인터넷 상점에서의 내용기반 추천을 위한 상품 및 고객의 자질 추출 성능 비교

안 형 준[†] · 김 종 우[††]

요 약

인터넷 쇼핑몰에서의 상품 추천을 위해 널리 사용되는 방식 중 한 가지는 상품의 특성과 고객의 특성을 비교하여 고객에 맞는 상품을 추천하는 방식이다. 이 방식은 상품이나 고객의 특성을 표현하는 자질(Feature)의 개수가 많을수록 그 중에 어떤 자질을 선택해야 더 좋은 추천 성과를 가져올 수 있는지 파악해 내는 것이 추천의 효과 및 효율성 측면에서 중요하지만 아직까지 충분히 연구되지 않은 실정이다. 본 연구에서는 인터넷 서점에서의 가상 구매실험을 바탕으로 사용자가 구매한 책 들에서 사용자를 잘 나타낼 수 있는 자질을 선택하는 방식에 대해서 벡터 스페이스 모형, TFIDF(Term Frequency-Inverse Document Frequency), Mutual Information, SVD(Singular Value Decomposition) 방식 등을 활용하여 실험하고 그 결과를 비교해본다. 실험 결과 SVD를 응용한 자질 추출 기법이 가장 좋은 성능을 나타내었다.

키워드 : 상품 추천, 내용 기반 필터링, SVD, 자질 선택

Comparison of Product and Customer Feature Selection Methods for Content-based Recommendation in Internet Storefronts

Hyung Jun Ahn[†] · Jong Woo Kim[††]

ABSTRACT

One of the widely used methods for product recommendation in Internet storefronts is matching product features against target customer profiles. When using this method, it's very important to choose a suitable subset of features for recommendation efficiency and performance, which, however, has not been rigorously researched so far. In this paper, we utilize a dataset collected from a virtual shopping experiment in a Korean Internet book shopping mall to compare several popular methods from other disciplines for selecting features for product recommendation: the vector-space model, TFIDF(Term Frequency-Inverse Document Frequency), the mutual information method, and the singular value decomposition(SVD). The application of SVD showed the best performance in the analysis results.

Key Words : Prouct Recommendation, Content-based Filtering, SVD, Feature Selection

1. 서 론

인터넷 상점에서의 상품이나 서비스를 추천하기 위해서 많은 연구가 이루어져왔다. 이러한 연구들은 크게 사용자들 간의 유사도를 활용한 협력적 필터링(Collaborative Filtering) 방식과 상품 및 사용자의 특성 값들을 활용하는 내용기반 필터링(Content-based Filtering) 방식으로 구분될 수 있다[1-6]. 협력적 필터링 방식은 많은 연구에서 성공적인 결과를 보여주는 것으로 제시되고 있으나 협력적 필터링에 사용되는 사용자들의 평가 정보를 얻기 힘든 경우가 많고, 또한 평가 정보의 수가 적을 경우 추천 성능이 크게 감소하는 희소성(Sparsity)의 문제를 갖고 있다[7-9]. 따라서 이러한 경우 내용 기반 필터링 방식이 현실적인 추천 도구가 될 수 있다.

내용 기반 필터링에서는 상품의 특성과 사용자의 특성이 모델링되고 각 특성들을 연결해 주는 학습 결과나 추천 모델에 따라 상품의 추천이 이루어진다. 상품 및 사용자의 특성은 인구 통계학적 특성, 사용자가 명시한 선호 특성 등 다양한 정보가 존재할 수 있으나 일반적인 상황에서는 등록되지 않은 사용자에 대해서 인구 통계학적 특성을 얻기 힘든 경우가 많고 사용자가 명시한 선호 특성들도 협력적 필터링에서의 평가 정보와 마찬가지로 구하기 힘든 경우가 많다. 따라서 그 보다는 사용자가 구매했거나 사용자가 관심

† 준 회 원 : Senior Lecturer, Waikato Management School, University of Waikato, New Zealand
†† 종신회원 : 한양대학교 경영학부 부교수(교신저자)
논문접수 : 2005년 5월 25일, 심사완료 : 2006년 2월 20일

을 표명한 상품의 특성을 바탕으로 사용자 프로파일을 구성하는 방식이 많이 사용된다[10, 11] .

그러나 비교적 제품의 속성이 간단한 경우를 제외하고는 인터넷 상점에서는 많은 종류의 상품과 서비스가 존재하게 되고, 따라서 다양하고 복잡한 종류의 속성 값을 갖게 된다. 예를 들어, 영화나 서적의 경우 이에 대한 속성은 간단한 장르나 주제 뿐 아니라 상품에 대한 서술적인 묘사나 평론 등도 상품의 속성이 될 수 있다. 영화나 서적과 같은 문화 컨텐츠 상품(contents product)이 아닌 경우에도 상품에 대한 서술적인 묘사는 상품에 대한 많은 정보를 제공해주므로 내용 기반 필터링 방식에서 주요한 자질로 사용될 수 있다.

