Support Vector Machine을 이용한 개인 사용자 선호 의상 추천[1)]

강한훈[O] 유성준
세종대학교 컴퓨터공학과
kangcom@paran.com[O], sjyoo@sejong.ac.kr

Recommendation of User Preferred Clothes using Support Vector Machine

Hanhoon Kang[0] SeongJoon Yoo
Dept. of Computer Engineering, Sejong University

요 약

본 논문에서는 의상에 대한 사용자 선호도를 찾아내는 기법에 대하여 기술한다. 의상에 대한 사용자 선호도를 찾기 위해서 의상 데이터에 대해 데이터 모델을 새롭게 제안한다. 이 데이터 모델을 기반으로 사용자의 의상관련 히스토리를 저장한다. 이렇게 저장된 히스토리 정보에 기계 학습 기법 중 최근 각광받고 있는 SVM 기법을 적용하여 사용자 선호도를 찾아내도록 하였다. 이 결과를 다른 학습 기법인 Naive Bayes 기법을 사용하여 의상에 대한 사용자 선호도를 검색한 성능과 비교하여 우리 모델이 더 좋다는 것을 확인하였다. 우리는 5명의 사용자에 대해서 동일한 취향을 갖는 사용자가 몇 명인지에 따라 A(모두 다름), B(2명), C(3명), D(4명), E(모두 같음) 형태별, 사용자별 1000건의 히스토리를 일정한 기준에 따라 생성했다. 그리고 이 중에서 900건을 학습용 데이터, 100건을 검증용 데이터로 선정하여 실험이 진행되었다.

1. 서 론

이 논문에서는 의상에 대한 사용자 선호도를 찾아내는 기법에 대해서 기술한다. 현재까지 진행된 개인화와 관련된 연구의 예를 들면 EPG(Electronic Program Guide)를 기반으로 한 TV 방송 프로그램 추천 방법과 사용자가 이전에 검색한 결과에 대해서 관심을 갖는 부분에 대해서 이후에 검색한 결과와 비슷할 경우, 이를 우선적으로 사용자에게 보여줄 수 있도록 하는 개인화된 검색 방법, 그리고 쇼핑몰 등을 통해서 이전에 구매했던 상품 정보를 기반으로 하여 신상품을 추천해 줄 수 있도록 하는 방법 등 다양한 분야에서 개인화와 관련된 활발한 연구가 되어왔지만 의상에 대한 사용자 선호도 기반 검색 기술은 거의 이루어지지 않고 있다.

우리는 의상에 대한 사용자 선호도를 찾아내기 위해서 의상 데이터를 모델링 하였는데, 우리가 선정한 의상은 남성용 정장으로 다음과 같이 구성하였다.

[재 킷]={스타일, 소재, 사이즈, 색상, 트기, 포켓모양 }
[셔 츠]={사이즈, 칼라모양, 색상}
[바 지]={앞면, 뒷면, 옆면스타일, 사이즈, 색상}
[넥타이]={배경색, 무늬}

그리고 이 특징 값들을 기계학습 기법에 적용하기 위해서는 정량화된 수치 데이터로 표현하여야 한다. 그래서 우리는 모델링한 정장 데이터에 대해서 Vector Space Model[1]에서 사용하는 벡터 형태로 표현하였다. 다음은 재킷 스타일 중 하나인 '홑자락 1단추 피크라펠' 이라는 특징 값을 벡터 형태로 표현한 예이다.

홑자락 1단추 피크라펠 = (1,0,0,0,0,0,0,0,0,0,0,0)

개인화와 관련된 기존 연구에서는 대부분이 많은 양의 단어가 포함된 문서를 기반으로 하여 주요 단어를 추출하고 그 단어에 대한 가중치 값을 추출하여 정량화된 수치 데이터로 사용된다. 이 가중치 값을 구하기 위해서는 별도의 추출 알고리즘[7]을 사용한다. 우리의 연구에서는 많은 양의 텍스트 데이터가 아닌 의상에 대한 간략한 정보만을 취급하므로 기존 연구에서 사용되었던 추출 알고리즘이 적합하지 않아, 의상에 대한 정량화된 수치 데이터를 새롭게 표현하여 사용한다.

