서 속성의 의미론적인 가중치를 산출하기 위해서 SVD 기법을 적용하고 있기 때문에 ENTROPY-SVD의 전 단계 중에서 SVD에 의한 입력 자료의 변환된 loss정보가 구성되어 있어야 한다. 따라서 ENTROPY-SVD 단계 이전에 군집화의 전처리 단계로서 SVD를 이용한 입력자료의 변환($U\Sigma V^T \approx SVD(A)$)과 차원축약이 필수적으로 이루어져야 한다.

다음은 ENTROPY-SVD의 과정별 세부 설명과 간단한 실험 내용이다. 입력 자료는 M과 C의 두 클래스 영어 단문 9문장을 사용하였으며 SVD를 통해 입력 행렬을 변환한 후 변환 된 SVD 문서벡터를 군집화 알고리즘을 통해 군집화하였다. 군집 수를 2개로 정하고 K-Means 알고리즘을 수행하여 표 1과 같은 결과를 얻었다.

표 1 실험 : 문서 군집 결과

Cluster Document	0	문서내용
M1	6	The generation of random, binary unordered trees
M2	7	The intersection graph of paths in trees
M3	8	Graph minors IV : Widths of trees and well-quasi-ordering
M4	9	Graph minors : A survey
Cluster Document	**1**	
C1	1	Human machine interface for Lab ABC computer applications
C2	2	A survey of user opinion of computer system response time
C3	3	The EPS user interface management system
C4	4	System and human system engineering testing of EPS
C5	5	Relation of user-perceived response time to error measurement

3.2 속성의 엔트로피 가중치 측정

군집 내 각 속성 가중치를 결정하기 위해서 일반적으로 정보이론이 사용된다. 속성의 엔트로피 가중치(Attribute Entropy Weighting) 방법은 속성의 빈도수를 기반으로 군집 내 응집도가 높은 속성을 발견하여 가중치를 높게 부여하는 방법이다[7]. 기본 구조는 정보검색의 가중치 기법인 TFIDF의 형태로 지역적 빈도 정보(TF)와 엔트로피를 사용한 전역적 빈도정보(IDF)와의 결합을 통해서 군집간에 공통으로 분포하는 속성의 가중치를 낮추고, 특정 군집에만 포함된 속성의 가중치를 높이는 효과를 가져온다. 이러한 효과는 군집 특징 선택에 있어서 중요한 여과방식을 제공한다. 즉 전체 군집에 균등하게 분포된 속성의 가중치를 낮게 책정함으로써 최종 군집 특징 선택에서 발견 가능성을 최소화한다.

엔트로피 가중치 기법의 세부 과정은 다음과 같다. 먼저 군집 내 각 속성의 빈도수를 계산한다. 즉 군집 내 객체들의 속성정보는 존재여부로 표현되기 때문에 군집 k 안의 속성 a의 빈도수를 계산할 수 있다.

다음단계는 군집 k에 존재하는 각 속성 i의 가중치를 엔트로피 가중치 식 (14)에 적용하여 최종 가중치를 부여한다.

$$a_{ik} = \log(f_{ik} + 1.0) * (1 + \frac{1}{\log(N)} \sum_{j=1}^{N} [\frac{f_{ij}}{n_i} \log(\frac{f_{ij}}{n_i})]) \qquad (14)$$

a_{ik} : 군집 k 내 속성 i의 가중치
f_{ik} : 군집 k 내 속성 i의 빈도수
N : 군집 개수
n_i : 속성 i의 전역적 빈도수

표 2 군집 내 속성 빈도수 정보

cluster 0	0	0	0	0	0	0	0	0	1	3	3	2
cluster 1	2	2	2	3	4	2	2	2	1	0	0	0

표 3 속성의 엔트로피 가중치 정보

cluster 0	0	0	0	0	0	0	0	0	0	1	0.861353116	0.861353116	0.682606194
cluster 1	0.682606194	0.682606194	0.682606194	0.861353116	1	0.682606194	0.682606194	0.682606194	1	0	0	0