그러나 위와 같이 텍스트 형태의 상품 묘사를 사용하는 경우, 서술적인 문장들에 포함 된 키워드의 숫자만큼 자질의 수가 많아질 수 있으며 이는 추천의 효과성과 효율성에 모두 문제를 가져올 수 있다. 효과성 측면에서는 부적절한 자질로 인해 추천 성능이 떨어질 수 있으며, 효율성 측면에서는 너무 많은 수의 자질을 유지하고 검색해야 하는 데서 오는 계산 및 시간 부담이 있을 수 있다. 따라서 전체 대상 자질 중에 더 의미 있고 유용한 소수의 자질을 파악 해 내는 것은 매우 중요한 일이다[12-14]. 본 연구에서는 이를 위해 인터넷 서점에서의 가상 쇼핑 실험을 통해 수집한 자료를 활용하여 사용자가 선택한 서적들에 대한 설명 문장들로부터 사용자의 프로파일을 구성해내고 이를 추천 대상 책들의 키워드와 비교하여 추천 실험을 수행하였다. 이때 정보 검색 분야에서 널리 쓰이는 키워드 벡터 스페이스 모형을 활용하며 이를 기반으로 TF(Term Frequency), IDF (Inverse Document Frequency), TFIDF(Term Frequency-Inverse Document Frequency)방식과 Mutual Information을 응용하여 근사치를 활용하는 방법, 그리고 SVD(Singular Vector Decomposition) 방식의 응용 등으로 자질을 선택하여 어떤 방식이 더 나은 추천 성과를 보여주는지 실험하도록 한다. 본 논문의 구성은 다음과 같다. 2장에서는 내용기반 추천 시에 활용가능한 자질 추출 방식들을 검토한다. 3장에서는 실험 내용과 분석 결과를 제시한다. 4장에서는 결론과 추후 연구 과제를 제시하도록 한다.

2. 관련 연구

2.1 상품 묘사를 사용한 자질 추출 방식

자질 추출(feature selection)에 관한 연구는 주로 패턴 매칭이나 기계 학습(machine learning), 텍스트 자동 분류 등의 분야에서 활발히 연구되어 왔다[12-14]. 이러한 연구들에서는 학습 대상 자료의 어떤 자질이 학습 성과에 큰 영향을 끼치는지를 연구해 왔다. 인터넷 상점의 상품 추천 분야에서는 그러나 자질 추출에 관한 연구가 많이 존재하지 않으며 따라서 패턴 매칭, 기계 학습 등의 분야에서 쓰이는 Mutual Information에 의한 방식[13, 14]이나 SVD[15] 등을 상품 추천에 응용 해 볼 수 있다. 그러나 주로 분류의 목적으로 쓰이는 자질 추출 기법들의 경우에는 본 연구의 대상인 인터넷 서점의 서적 추천에 적합하지 않다.

상품에 대한 글로 이루어진 묘사에서 자질을 추출하기 위해서는 정보 검색(Information Retrieval) 분야에서 쓰이는 기법과 도구들이 활용되어야 한다. 보통 일반적인 문장에는 키워드로 적합하지 않은 단어들이 많이 포함되어 있으며. 또한 명사 등 키워드에 적합한 단어에서 조사나 어미를 분리해 내는 작업도 필요하다. 이러한 작업은 형태소 분석기를 거쳐서 이루어지게 되며 작업 결과로 문서나 사용자에 대한 키워드 자질 집합을 갖게 된다. 이러한 키워드 자질 집합은 보통 <w1, w2, w3, …>와 같이 각 키워드에 대한 가중치로 이루어진 벡터로 표현되어진다. 따라서 사용자에 대한 프로파일 벡터와 추천 대상 상품에 대한 벡터의 유사도를 계산하여 더 높은 유사도를 가진 상품을 추천하는 것이 하나의 가능한 추천 방법이 된다. 이때 유사도 계산의 방법으로 두 벡터의 코사인 값을 계산하거나 선형 correlation 값을 계산하는 방법 등이 널리 쓰이고 있다.

2.2 벡터 스페이스(Vector Space) 모형

벡터 스페이스 모형[16]은 정보 검색 분야에서 문서와 질의(Query)를 나타내기 위해 널리 쓰이는 방법이다. 문서나 질의에는 키워드가 등장하며 키워드는 각 문서와 질의에서 서로 다른 중요성을 갖게 된다. <표 1>은 벡터 스페이스 모형으로 표현한 문서와 질의의 예가 나와있다. 만약 키워드가 세 개 <정보, 노트북, 경영>만 존재한다면 문서는 3차원 벡터인 <0.3, 0.7, 0.1>로 표현되고 질의는 <0.5, 0, 0.8>로 표현된다. 이때 각 숫자는 각 단어(자질)가 문서 및 질의에서 갖는 가중치가 된다. 이때 이 문서벡터 d와 질의 q의 유사도를 코사인 값을 이용해 계산하면 다음과 같다.

$$\text{Cosine distance(d, q)} = \frac{d \bullet q}{|d||q|} \quad (1)$$

〈표 1〉 벡터 스페이스 모형의 표현

단어	문서	질의
정보	0.3	0.5
노트북	0.7	0
경영	0.1	0.8

2.3 TFIDF

앞서 <표 1>의 예에서 가중치의 예가 제시되었다. 이러한 가중치는 여러 가지 방식으로 계산될 수 있으나 정보 검색 분야에서 가장 널리 쓰이는 방식은 TFIDF[16] 방식이다. TFIDF 방식은 TF(Term Frequency)와 IDF(Inverse Document Frequency)의 결합 방식이며 이때 TF는 단어의 단순한 출현 빈도수를 나타내고 IDF는 단어의 변별력을 나타낸다. 즉 해당 문서에서 많이 등장하는 단어는 높은 TF 값을 갖게 되고, 전체 문서의 집합에서 적은 수의 문서에만 등장하는 단어는 높은 IDF 값을 갖게 된다. 이때 TF 값과 IDF 값이 둘 다 높을수록 높은 가중치를 주는 방식이 TFIDF 방식이다. TF와 IDF는 다음과 같이 계산된다.