우리는 이 실험에서 5명의 사용자에 대해 동일한 취향을 갖는 사용자의 수에 따라 A(모두 다름), B(2명), C(3명), D(4명), E(모두 다름) 와 같이 분류하였다. 사용자별 1,000건의 벡터 형태로 된 히스토리를 만들되, 사용자 마다 설정된 구매 취향의 범위 안에서 임의로 골라내어 데이터를 만들고 900건을 학습용 데이터, 100건을 검증용 데이터로 선정하였다. 이 데이터를 이용하여 SVM[2]과 Naive Bayes[3]에서 비교 실험한 결과, A형태에서 가장 잘 분류하였고, B,C,D,E 형태로 갈수록 동일한 취향(16개의 속성 모두가 같음)을 가진 사용자가

1) '서울시 전략 산업 혁신 클러스터 육성 지원 사업'의 지원을 받아 수행된 연구임.

많아짐에 따라 분류 성능이 낮아지는 것을 확인할 수 있었다. 또한 우리가 채택한 SVM이 Naive Bayes 보다 전반적으로 나은 성능을 보였다.

2. 관련 연구

[4][5]는 TV 방송 프로그램에 대한 추천 기법에 대해서 기술하고 있다. 특히 [4]는 EPG를 이용한 추천 방법에 대해서 기술하고 있는데, 이 기법은 사용자가 기존에 시청하였던 프로그램에 대해서 제목, 시간 등의 간단한 정보가 저장되고, 이 정보로부터 bi-gram 단위로 어휘를 추출[12]하여 추출된 어휘로부터 feature 값을 구한 후에 예측을 통해서 방송 프로그램을 추천한다.

[6][7][8]은 텍스트 정보에 대해 개인화된 검색 기술을 제공할 수 있도록 하기 위해서 텍스트 문서에 대한 분류 기법을 기술하고 있다. 이미 분류된 특정 카테고리 문서 내의 주요 단어를 추출하고 학습 시킨 후 새로운 문서 내의 주요 단어를 추출하여 이 문서가 어느 카테고리로 분류되는지 예측할 수 있는 방법들을 제안한다.

[9][10]은 상품 정보에 대한 추천 기법에 대해서 기술하고 있는데, [9]는 내용 기반 추천(Content-Based Recommendation) 방식으로 기존 구매 상품에 대한 사용자의 선호도와 사용자가 구매했던 상품의 정보를 통해 유사한 상품을 추천해주도록 하는 방식으로 상품에 대한 특징 요소들을 이용하기 때문에 단순하다는 것이 장점이지만, 이전에 구매했던 상품에 대해서만 추천할 수 있다는 단점을 가진다.

[10]은 협업 필터링(Collaborative Filtering) 방법으로 소비하는 취향이 유사한 사용자를 그룹으로 묶고, 새로운 사용자의 성향과 비슷한 성향을 지닌 기존 고객을 찾아 그들이 선호하는 상품을 추천해 주도록 하는 방법이다. 기존의 선호하는 데이터만을 가지고 분석했던 내용 기반 추천 방식에 비하면 협업 필터링은 다양한 상품이 추천될 수 있지만, 새로운 사용자의 취향과 비슷한 기존의 사용자 그룹이 존재하지 않는다면 추천 정보를 얻어오지 못한다는 단점을 가진다.

특허 [11]은 규칙 기반 필터링(Rule-Based Filtering) 방법을 이용하여 사용자의 취향을 이용하여 미리 생성된 규칙(Rule)을 생성한다. 이 방법은 추천 엔진이 간단하지만 정해진 규칙에 대해서만 적용가능하기 때문에 추천 기법이 한정적이라는 단점을 가진다.