표 4 군집과 속성의 유사도 정보

cluster 0	0.000	0.209	0.209	0.180	0.084	0.293	0.293	0.016	0.601	0.775	0.634	0.822
cluster 1	0.892	0.965	0.906	0.644	0.433	0.783	0.783	0.764	0.477	0.099	0.114	0.119

식 (14)는 크게 두 부분으로 구성되어 있는데 지역적 빈도 정보와 전역적 빈도 정보를 사용하여 속성의 엔트로피 가중치를 계산한다. 전역적 빈도 정보를 반영하기 위해서 균등분포 속성의 가중치를 "-1"에 근접하게 만들고 한 군집 속에만 속성이 분포될 경우 "0"에 근접한 값을 가지게 한다. 이러한 특성을 사용하여 군집의 대표 속성을 결정하게 된다.

표 2는 표 1의 군집화 결과를 사용하여 군집 내 단어의 빈도수를 계산한 내용이며, 표 3은 표 2의 빈도수를 사용하여 군집 내 속성의 엔트로피 가중치를 계산한 내용이다. 표 2의 내용을 살펴보면 단어 "survey"가 두 군집에 균등하게 분포되어 있고, 표 3에서 속성의 엔트로피 가중치에 의해 제거된 것을 볼 수 있다.

3.3 SVD 변환

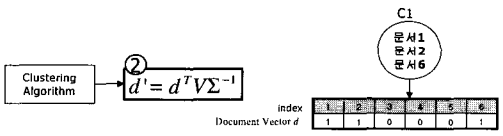

그림 4 유사 객체군의 SVD 벡터로의 변환

군집과 속성간의 유사도를 측정하기 위해서 먼저 군집 내 유사 객체군을 Pseudo-Term 벡터로 변환하는 과정이 필요하다. 그림 4에서처럼 C1 군집에 3개의 문서 "문서1","문서2","문서6"이 있을 때 문서벡터 d를 만들 수 있다. 여기에서 문서벡터 d를 식(10)을 사용하여 Pseudo-Term 벡터 d'를 생성한다.

3.4 속성의 유사도 측정

속성의 존재여부 만을 사용하는 속성 빈도 정보 기반의 단점은 속성을 포함하는 객체들의 구조적인 관계 정보를 제공하지 못하고 있다. 이러한 문제를 해결하기 위해서 LSI/LSA(Latent Semantic Indexing/Latent Semantic Analysis) 방법을 사용하여 객체들의 구조적 정보에 내포된 속성의 개념적 정보를 획득한다. 군집화 알고리즘을 수행한 후 얻어진 정보는 유사 객체군이다. 따라서 군집 내 객체들은 공통 속성들을 서로 공유하게 되는데 객체간의 잠정적인 구조 정보를 획득하기 위해서 LSI/LSA를 사용한다[4][5][8].

다음은 문서집합을 예로 속성의 유사도를 측정하는 방법을 설명한다. 군집화 알고리즘에 의해 유사 객체군 혹은 유사 문서 집합 정보를 획득할 수 있다. 이러한 객체군집 정보를 사용하여 객체군(즉, 유사 문서 집합)과 가장 근접한 속성(즉, 단어)을 추출할 수 있다. 변환된 유사 객체군의 Pseudo-Term 벡터 d'를 사용하여 SVD 문서 벡터 공간상에서 근접한 유사 속성들을 발견하기 위해서 Pseudo-Term 벡터 d'와 SVD 단어 벡터 U 간의 유사도를 측정한다. 유사도 측정은 Cosine, Euclidean Distance, Jaccard와 같은 함수를 사용하여 측정할 수 있다[6].

표 4는 표 1의 군집화 결과 클래스 M의 문서들 d={ M1, M2, M3, M4 }를 Pseudo-Term 벡터로 변환하고 변환된 벡터와 SVD 단어 벡터간의 유사도 함수를 사용해서 측정한 내용이다.