$$TF(t,d)=\frac{Occ(t,d)}{MaxOcc(d)} \quad (2)$$

이때 $Occ(t,d)$는 문서 d에서 단어 t의 등장 횟수, $MaxOcc(d)$는 문서 d에서 가장 많이 등장한 단어의 등장 횟수이다.

$$IDF(t)=\frac{N}{N(t)} \quad (3)$$

이때 N은 전체 문서의 개수, $N(t)$는 단어 t가 등장한 문서의 개수임.

따라서 (2)와 (3)을 결합하여 TFIDF 값은 다음과 같이 계산된다.

$$TFIDF(t,d)=\frac{Occ(t,d)}{MaxOcc(d)}\cdot\frac{N}{N(t)} \quad (4)$$

2.4 Mutual information

Mutual Information(MI)은 Shannon의 정보 이론에서 제시된 도구로 두 개의 확률적 사건(event)이 서로 제공하는 상대 사건에 대한 정보 값이며 두 사건에 대해 대칭적인 값을 갖는다[8, 13]. 즉 MI(a, b)를 a 사건과 b 사건의 Mutual Information이라고 하면 두 개의 사건이 서로의 발생에 대해 주는 정보가 클수록 이 값은 커지게 되며, 이것은 이 값이 클 수록 한 사건의 발생이 다른 사건의 발생을 더 큰 확률로 예측할 수 있음을 의미한다. 결국 이 값이 클 수록 두 사건 a, b가 더 밀접하게 관련되어 있다고 볼 수 있다.

추천을 위한 자질 추출 연구에서는 특정 자질의 존재와 실제 상품의 추천을 두 사건으로 하여 특정 사용자에 대해 어떤 자질과 실제 구매 경력이 서로 관련 있는지를 MI 값을 활용하여 계산할 수 있다. 두 사건 a, b에 대한 MI는 다음과 같이 계산될 수 있다.

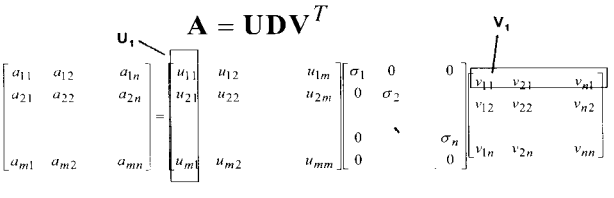
$$MI(a,b)=\log_2\frac{P(ab)}{P(a)P(b)} \quad (5)$$

이때 P(a)는 a의 발생 확률, P(b)는 b의 발생 확률, P(ab)는 a와 b사건이 모두 발생할 확률이다.

2.5 Singular Value Decomposition

Singular Value Decomposition(SVD)는 선형대수학의 행렬 분할 방식 중의 하나로 정보 검색, 패턴 인식 등의 분야에서 원 자료의 차원을 줄이기 위해 많이 사용되는 방식이다. 이 방식의 기본 아이디어는 원 자료의 행렬을 SVD 방식으로 분할했을 때 생기는 singular value들의 값이 차원을 증가시킴에 따라 급격히 줄어든다는 점을 이용한 것이다. 따라서 가장 큰 소수의 singular value와 그에 해당하는 직교 벡터(orthogonal vector)들만 가지고도 원래 행렬에 근접한 행렬을 복구해 낼 수 있다. 따라서 큰 이미지를 압축하거나 대량의 문서 집합의 차원을 줄여서 색인(indexing) 하는 등의 용도로 널리 쓰이고 있다. 특히 정보 검색 분야에서 문서를 색인 하는 경우 키워드들이 가진 잠재적인 의미까지 활용하게 되어 정보 검색의 성능이 좋아지는 것으로 알려져 있으며, 이를 Latent Semantic Indexing(LSI)라고도 부른다[15, 17] .

(그림 1)은 SVD를 통해 행렬을 분해하는 과정을 보여주고 있다. 원 자료 행렬인 A는 SVD 과정을 통해 U, D, V 행렬로 분해되며 이때 D는 singular value들로 이루어진 대각 행렬이며 각 singular value인 σ_i는 i가 클수록 작은 값을 갖게 된다. U와 V는 각각 직교 벡터로 이루어지게 된다. 이때 값이 큰 소수의 singular value만을 선택함으로써 원 자료 행렬의 차원을 크게 줄일 수 있으며 원 자료에 근사한 행렬을 얻을 수 있다. 이때 차원은 벡터 스페이스 모형의 자질의 개수와 정확히 일치하는 것은 아니며, 행렬이 표현하는 벡터 공간의 차원이라고 할 수 있다. SVD의 경우 축소된 차원이나 추출된 자질이 어떤 의미를 갖는지는 명확히 알 수 없는 단점이 존재하지만 그러한 의미의 파악이 목적이 아니고 차원 축소를 통한 패턴 인식 및 추천 등의 용도로 활용할 경우엔 문제가 되지 않는다.