위에서 기술한 [4][5][6][7][8]은 모두 텍스트 정보로부터 적절한 어휘를 추출하고 그 어휘에 대한 정량화된 특징(feature) 값을 구한다. 그리고 그 값을 이용, 학습 및 예측하는 기법에 대해서 소개하고 있다. 학습 기법에서는 이 feature 값에 따라 성능이 달라질 수 있기 때문에 feature를 추출하고 선택하는 방법에 대한 부분도 활발하게 연구되어 오고 있다. [9][10]은 전형적인 상품에 대한 추천 기법에 대해서 서술하고 있기 때문에 우리가 선정한 의상에 대해서도 이 기법을 적용할 수도 있었으나, 앞서서 설명한 것처럼 [9][10]은 여러 가지 단점을 내포하고 있기 때문에 우리는 학습/예측 기법을 통해서 추천/분류 실험을 하였는데, 우리는 텍스트 기반이 아닌 의상 모델에 대한 것이므로 특징 값을 추출하는데 있어서 텍스트 분류 기법에서 소개하는 데이터표현과 선호도 추출기법이 아닌, [부록]에서 명시하고 있는 99차원의 0 또는 1의 벡터 형태로 단순하게 표현하여 이를 학습 알고리즘(SVM, Naive Bayes)의 입력 값으로 한 추천/분류 실험을 하여 성능을 측정하였다.

3. 데이터 모델

정장 한 벌을 구성하기 위한 요소는 사람들의 기준에 따라 차이를 둘 수 있다. 우리는 그 요소를 [표 1]과 같이 구성하여 정장에 전체에 대한 데이터 모델을 만들었다.

표 1. 정장 요소

[재 킷]={스타일, 소재, 사이즈, 색상, 트기, 포켓모양 }
[셔 츠]={사이즈, 칼라모양, 색상}
[바 지]={앞면, 뒷면, 옆면스타일, 사이즈, 색상}
[넥타이]={배경색, 무늬}

여기에선 정장을 구성하는 네 가지 유혀의 의상에 대해서만 특징 값을 벡터 형태로 표현하였는데 이를 모든 의상의 데이터와 그의 특징 값을 나타내는 형태로 확장할 수 있다. 이 때 일반식은 다음과 같다.

$$C = \{F_1, F_2, \ldots, F_n\}$$

여기서 C 는 의상 종류, F_i 는 해당 의상의 특징 값을 의미한다.

이 데이터 모델은 추천/분류 실험을 위해서 정량화된 특징(feature)값으로 변환되어야 할 필요가 있다. 즉, 기계학습을 통해서 학습하고 예측하기 위한 입력 값을 의미하는데, 우리는 이 값의 표현 방법을 VSM에서 사용하는 벡터 표현 방법을 이용하여 [표 2]와 같은 형태로 표현한다. 기존 연구에서 다루어졌던 데이터 표현 방법은 많은 양의 텍스트 정보를 기반으로 하였기 때문에 이 연구에서 다루는 데이터 모델과 같이 간략한 의상 정보만을 표현하는 연구에는 적합하지 않다. 따라서 우리의 연구에서는 VSM에서 사용하는 벡터 표현 방법을 사용한다.

표 2. 재킷에 대한 벡터 표현 예

속성	특징 값	벡터 표현
스타일	홑자락 1단추 피크라펠	(1,0,0,0,0,0,0,0,0,0,0,0)
소재	색소니	(1,0,0,0,0,0,0,0)
사이즈	95	(1,0,0,0)
색상	청색	(1,0,0,0)
트기	벤트리스	(1,0,0,0)
포켓모양	플랩	(1,0,0,0,0,0)

각각의 속성 별로 벡터 표현의 자리 수는 해당 속성이 지닐 수 있는 특징 값의 수에 따른다. 즉 재킷 요소에서 스타일 속성은 12가지이므로 벡터 표현의 자리 수는 12자리이다. 그리고 [부록]을 참조하면 해당 속성의 특징 값 별로 순서대로 이루어져있는데, 그 순서를 기준으로 벡터 값에 '1'을 설정한다. 예를 들어 '홑자락 4단추'는 번호가 2번이며 세 번째에 위치한다. 따라서 세 번째에 '1'을 설정하여 벡터 형태로 표현한다.