3.4 군집 특징 선택

군집 특징 선택 방법은 이전 단계에서 얻어진 속성의 엔트로피 가중치와 군집과 속성의 유사도(Attribute Similarity)의 각 속성들의 값을 곱한 후 임계치 이상의 값을 취하여 결정한다. 즉 식 (15)와 같이 i 번째 군집의 속성 j의 대표 속성의 최종 값 FS_{ij}는 속성의 엔트로피 가중치 EW_{ij}와 속성과 군집과의 유사도 AS_{ij}의 곱을 사용하여 결정한다.

$$FS_{ij} = EW_{ij} \times AS_{ij} \quad (15)$$

표 1과 같이 두 개의 군집으로 분류된 내용으로부터

표 5 군집 특징 선택 결과

cluster 0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.667	0.546	0.561
cluster 1	0.609	0.659	0.618	0.555	0.433	0.534	0.534	0.522	0.000	0.000	0.000

군집 특징 선택을 한 결과 표 5의 내용처럼 먼저 엔트로피 가중치 기법에 의해 "survey"와 같은 균등 분포 단어들이 제거되었으며 "군집 1"의 중요한 속성으로 "human", "interface", "computer"를 선택할 수 있게 되었다. 표 2와 표 3의 내용에서 실제 엔트로피 가중치 방법에 의해 가장 높은 값을 가지는 단어는 "system", "user" 와 같은 단어들이지만, 표 4의 내용에서처럼 군집과 속성의 유사도 측정에서 낮은 값이 취해지므로 최종 군집 특징 선택에서는 순위가 바뀐 표 5의 내용으로 나타났다.

4. 실험 결과 및 분석

본 연구에서는 후처리 단계에서의 ENTROPY-SVD 군집 특징선택 방법의 평가를 위해서 군집 대표 속성 집합을 사용한 소속 군집 판정 정확도 측정과 모델 기반 협력적 여과 기법의 CFS-CF 추천 시스템을 통하여 평가하였다. 소속 군집 판정 정확도 측정은 군집의 대표 속성 집합에 가중치를 높게 주었을 경우 군집화의 군집 중심점과 같은 역할을 할 수 있음을 보이는 내용이며 CFS-CF와 같은 추천 시스템에서는 상품을 속성으로 인식하는 자료에서 선택된 군집의 대표 속성이 가치가 있음을 증명하는 내용이다.

본 연구에서 사용되는 실험 자료는 EachMovie 자료로서 고객이 본 영화목록과 시청 영화에 대한 점수(Score)가 포함된 자료이다. 대부분의 추천 시스템에서 사용되는 평가 방법을 살펴보면, 추천의 정확도를 평가하기 위해서 고객 시청영화 목록의 일정 부분을 취하여 입력자료로 사용하고 나머지 부분은 추천 시스템의 예측 및 추천의 정확도를 판별하기 위한 테스트 자료로서 활용한다. 예측 정확도는 실제 점수 R_i와 예측 값 P_i값을 통해서 식 (16)의 MAE(Mean Absolute Error)로 판별하고 추천의 정확도는 식 (17)의 재현율(Recall), 식 (18)의 정확율(Precision)을 통해서 평가한다. 식 (19), F1 측정방법은 시스템에 따라 정확율과 재현율이 차이를 보이기 때문에 두 평가치의 가치를 하나의 측정값으로 나타내기 위한 방안으로 사용된다. F1값이 "1"에 근접해지면 정확율과 재현율의 값이 모두 좋아짐을 나타내며 "0"에 근접해지면 정확율과 재현율의 한쪽이 나빠짐을 나타낸다.

$$MAE = \frac{\sum_{i=1}^{N}|R_i - P_i|}{N} \tag{16}$$

$$Recall = \frac{size\ of\ hit\ set}{size\ of\ test\ set} = \frac{|test \cap topN|}{|test|} \tag{17}$$

$$Precision = \frac{size\ of\ hit\ set}{size\ of\ topN\ set} = \frac{|test \cap topN|}{N} \tag{18}$$

식 (17)과 식 (18)에서 사용된 각 변수의 의미는 다음과 같다.

- hit set : 고객이 제공한 0.8 이상의 점수를 가진 상품과 추천된 Top N 상품의 교집합
- topN set : 추천 시스템에서 예측 값이 높은 N 개의 추천 상품 집합
- test set : 테스트를 위한 고객의 상품 집합

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{19}$$

다음은 본 연구에서 제안한 ENTROPY-SVD에 의해 획득된 군집 대표 속성들의 효용성을 평가하기 위한 CFS-CF의 세부 내용이다(그림 5). CFS-CF은 크게 모델 구축과 추천 시스템으로 구분된다.