(그림 1) SVD의 도해

3. 실 험

3.1 자료 수집

분석을 위한 실험 자료는 특정 인터넷 서점에서의 가상 구매 실험 및 평가 대상 책들에 대한 실험 참가자들의 5점 척도 평가 점수로 이루어져 있다. 가상 구매 실험은 실험 참가자 140명에게 해당 인터넷 서점 사이트에서 원하는 만큼의 선호하는 책을 구매하여 장바구니에 넣도록 하였다. 가상 구매 실험을 통해 구매한 책들에 관한 기본 통계는 <표 2>와 같다.

〈표 2〉 가상 구매 실험

평균 구매 수량	10.93
표준 편차	7.01
최대 구매자 구매 수량	50
최소 구매자 구매 수량	2

가상 구매 실험을 통해 모아진 자료는 각 실험 대상자들의 사용자 프로파일을 구성하기 위해 사용되었다. 그 다음, 구성된 사용자 프로파일을 사용하여 서적을 추천하고 추천 성과를 비교하기 위해 <표 3>에 나온 16개의 대분류마다

〈표 3〉 추천 대상 서적들의 대 분류

컴퓨터/인터넷	경영/경제	외국어/어학	어린이
취미/건강	만화/애니메이션	소설	시
인문	에세이	고전	사회과학
과학	역사	예술	잡지

각 2권의 책을 선정하여 총 32권의 책에 대하여 사용자들이 1~5 까지의 선호도를 부여하도록 하였다. 이때 5의 선호도가 해당 책에 대한 가장 높은 선호도를 의미한다. 선정된 책들이 사용자가 가상 구매한 책들과 중복되는 것을 피하기 위하여 선정 범위를 16개 대 분류의 신간으로 제한하였으며 또 시험 시점은 가상 구매 실험으로부터 1개월 경과 한 이후로 하였다.

각 서적 별 자료는 서적의 ISBN 혹은 ISSN, 제목, 그리고 책에 대한 설명으로 이루어져 있다. <표 4>는 구매 및 추천 대상 서적 정보의 예이다.

〈표 4〉 구매 및 추천 대상 서적 정보의 예

ISBN	89-7629-317-7
분류	컴퓨터/인터넷
제목	세상에서 가장 쉬운 Flash MX
설명	전반부는 플래시를 통해서 그림을 그리고 애니메이션하는 예시를 많이 보여주고 있다. (후략)

3.2 키워드 추출

내용 기반 추천을 위해, 구매 실험을 통한 서적들과 추천 대상 서적들에서 키워드를 추출해 내었다. 키워드 추출 과정에서는 한글의 형태소 분석 도구를 활용하여 각 서적의 제목과 설명에 등장하는 명사들만을 추출하였다. <표 5>는 실험자 별 그리고 추천 대상 서적 별 추출해 낸 키워드에 관한 기초 통계량을 보여주고 있다. <표 6>은 실제 추출해 낸 키워드의 예를 보여준다.

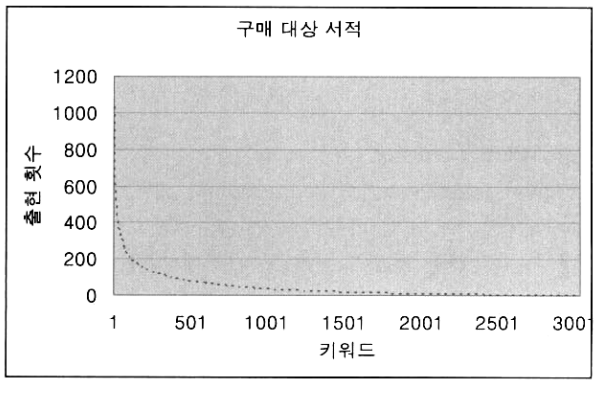
〈표 5〉 추출해 낸 키워드의 기초 통계량

	평균 키워드 수	표준편차	최대 키워드	최소 키워드
실험자 별	551.2	238.8	1374	64
추천 대상 서적 별	150.1	132.5	721	15

〈표 6〉 추출해 낸 키워드의 예 (일부 키워드)

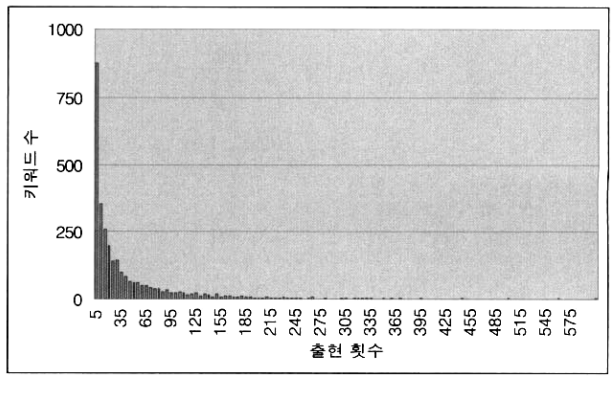
서적	키워드	출현 횟수
베네치아에서 비발디를 추억하며 : 건축가 정태남의 이탈리아 음악여행	음악	24
	이탈리아	22
	도시	16
	여행	13
	유럽	12
	예술	8
	건축가	6
	오페라	5