다음은 [표 1]의 구성대로 특징 값을 적용한 예이다.이 예에서는 총 16가지의 특징을 나타내고 있다.

```
재 킷 = {홀자락 4단추, 린넨, 95, 회색, 벤트리스, 플랩}
셔 츠 = {95, 핀, 흰색}
바 지 = {앞면1, 뒷면3, 옆면2, 95, 청색}
넥타이 = {흰색, 도트}
```

이를 벡터형태로 표현하면 아래와 같이 된다.

```
재 킷={ (0,0,1,0,0,0,0,0,0,0,0,0),(0,0,0,0,0,0,1,0),
        (1,0,0,0),(0,1,0,0),(1,0,0,0),(1,0,0,0,0,0) }
셔 츠={ (1,0,0,0),(1,0,0,0,0),(1,0,0,0,0,0,0,0) }
바 지={ (1,0,0,0,0),(0,0,1,0,0),(0,1,0,0,0),(1,0,0,0),
        (1,0,0,0) }
넥타이={(1,0,0,0,0,0,0,0),(0,0,0,1,0,0,0,0,0,0,0,0)}
```

정장을 이루는 16개의 속성에 대해서 각각의 벡터 값을 조합하여 1차원 배열 형태로 구성한다. 즉 위의 예에서 괄호만 제거해주면 99차원의 벡터로 표현되는 것을 알 수 있다. 이 값이 정장에 대한 정량화된 수치 데이터이며 학습 기법의 입력 값으로 쓰여 추천/분류한다.

4. 실험

4.1 기계 학습

우리는 자바로 구현된 SVM Library[13]와 Naive Bayes Libary[14]를 이용하여 비교 실험 하였다. 두 학습 Library 모두 기계학습 기법의 종류로써 학습(train)과 예측(predict)의 과정을 수행하게 된다.

우리는 앞서서 정장 한 벌에 대해서 벡터 형태로 표현하는 법을 알 수 있었다. 이 벡터 형태로 구성된 값을 기계학습 알고리즘의 실제 입력 값으로 한다. 특별히 SVM Library 같은 경우는 입력 값 이외의 parameter 값을 요구로 하는데 kernel type이 그 한 예이고, 우리는 kernel type을 라이브러리에서 기본 설정 값으로 제공하는 rbf type으로 하였다. 이외의 다른 parameter 값에 대해서도 라이브러리 자체에서 설정된 기본 값에 의거하여 실험을 하였다.

[표 3]과 [표 4]는 각각의 학습 Library별로 사용되는 입력 데이터 형식의 예를 나타내고 있다. 벡터 형태로 표현된 값을 이용하여 각각의 학습 Library에 알맞은 형식으로 변환하여 사용한다.

표 3. SVM Library 입력 데이터의 예 (일부)

```
1 1:0 2:0 3:1 4:0 5:0 6:0 7:0 8:0 9:0 10:0 ... 99:0
1 1:1 2:0 3:0 4:0 5:0 6:0 7:0 8:0 9:0 10:0 ... 99:1
2 1:0 2:0 3:0 4:1 5:0 6:0 7:0 8:0 9:0 10:0 ... 99:0
5 1:1 2:0 3:0 4:0 5:0 6:1 7:0 8:0 9:0 10:0 ... 99:1
```

정장 한 벌에 대해서 99개의 벡터 값을 가지므로 위의 데이터에서 99까지 설정되어 있는 것을 확인할 수 있다. 가장 앞에 설정된 숫자는 사용자를 구분 짓는다.