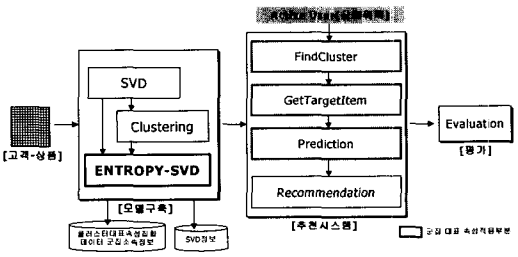

그림 5 평가 시스템 구조도

모델 구축 부분에서는 고객들과 상품 구입목록들로 표현된 입력자료의 전처리를 위하여 SVD 처리를 하고 변환된 자료를 이용하여 군집화를 수행한다. 군집화 결과인 입력 자료의 클래스 정보와 SVD 처리로 변형된 자료를 바탕으로 각 군집의 대표 속성을 결정하고 SVD에 의해 변환된 자료와 입력자료의 클래스 정보, 군집 중심점과 같은 군집 모델, 군집 대표 속성들을 데이터베이스화한다.

추천 시스템 부분에서는 이러한 자료들을 이용하여 고

객의 구입 상품 목록이 입력으로 주어졌을 때 고객의 구입 가능성이 높은 상품을 추천한다. 추천 시스템의 각 단계를 세부적으로 살펴보면 입력 고객의 소속군집발견(FindCluster), 목표 항목 선정(GetTargetItem), 예측(Prediction), 추천(Recommendation)의 순서로 진행된다.

소속군집의 발견은 입력고객의 구입 상품 목록을 사용하여 고객이 어느 군집에 소속되는지 판단하는 부분을 말하며 모델 기반 협력적 여과 기법의 경우 군집 중심점을 통해서 고객의 군집을 결정한다. 군집 중심점을 표현하는 각 벡터의 성분은 SVD 벡터 공간상에 존재하는 임의의 성분이 되기 때문에 군집 중심점을 통해서 원본 자료 속성의 직접적인 해석은 불가능하다. 다만 군집 중심점을 이용한 입력 고객의 소속 군집을 결정할 수 있을 뿐이다. 따라서 원본 자료 속성의 직접적인 해석을 가능하게 하기 위해서 SVD를 통한 군집화 방법의 후처리 단계에서 속성의 엔트로피 가중치와 SVD 정보를 기반으로 속성과 군집의 유사도를 측정하여 군집 특징 선택을 한다. 군집 특징 선택에 의해 선택된 군집 대표 속성 집합의 효용성을 평가하기 위해서 SVD 벡터 성분으로 구성된 군집 중심점을 사용하지 않고 선택된 군집 대표 속성 가중치를 기반으로 입력 자료의 군집을 결정하였다. 발견된 군집의 대표 속성 집합의 가치를 평가하기 위한 실험방법으로 입력 자료의 소속 군집 발견 정확도 측정하였으며 이를 통해서 ENTROPY-SVD와 기존의 특징 선택 기법들과 성능을 비교하였다.

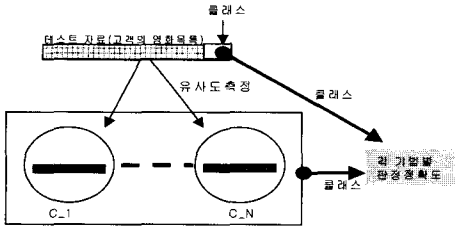

그림 6 군집 대표 속성 집합을 이용한 소속 군집 선택

그림 6에서처럼 임의의 클래스가 주어진 입력 자료를 소속 군집 발견 정확도를 측정하기 위한 입력 자료로 사용하고 특징 선택 기법에 의해 선택된 군집 대표 속성 집합을 군집의 중심점 대신 사용하여 주어진 클래스로 재분류되는 정확도를 측정하였다. 실험에 사용된 특징 선택 기법들로는 ENTROPY-SVD, 엔트로피 가중치(ENTROPY), Mutual Information(MI), x^2 Statistic (CHI)을 사용하였다. 이러한 특징 선택 기법은 각각의 클래스에 포함된 속성과 클래스의 상관관계 혹은 클래스 내에서 속성의 중요도 가치를 가중치로 나타낼 수 있다. 표 6은 여러 특징 선택 기법들과 ENTROPY-SVD의 성능 비교 실험 결과이다. ENTROPY-SVD 방법이 다른 방법에 비해 재 분류율이 높았다.