전체 키워드 중에서 실제로 사용되는 키워드는 실험자의 프로파일과 추천 대상 서적들 양 쪽에서 적어도 한 번 이상 출현한 키워드여야 한다. 이는 한 쪽에서만 출현하는 키워드는 추천할 때에 영향을 끼치지 못하기 때문이다. 이렇게 하여 총 3034개의 키워드가 남겨졌다. 각 키워드는 (그림 2)에서 보이듯이 지수적인 분포를 보이며 구매 실험에서 등장한 서적 들에 나타난다. 즉 매우 흔히 등장하는 단어 군이 왼쪽에 존재하며 가장 우측에는 반대로 매우 소수의 책들에서만 등장하는 키워드가 존재한다. 이때 너무 많이 등장하는 최 우측의 단어군은 변별력이 적으며 너무 적은 출현 빈도를 갖는 가장 우측의 단어군은 잘 쓰이지 않는 키워드로 효용이 떨어질 수 있다.



(그림 2) 키워드를 구매 대상 책에 등장한 횟수에 따라 정렬하고 횟수를 Y축에 표시

반대로 출현 횟수 별 키워드의 개수를 히스토그램의 형태로 보면 (그림 3)과 같다. 여기서는 역시 5개 미만의 서적에서 등장한 키워드가 가장 많고 대부분의 키워드들이 100개 이하의 문서에서 등장함을 확인할 수 있다.



(그림 3) 출현 횟수 별 키워드 개수 (히스토그램)

3.3 실험 방식

자질을 추출하여 추천 성능을 시험하기 위해 5가지 실험이 수행되었다. TF, IDF, TFIDF, MI, SVD 등의 방법이 사용되었으며 각각 키워드나 차원의 개수를 1에서부터 증가시켜 가며 추천 성능을 비교하였다. TF, IDF, TFIDF, MI 방

〈표 7〉 실험 방식의 정리

방식	자질(차원)의 범위	유사도 계산 방식
TF	1~200	코사인 방식
IDF	1~200	코사인 방식
TFIDF	1~200	코사인 방식
MI	1~200	정보(bit) 합산
SVD	1~140	코사인 방식

식에서는 추천 대상 서적의 평균 키워드 개수가 150개인 점을 감안하고 또 실험 목표가 적은 수의 훌륭한 자질을 추출해 내는 데 있으므로 키워드 자질의 개수는 200개 까지만 증가시켜가며 실험하였다. SVD에서는 원 자료 행렬이 140 X 3034의 크기를 가지므로 최대 140개의 singular value가 존재하며 따라서 1에서 140까지의 차원에 대해 실험하였다. TF, IDF, TFIDF, SVD 방식에서는 기본적으로 앞서 설명된 코사인 방식의 유사도 척도를 활용하였으며 MI 방식에서는 상호 정보의 값이 비트(bit)로 계산되므로 이를 각 키워드에 대해 합산하는 방식을 사용하였다.

3.3.1 TF

TF 방식은 가장 간단한 방식으로 각 실험 대상자 별로 프로파일에 가장 많이 등장하는 키워드의 순으로 자질을 추출하는 방식이다. 이때 자질이 추출된 후의 유사도 계산은 원래대로 TFIDF 가중치를 활용하여 코사인 방식으로 계산된다.

3.3.2 IDF

IDF 방식은 단어들의 변별력을 중심으로 자질을 추출하는 방식이다. 따라서 가장 변별력이 높은 단어부터 차례로 자질이 추출되고 이 후에는 TF 방식과 마찬가지로 원래의 TFIDF 가중치 및 코사인 유사도를 활용해 추천이 이루어진다. 이때 단어들의 IDF 값은 사용자들이 구매 선택하였던 모든 책들을 문서집합(corpus)로 하여 계산되어졌다.

3.3.3 TFIDF

TFIDF 방식은 각 사용자 별로 미리 계산되어진 TFIDF 값 중 높은 값을 갖는 단어들을 차례로 자질로 추출하는 방식이다. 이때 역시 TFIDF 가중치와 코사인 유사도가 활용되어졌다.

3.3.4 MI

MI 방식에서는 자질들이 실험자의 구매한 책과 자질을 포함한 전체 책 사이의 연관 관계, 즉 상호 정보를 얼마나 크게 해 주는지를 활용하는 방식이다. 즉 단어 w가 좋은 자질이 되기 위해서는 w를 포함한 책들을 실험자가 많이 구매해야 하며, 반대로 실험자가 구매한 책 들에는 w가 포함된 책 들이 많이 존재해야 한다. 전체 존재하는 책 중에 사용자가 구매한 책의 비율을 P(B)라고 하고, 또 전체 책 중에서 단어 w를 포함한 책의 비율을 P(w)라고 하고, 둘의 교집합을 P(Bw)라고 하면 MI는 다음과 같이 계산될 수 있다.

$$MI(B,w)=\log_2\frac{P(Bw)}{P(B)P(w)} \tag{6}$$

그러나 본 논문의 실험 환경에서는 실험자들에게 노출 된 전체 서적의 개수를 알 수 없으며 따라서 각각의 확률 P(B), P(w), P(Bw)를 정확히 계산하기 힘들다. 따라서 대신 다음과 같은 근사 방식을 사용하였다.