표 4. Naive Bayes Library 입력 데이터의 예(일부)

```
@relation suit
@attribute attr1 { 0, 1 }
@attribute attr2 { 0, 1 }
...
@attribute attr99 { 0, 1 }
@attribute class { 1, 2, 3, 4, 5 }

@data
0,0,0,0,1,0,0,0,0,0, ... ,1
0,0,1,0,0,0,0,0,0,0, ... ,1
0,0,0,0,0,0,1,0,0,0, ... ,2
0,0,0,0,1,0,0,0,0,0, ... ,5
```

Naive Bayes Library에서 사용되는 데이터 역시 '@data' 부분에 콤마로 구분하여 99개의 숫자로 이루어진다. 이 데이터에서는 가장 마지막이 사용자를 구분할 수 있는 값이다.

이러한 형식에 맞춘 900건의 학습용 데이터를 학습시키면 각각의 사용자별로 학습 모델이 생성된다. 각각의 알고리즘에 따라 계산된 결과 값이다. 이 모델과 검증용 데이터 100건(학습용 데이터와 형식이 일치)을 입력 값으로 하여 예측을 하게 되면 각각의 Library는 예측한 결과를 화면상으로 보여주거나 파일로 생성하게 된다.

4.2 성능 측정

텍스트 기반 분류에서 사용되는 정확률(Precision), 재현율(Recall), F_1-measure를 통해서 성능을 측정한다.

$$\text{정확률}(P) = TP / (TP + FP)$$
$$\text{재현률}(R) = TP / (TP + FN)$$
$$F_1\text{-Measure} = (2*R*P) / (R+P)$$

표 5. 위 식에서 표현된 기호의 의미

분류 \ 예측	TRUE	FALSE
TRUE	TP	FP
FALSE	FN	TN

사용자 5명에 대해서 각각 정확률, 재현율, F_1-Measure를 구하고 그 값들에 대한 평균값을 구하기 위한 방법으로 정확률과, 재현율, 그리고 F_1-Measure에 대해서 각각 micro-averaging과 macro-averaging값을 구한다.

$$R = \frac{\sum_{i=1}^{n} TPi}{\sum_{i=1}^{n}(TPi + FPi)} \qquad R = \frac{\sum_{i=1}^{n} Ri}{n}$$

그림 1 micro-averaging과 macro-averaging

4.3 학습 데이터의 선정

우리는 [표 6]과 같이 사용자별 구매 취향을 설정하고, 그 범위 안에서 다시 형태(A,B,C,D,E)를 고려하여 5

명의 사용자에 대해서 각각 임의로 1,000건의 히스토리를 만들었다. 단, 정장 한 벌을 이루기 위한 16개[2]의 속성 중에서 15개의 속성이 5명 모두가 같고, 넥타이 '무늬' 속성만 모두가 다를 때 (A형태), 2명 같을 때 (B형태), 3명 같을 때(C형태), 4명 같을 때(D형태), 모두 같을 때(E형태) 분류 실험을 하되 900건을 학습용 데이터로 사용했으며, 나머지 100건을 검증용 데이터로 사용했다. 16개의 속성 중 단지 1개의 속성만 다르게 한 이유는 1개의 속성만 달리하여도 분류할 수 있는 최소 단위가 될 수 있기 때문이다.