표 6 군집 소속 판정 정확도

군집소속판정 정확도			
ENTROPY-SVD	ENTROPY	MI	CHI
73.46%	70.98%	68.92%	31.92%

목표 항목(Target Item)의 선정은 추천 시스템 내부에서 고객(Active User)의 비구입 상품(Target Item) 목록을 작성하는 단계로서 군집 대표 속성 집합 기준으로 고객 구입 상품 목록을 뺀 집합을 사용하여 목표 항목으로 선정한다.

예측은 목표 항목이 선정되면 고객이 상품을 구입할 것인지 혹은 얼마나 상품을 선호할 것인지 판단하는 부분 상품들의 선호도 계산 부분이다.

다음은 SVD 속성 선택 범위의 변화에 따라 추천 정확도가 변화되는 내용을 보인다. 즉 SVD 속성 범위가 증가한다는 것은 군집화의 입력으로 차원 정보를 충분히 제공하는 상태가 된다. 따라서 군집화 정확도가 향상되며 후처리 단계에서의 군집 특징 선택 기법의 결과가 좋아지게 된다. 그림 7은 입력 고객의 군집을 발견하고 해당 군집에서 10명의 가장 근접한 이웃을 찾아 추천한 결과이다. 그림 7의 F1 측정값의 변화를 살펴보면 SVD의 속성 벡터가 증가함에 따라 F1 측정값이 좋아짐을 보였다. 결과적으로 군집 특징 선택 기법의 질은 군집화 결과의 질에 의해 결정됨을 알 수 있다.

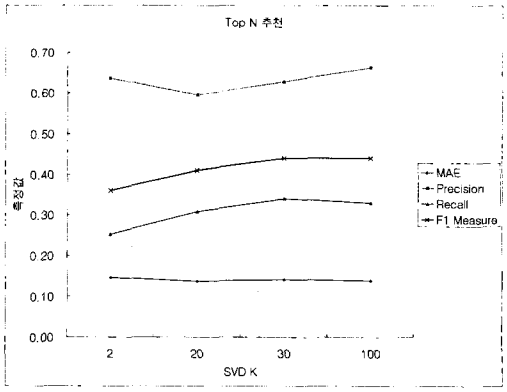

그림 7 소속 군집 내 K-Nearest Neighborhood의 Top

N 추천

표 7은 기존 모델 기반의 협력적 여과 기법(Original-CF)의 추천 시스템과 본 연구에서 군집 대표 속성 집합을 사용하는 추천 시스템 CFS-CF의 성능을 비교한 실험 결과이다. 본 기법에 의해 선택된 군집 대표 속성을 기반으로 추천항목을 정한 추천 시스템의 성능이 기존의 방법 보다 향상된 결과를 보였다.

표 7 추천 시스템 성능 비교

예측평가		
평가방법	Original-CF	CFS-CF
MAE	0.213436	0.146278405
추천평가		
평가방법	Original-CF	CFS-CF
Top N 평균 Precision	66.00%	73.68%
Top N 평균 Recall	32.24%	37.05%
Top N 평균 F1 Measure	39.23%	46.00%

5. 결론 및 향후연구

본 연구에서는 군집화 알고리즘 결과를 통해서 군집에서 대표할 수 있는 속성들을 발견하는 기법으로 군집 특징 선택 기법 ENTROPY-SVD를 제안한다.

ENTROPY-SVD는 엔트로피 기반의 가중치 기법과 SVD를 사용하여 군집내 응집도가 높은 속성과 군집과 속성의 유사성이 높은 속성들을 발견하는 특징 선택 기법이다. ENTROPY-SVD의 효용성을 평가하기 위하여 군집의 대표 속성 집합을 이용한 추천 시스템, CFS-CF를 구현하였다. ENTROPY-SVD에 의해 선택된 군집의 대표 속성을 유사 고객군의 대표 상품으로 인식하고 이러한 상품을 대상으로 추천 시스템은 상품을 추천한다. 상품 추천에 있어서 ENTROPY-SVD 특징 선택에 의한 상품 추천 시스템 CFS-CF의 추천 정확도가 기존 추천 시스템(Original-CF)에 비해 높았으며 또한 군집의 성향 분석을 통한 추천 시스템으로도 발전할 수 있는 특성을 가지고 있었다.