P(w)는 샘플 크기가 충분히 클 경우에 실험에서 관찰한 값 P'(w)로 치환이 가능하다. 실험에서 관찰한 w를 포함한 서적의 개수 N'(w), 실험에서 사용된 서적의 개수 N이라 하면 다음과 같이 근사할 수 있다.

$$P(w)\approx\frac{N'(w)}{N}=P'(w) \tag{7}$$

서점에 존재하는 전체 책의 개수를 N(U)라고 하고 실험자가 전체 책 들에 대해 노출되었을 경우에 선호하는 책으로 구매 한 서적의 개수를 N(B)라고 하면 P(B)는 다음과 같이 구할 수 있다.

$$P(B)=\frac{N(B)}{N(U)} \tag{8}$$

그러나 이때 실험 상황에서는 실험자가 전체 서적에 대해 노출되었다고 가정할 수 없으며 따라서 실제 구매 실험에서 구매한 책의 개수 N'(B)는 아래와 같이 이상적인 상황의 N(B)의 일부라고 생각할 수 있다.

$$N'(B)\approx\frac{N(B)}{k}\quad (\text{이때 } k>1) \tag{9}$$

이렇게 가정을 할 경우 B와 w의 교집합의 크기인 N(Bw)도 다음과 같이 근사되어질 수 있다.

$$N'(Bw)\approx\frac{N(Bw)}{k} \tag{10}$$

따라서 특정한 키워드 w의 MI 값은 아래와 같이 모두 실험에서 관측 가능한 값을 바탕으로 계산된다.

$$MI(B,w)=\log_2\frac{P(Bw)}{P(B)P(w)}\approx\log_2\frac{\frac{N'(Bw)}{N(U)}}{\frac{N'(B)}{N(U)}P'(w)}$$

$$=\log_2\frac{kN'(Bw)}{kN'(B)\frac{N'(w)}{N}}=\log_2\frac{N\cdot N'(Bw)}{N'(B)N'(w)} \tag{11}$$

개별 키워드에 대한 MI값은 위처럼 비트 단위의 정보로 계산되며 이를 실험자 프로파일과 추천대상 서적 프로파일

$$\begin{bmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ & & & \\ a_{m1} & a_{m2} & & a_{mn} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & & u_{1m} \\ u_{21} & u_{22} & & u_{2m} \\ & & & \\ u_{m1} & u_{m2} & & u_{mm} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & & 0 \\ 0 & \sigma_2 & & \\ & & & \\ 0 & & & \sigma_n \\ 0 & & & 0 \end{bmatrix} \begin{bmatrix} v_{11} & v_{21} & & v_{n1} \\ v_{12} & v_{22} & & v_{n2} \\ & & & \\ v_{1n} & v_{2n} & & v_{nn} \end{bmatrix}$$

3034 X 140 = 3034 X m × m X m × m X 140

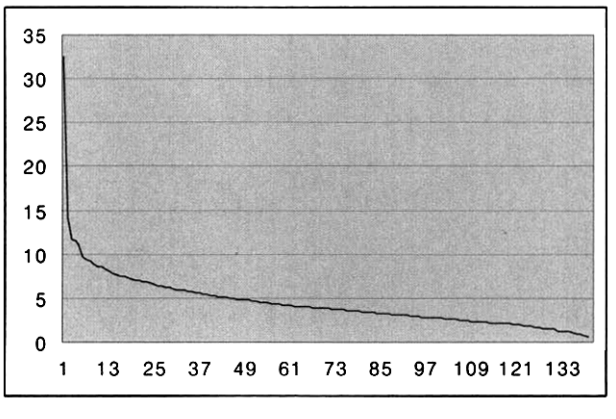
(그림 4) SVD 분할 (이때 m은 차원으로 원 행렬의 차원이 140을 넘을 수 없으므로 m이 140 이상의 경우에는 해당 차원의 singular value는 0이 된다.)

에 공통된 키워드의 집합에 대해 합산하여 유사도로 활용하였다.

3.3.5 SVD

SVD를 활용하기 위한 행렬은 (키워드 수 X 실험자 수)의 크기를 갖는다. 따라서 3034 X 140 크기의 행렬이 생성되었으며 이를 SVD 방식에 의해 (그림 4)와 같이 각각 U, S, V의 행렬로 분할하였다. 이때 각 사용자별 열 벡터들은 TFIDF 방식에 의해 구한 가중치 값을 활용하였다.

(그림 5)는 실험 자료를 통한 SVD 분할 이후에 singular value의 값을 그래프로 표시한 것이다. 이 그래프에서도 역시 초기의 소수 값들이 원 행렬의 자료를 상당 부분 설명해 주는 것을 알 수 있다.