표 6. 5명 사용자에 대한 구매 취향

재킷	속성	스타일	트기	소재	주머니	색상	사이즈
	번호[3]	0,1,2,3	2,3	0,1,3	2,3	0,1	0

바지	속성	앞면	뒷면	옆면	색상	사이즈
	번호	0,1,2	0,1,2	0,1,2	0,1	1

셔츠	속성	칼라모양	색상	사이즈
	번호	0,1,3	0,2,3	0

넥타이	속성	무늬	배경색
	번호	(사용자 A) 0,1,2 (사용자 B) 0,4,5 (사용자 C) 0,7,8 (사용자 D) 0,9,10,11 (사용자 E) 0,3,6,12	0,2,4

예를 들어, 이 실험에서 5명 모두가 다를 때의 '무늬' 속성은 A={0,1,2}, B={4,5}, C={7,8},D={9,10,11}, E={3,6,12}가 되며, 서로 중복되는 속성이 없다.
2명이 같을 경우에는 A={0,1,2}, B={0,4,5},C={7,8}, D={9,10,11}, E={3,6,12}이다. 즉 '0' 속성이 A와 B, 2명에게만 속해있는 것을 의미한다. 3명, 4명, 5명이 같을 때도 각각의 사용자마다 '0' 속성만 추가해주면 된다.

4.4 학습 데이터에 대한 유사도

임의로 만든 데이터라고 하더라도 넥타이 '무늬' 속성 하나에 대해서 형태(A,B,C,D,E)별 유사도는 구매 성향이 비슷한 사람이 많아질수록 높아지게 된다. 이 실험은 유사 고객이 얼마나 많으냐에 따라 분류 성능이 얼마나 되는지를 알아보기 위한 것이기 때문에 사용자별로 구매 취향에 의해 생성된 데이터에 대해서 각각의 속성별로 구매 횟수를 구하였다. 그리고 사용자별로 얻어진 각각의 속성별 구매 횟수를 가지고 Normalized Euclidean Distance[4] 계산 방법을 통해서 형태(A,B,C,D,E)별로 [표 7]과 같이 계산하여 결과로 얻은 유사도 25개에 대한 평균값을 [표 8]과 같이 구하였다.

2) 재킷(6개), 셔츠(3개), 바지(5개), 넥타이(2개)
3) 번호에 대한 상품 정보는 [부록] 참조
4) $d=\sqrt{\sum_{i=1}^{v}\left(\frac{(d1-d2)^2}{v}\right)}$, v=벡터표현 속성의 개수(99)
*Normalized Euclidean Distance는 값의 범위를 줄여준다.

표 7. 5명의 사용자간 유사도

사용자	A	B	C	D	E
A	0	x	x	x	x
B	x	0	x	x	x
C	x	x	0	x	x
D	x	x	x	0	x
E	x	x	x	x	0

Normalized Euclidean Distance 계산 결과 값이 0인 것은 완전히 일치한다는 의미이다. A와 A는 동일하기 때문에 0이고, 나머지 동일한 것끼리의 유사도 역시 0이 된다. x는 임의의 값을 나타낸다.

표 8. [표 7]의 25개 유사도 평균 값

A형태	B형태	C형태	D형태	E형태
66.8	63.1	57.6	54.7	50.1

구매 성향이 비슷한 사용자가 많아질수록 유사도 값은 0에 가까워지고 있으며 이것은 유사도가 높다는 의미이다. 즉, 유사도가 높으면 높을수록 분류되는 성능이 낮아진다. 이는 실험 결과에서 확인할 수 있다.

4.5 결과

[4.2]에서 소개한 측정 방법을 통하여 측정한 실험 결과는 [표 9],[표 10],[그림 2]과 같다.

표 9. SVM을 이용한 실험 결과

측정률 / 형태	Micro-Avg.			Macro-Avg.		
	정확률	재현률	F_1	정확률	재현률	F_1
A형태	100%	100%	100%	100%	100%	100%
B형태	95.6%	95.6%	95.6%	96.4%	95.6%	96.0%
C형태	90.4%	90.4%	90.4%	93.5%	90.4%	91.9%
D형태	84.2%	84.2%	84.2%	91.2%	84.2%	87.5%
E형태	82.0%	82.0%	82.0%	90.5%	82.0%	86.1%