또한, ENTROPY-SVD에 의해 선택된 군집의 대표 속성을 사용하여 입력 고객의 군집 소속 판정 정확도 실험을 수행하였으며 그 결과 선택된 대표 속성의 가중치가 각 군집의 중심점 역할을 수행할 수 있는 프로파일로 효용성이 있음을 발견하였다.

참 고 문 헌

[1] Yang, Y., Pedersen, J.O., A Comparative Study on Feature Selection in Text Categorization, Proc. of the 14th International Conference on Machine Learning ICML97, pp. 412-420, 1997.

[2] Joachims, T., A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, Proc. of the 14th International Conference on Machine Learning ICML97, pp. 143-151, 1997.

[3] Lewis, D. D., Feature selection and feature extraction for text categorization, Proceedings of Speech and Natural Language Workshop, pp. 212-217, 1992.

[4] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R., Indexing by latent semantic analysis, Journal of the American Society for Information Science, 41(6), pp. 391-407, 1990.

[5] Berry, M. W., Dumais, S. T., and O'Brien, G. W., Using linear algebra for intelligent information retrieval, SIAM Review, 37(4), pp. 573-595. 1995.

[6] Kolda, T. G. and O'Leary, D. P., A semidiscrete matrix decomposition for latent semantic indexing in information retrieval, ACM Trans. Inf. Syst., 16, pp. 322-346, 1998.

[7] M.W. Berry, Z. Drmac, and E.R. Jessup, Matrices, vector spaces, and information retrieval, SIAM Rev., 41(2), pp. 335-362, 1999.

[8] Landauer, T. K., Foltz, P. W., and Laham, D., An introduction to Latent Semantic Analysis, In Discourse Processes 25, pp. 259-284, 1998.

[9] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J., Application of Dimensionality Reduction in Recommender System-A Case Study, In ACM WebKDD 2000 Web Mining for E-Commerce Workshop, 2000.

[10] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim., ROCK: a robust clustering algorithm for categorical attributes, In Information Systems, 25(5), pp. 345-366, 2000.

[11] Strehl, A., Ghosh and J., Mooney, R., Impact of similarity measures on web-page clustering, In Proc. AAAI Workshop on AI for Web Search, pp. 58-64, 2000.

[12] M. Devaney and A. Ram., Efficient feature selection in conceptual clustering, In Machine Learning: Proceedings of the Fourteenth International Conference, pp. 92-97, Nashville, TN, 1997.

[13] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom and John Riedl, GroupLens: an open architecture for collaborative filtering of netnews, Proceedings of the conference on Computer supported cooperative work, pp. 22-26, October 1994.

[14] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J., Item-based Collaborative Filtering Recommender Algorithms, In WWW10 Conference, pp. 285-295, May 2001.
[15] D. Billsus and M. J. Pazzani, Learning collaborative information filters, In Proceedings of the Fifteenth International Conference on Machine Learning, pp. 46-54, July 1998.
[16] Sonny HS Chee, RecTree: A Linear Collaborative Filtering Algorithm, M.S thesis, Computing Science, Simon Fraser University, 2000.

이 영 석
1998년 원광대학교 컴퓨터공학과 학사. 2001년 숭실대학교 컴퓨터학과 석사. 관심분야는 Data Mining, CRM, Planning, Machine Learning, AI

이 수 원
1992년 서울대학교 자연과학대학 계산통계학과 학사. 1984년 한국과학기술원 전산학과 석사. 1984년 ~ 1987년 LG 중앙연구소 주임 연구원. 1994년 University of Southern California 전산학과 박사. 1995년 ~ 현재 숭실대학교 컴퓨터학부 조교수. 관심분야는 Data Mining, CRM Agent, Machine Learning, Expert System, AI