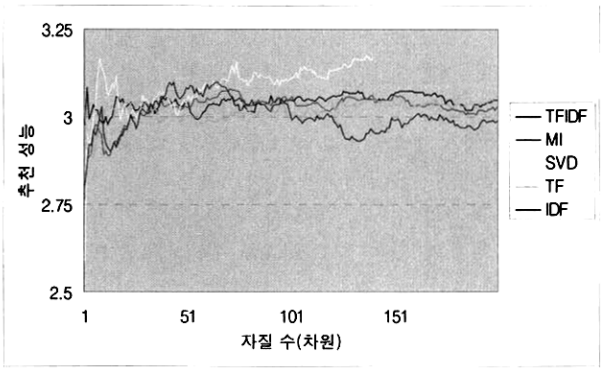
(그림 5) Singular value의 값

SVD를 통해 얻은 행렬을 추천에 사용할 때에는 다음과 같은 방식으로 사용자 i 와 추천 대상 서적의 유사도를 계산한다. 추천 대상 서적의 벡터를 X_q라고 하면 다음 행렬의 i 번째 열의 값이 둘 사이의 유사도가 된다.

$$X_q^{'}USV^{'} \tag{12}$$

3.3 결과

앞서 소개한 자질 추출에 따라 자질 수 별로, 혹은 차원(SVD의 경우)의 수 별로 추천 실험한 결과는 다음 (그림 6)과 같다. 이 값은 추천 대상 32권의 책을 각각 실험 참여자와의 유사도를 계산한 후 유사도 상위 3권의 책을 추천하는



(그림 6) 자질 수 및 차원 수 별 추천 성능 비교

것으로 가정하여 이때 실험 참여자가 실제 평가한 5점 척도의 평가 점수를 평균한 것이다.

실험 결과는 다음과 같이 정리해 볼 수 있다. 전반적으로 SVD에 의한 추천 결과가 다른 방식에 비해 높은 추천 점수를 나타내고 있다. TFIDF 방식은 TF 방식보다 좋은 결과를 내지만 TF, TFIDF, MI, IDF 방식 등의 추천 점수는 큰 차이를 보이지는 않는다. TFIDF 및 TF 방식의 경우 초기 매우 소수의 자질을 사용한 추천 결과가 뒤 부분의 많은 수의 자질을 사용한 추천 결과에 뒤쳐지지 않는 것으로 나타났다. 정도의 차이는 있으나 MI, IDF 방식 등에서도 유사한 형태로 초기 소수의 자질을 활용하여도 큰 추천 성과의 하락은 없다는 것을 관찰할 수 있다. SVD의 경우는 흥미로운 결과를 보여주고 있는데 차원의 개수가 약 9개 가량일 때 매우 높은 추천 성능을 보이고 이후 추천 성능이 하락한 이후에 지속적으로 추천 성능이 향상하여 초기 9개 차원으로 추천했을 때와 유사해 지고 있다. 이는 Latent semantic 등의 효과에 의한 것으로 추측할 수 있고 SVD 방식에서 소수의 차원만 사용하고도 좋은 추천 성능을 얻는 것이 가능함을 보여주고 있다.

하지만 본 실험 결과를 해석하는데 다음과 같은 점들을 유의해야 한다. 첫째, 실험 결과를 통해 대체적으로 SVD를 활용한 방식이 더 나은 추천 성능을 보인다는 점을 확인할 수는 있으나 구체적으로 어느 정도 수의 차원을 활용하는 것이 추천에 도움이 되는지는 실제 적용되는 상황에 따라 다를 수 있다. 예를 들어 (그림 6)의 경우에서 IDF의 경우 자질의 수가 약 50개일 때의 성능이 SVD 방식이 50개의 차원을 활용했을 때 보다 더 높게 나타나고 있다. 둘째, SVD

에서의 차원이 정확히 자질의 수와 같은 개념은 아니다. 실제로 자질의 수를 활용할 경우 추천을 위해 추천 시스템이 유지해야 하는 정보의 양은 사용자 수 곱하기 자질의 수가 된다. 그러나 SVD에서 필요한 정보의 양은 {(키워드 수×차원)+(사용자 수 X 차원)+차원} 이 된다. 따라서 SVD 방식이 더 많은 정보를 요구한다고 할 수 있으나 이때 정보의 양도 차원에 정비례하는 수준이다. 그러나 추천 성능 보다 추천에 필요한 정보의 양이 매우 중요한 경우 두 가지 요소를 모두 고려하여 자질 추출 방식을 선택하는 것이 필요하다.

4. 결론 및 추후 연구

내용 기반 추천의 경우 상품과 고객의 특성을 비교하여 고객에 적합한 상품을 추천한다. 상품과 고객의 특성을 표현하는 자질의 수가 늘어남에 따라서 추천 시스템의 성능에 나쁜 영향을 준다. 본 연구의 의의는 기계 학습, 패턴 인식, 또는 정보 검색 등의 분야에서 널리 쓰이는 자질 추출 방법들을 상품 추천 분야에 적용해 보고 각 방식들의 결과를 비교 분석한 데 있다. 이를 위해 국내의 한 인터넷 서점에서의 가상 구매 실험 및 추천 대상 서적들에 대한 평가 자료를 수집하여 소비자가 과거 구매한 상품의 특성을 활용하는 내용 기반 추천에 대한 실험을 수행하였다. 실험에서는 상품과 고객의 자질들의 수가 많을 때 그 중 효과적인 소수의 자질 혹은 적은 양의 정보만을 추출해 내기 위해 여러 방법들을 적용하여 추천 성과를 비교하였다. TF, IDF, TFIDF, Mutual Information, SVD(Singlular Vector Decomposition) 기법 등 5가지 방법을 비교하였으며, 실험 결과 SVD 기법이 내용 기반 추천 시에 전체적으로 좋은 추천 성능을 보여주었으며 적은 수의 차원을 사용했을 때에도 높은 성능을 낼 수 있음을 보여주었다.