표 10. Naive Bayes를 이용한 실험결과

측정률 / 형태	Micro-Avg.			Macro-Avg.		
	정확률	재현률	F_1	정확률	재현률	F_1
A형태	100%	100%	100%	100%	100%	100%
B형태	96.2%	96.2%	96.2%	96.3%	96.2%	96.3%
C형태	90.2%	90.2%	90.2%	91.2%	90.2%	90.7%
D형태	82.8%	82.8%	82.8%	85.2%	82.8%	84.0%
E형태	80.2%	80.2%	80.2%	82.6%	80.2%	81.4%

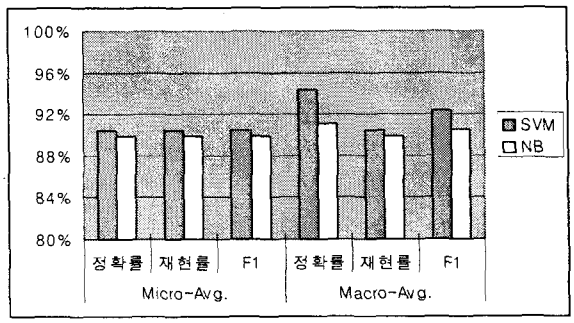

그림 2 SVM과 Naive Bayes의 성능 비교

A형태는 5명의 사용자에 있어서 정장 한 벌에 대한 16개의 속성 중 한 가지의 속성만 달리하여도 잘 분류하지만 B->C->D->E형태로 진행할수록 16개의 속성 값 모두가 같은 사용자가 많아질수록 즉, 구매 유사도가 높아지기 때문에 분류 성능이 낮아지는 현상을 발견할 수 있다. 또한 B형태에서는 Naive Bayes를 이용했을 때 SVM 보다 높은 결과를 나타나지만 [그림 2]와 같이 전반적으로 SVM이 Naive Bayes 보다 나은 측정 결과를 나타내고 있다. [그림 2]는 SVM과 Naive Bayes로 실험한 각각의 5개의 형태에 대한 평균값을 비교한 것이다.

4.6 순위 결정

기계학습 기법을 통한 분류 실험에서 동일한 취향을 가진 사용자가 많으면 많아질수록 분류 성능이 낮아지는 것을 확인할 수 있었다. 그러나 16개의 속성 모두가 같은 사용자가 몇 명이나 될 것인가에 대해서 이항분포 이론[15]는 극히 낮을 것이라고 확률적으로 예측한다. 그러나 실제로 이러한 기법을 이용한다고 했을 때 동일한 취향을 갖는 사용자가 몇 명이나 될 것인지는 예측할 수 없다. 따라서 향후 연구에서는 기계학습 기법에서 잘못 예측하여 잘못 분류하더라도, 2차적인 방법으로 사용자의 히스토리와 추천/분류된 결과와의 유사도 비교를 거친 후에 유사도 가장 높은 순으로 순위를 결정하여 사용자에게 보여주도록 하는 방법도 고려할 수 있다.

5.결론 및 향후 연구

본 실험에서는 사용자간 구매 유사도가 낮을수록 잘 분류되는 것을 확인할 수 있었고, SVM이 Naive Bayes 보다 더 나은 분류 성능을 보여주었다. 그러나 동일한 성향(16개의 속성 모두가 같은 것)의 사용자가 많으면 많아질수록 잘못 분류될 가능성이 있으므로, 2차적인 방법으로 VSM 등을 이용한 유사성 비교를 통해서 사용자에게 적합한 상품별로 순위를 설정하여 보여주도록 하는 방법도 고려할 수 있다.

현재 연구는 다양한 의상 중에서 특징 값들이 많은 남성 정장에 대해서만 실험을 하였다. 이외에도 더 많은 사용자가 관심을 갖는 더 많은 의상에 대해서 추천할 수 있는 기법의 연구가 더 필요하다.