연구의 한계 및 추후 연구 과제는 다음과 같다. 현재의 실험에서는 자질 추출 기법의 성능을 비교하였을 뿐, 몇 개의 자질을 선택하는 것이 좋은 지에 대한 일반적인 분석을 제시하고 있지는 못하다. 따라서 최적의 자질 개수 선정에 대한 추가적인 연구가 필요하다. 또한 본 연구 결과를 일반화하기 위해서는 다른 상황의 데이터에 대한 적용이 필요하다. 예를 들어, 인터넷 뉴스 사이트와 같이 내용기반 추천이 유용하게 사용될 수 있는 사이트에 대한 추가적인 실험 및 분석도 흥미로운 연구과제이다.

참 고 문 헌

[1] J. A. Konstan, B. N. Miller, et al., "GroupLens: Applying Collaborative Filtering to Usenet News," Communications of the ACM, Vol.40, No.3, pp.77-87, 1997.

[2] A. Ansari, S. Essegaier, et al., "Internet Recommendation Systems," Journal of Marketing Research(JMR), Vol.37, No.3, pp.363-375, 2000.

[3] W. W. Cohen and W. Fan, "Web-collaborative Filtering : Recommending Music by Crawling the Web," Computer Networks, Vol.3, No.1-6, pp.685-698, 2000.

[4] G. Greco, S. Greco, et al., "Collaborative Filtering Supporting Web Site Navigation," AI Communications, Vol.17, No.3, pp.155-167, 2004.

[5] M. Vozalis and K. G. Margaritis, "On the Combination of User-based and Item-based Collaborative Filtering," International Journal of Computer Mathematics, Vol.81, No. 9, pp.1077-1096, 2004.

[6] K. Tout, D. J. Evans, et al., "Collaborative Filtering : Special Case in Predictive Analysis," International Journal of Computer Mathematics, Vol.82, No.1, pp.1-11, 2005.

[7] D. C. Wilson, B. Smyth, et al., "Sparsity Reduction in Collaborative Recommendation : A Case-Based Approach," International Journal of Pattern Recognition & Artificial Intelligence, Vol.17, No.5, pp.863-884, 2003.

[8] Z. Huang, H. Chen, et al.,. "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering," ACM Transactions on Information Systems, Vol.22, No.1, pp.116-143, 2004.

[9] D. O. Sullivan, B. Smyth, et al., "Preserving Recommender Accuracy and Diversity in Sparse Datasets," International Journal of Artificial Intelligence Tools, Vol.13, No.1, pp.219-236, 2004.

[10] B. Mobasher, R. Cooley, et al., "Automatic Personalization Based on Web Usage Mining," Communications of the ACM, Vol.43, No.8, pp.142-151, 2000.

[11] S.-S Weng. and M.-J. Liu, "Feature-based Recommendations for One-to-one Marketing," Expert Systems with Applications, Vol.26, No.4, pp.493-508, 2004.

[12] J. S. Deogun, S. K. Choubey, et al., "Feature Selection and Effective Classifiers," Journal of the American Society for Information Science, Vol.49, No.5, pp.423-434, 1998.

[13] A. Al-Ani, M. Deriche, et al., "A New Mutual Information Based Measure for Feature Selection," Intelligent Data Analysis, Vol.7, No.1, pp.43-57, 2003.

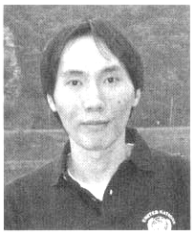
[14] D. Huang, and T. W. S. Chow, "Effective Feature Selection Scheme Using Mutual Information," Neurocomputing, Vol.63, No.1-4, pp.325-344, 2005.

[15] M. D. Gordon and S. Dumais, "Using Latent Semantic Indexing for Literature Based Discovery," Journal of the American Society for Information Science, Vol.49, No.8, pp.674-685, 1998.

[16] D. L. Lee, and H. Chuang, "Document Ranking and the Vector-space Model," IEEE Software, Vol.14, No.2, pp.67-76, 1997.

[17] T. A. Latsche and M. W. Berry, "Large-scale Information Retrieval with Latent Semantic Indexing," Information Sciences, Vol.100, No.1-4, pp.105-137, 1997.

안 형 준



e-mail : hjahn@waikato.ac.nz
1995년 한국과학기술원 경영과학과(학사)
1997년 한국과학기술원 산업경영학과 (공학석사)
2004년 한국과학기술원 경영공학과 (공학박사)
2004년 한국과학기술원 테크노경영대학원 대우교수
2005년~현재 Lecturer, Senior Lecturer(from Feb. 2006), Waikato Management School, University of Waikato, New Zealand
관심분야 : 멀티 에이전트 시스템, 데이터마이닝, e-Supply Chain Management

김 종 우

e-mail : kjw@hanyang.ac.kr
1989년 서울대학교 수학과(학사)
1991년 한국과학기술원 경영과학과 (공학석사)
1995년 한국과학기술원 산업경영학과 (공학박사)
1995년~1996년 한국과학기술원 연수연구원(Post Doc.)
1996년~2003년 충남대학교 통계학과 전임강사, 조교수, 부교수
2003년~현재 한양대학교 경영학부 부교수
1999년~2000년 University of Illinois at Urbana-Champaign, Visiting Scholar
관심분야 : 이비즈니스 추천기술, B2B 프로세스 모델링, 데이터마이닝, 의사결정지원시스템