6.참고문헌

[1]Salton, Gerard. Introduction to Modern Information Retrieval. McGraw-Hill, 1983
[2] Christopher J. C. Burges, "A tutorial on support vector machine for pattern recognition",Data Mining and Knowledge Discovery, 1998
[3] Markus Forsberg, Kenneth Wilhelmasson, "Automatic Text Classification with Bayesian Learning", 2003
[4]Jin An Xu, Kenji Araki, "A Personalized Recommendation System for Electronic Program Guide", AI 2005, LNCS 3809, pp.1146-1149, 2005.
[5]L. Ardissono, C. Gena, P. Torasso, F. Bellifemine, A. Chiarotto, A. Difino, B. Negro, "Personalized Recommendation of TV Programs", AI*IA 2003, LNCS 2829, pp. 474-486, 2003
[6]Thorsten Joachims, "Text Categorization with Support Vector Machine : Learning with Many Relevant Features", ECML 1998
[7]Subramanian Arumugam, "Classification Techniques for Categorization of Hypertext Documents", Technical Report, 2005
[8]Pencheng Wu, Thomas G. Dietterich, "Improving SVM Accuracy by Training on Auxiliary Data Sources", ICML 2004, July 4-8 2004, Banff, Alberta, Canada
[9]M.Balabanovic and Y. Shoham, "Fab: Content-Based Collaborative Recommendation", Communication ACM, Vol. 40, No. 3, 1997
[10]Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Item Based Collaborative Filtering Recommendation Algorithms", International World Wide Web Conference 2001
[11] "한 벌의 옷을 선택하기 위해 추천을 발생시키는 방법 및 시스템", 공개특허 특2002-0067507
[12] Nakagawa H., Mori T., "A Simple but Powerful Automatic Term Extraction Method."Proc.of 2nd International Workshop on Computational Terminology, COLING-2002 WORKSHOP, Taipei, 2002, 29-35.
[13] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[14] http://www.cs.waikato.ac.nz/ml/weka/
[15] http://en.wikipedia.org/wiki/Binomial_distribution
[16]남자의 옷 이야기 1(정장편), 시공사, 1997

7. 부록 - 남성 정장 요소 및 벡터 값

7.1 재킷

7.1.1 스타일

번호	특징 값	벡터 값
0	홑자락 1단추 피크라벨	(1,0,0,0,0,0,0,0,0,0,0)
1	홑자락 1단추 세미노치라벨	(0,1,0,0,0,0,0,0,0,0,0)
2	홑자락 4단추	(0,0,1,0,0,0,0,0,0,0,0)
3	홑자락 5단추 스텐칼라	(0,0,0,1,0,0,0,0,0,0,0)
4	홑자락 2단추 피크라벨	(0,0,0,0,1,0,0,0,0,0,0)
5	홑자락 2단추 노치라벨	(0,0,0,0,0,1,0,0,0,0,0)
6	홑자락 3단추 노치라벨	(0,0,0,0,0,0,1,0,0,0,0)
7	홑자락 3단추 피크라벨	(0,0,0,0,0,0,0,1,0,0,0)
8	겹자락 4단추 2장금 피크라벨	(0,0,0,0,0,0,0,0,1,0,0)
9	겹자락 4단추1장금 세미피크라벨	(0,0,0,0,0,0,0,0,0,1,0)
10	홑자락 1단추 피크라벨	(0,0,0,0,0,0,0,0,0,0,1)
11	겹자락 4단추 1장금 세미피크라벨	(0,0,0,0,0,0,0,0,0,0,1)

7.1.2 소재

번호	특징 값	벡터 값
0	색소니	(1,0,0,0,0,0,0,0)
1	태즈 메이니아	(0,1,0,0,0,0,0,0)
2	플란넬	(0,0,1,0,0,0,0,0)
3	알파카	(0,0,0,1,0,0,0,0)
4	트로피컬	(0,0,0,0,1,0,0,0)
5	홉색	(0,0,0,0,0,1,0,0)
6	린넨	(0,0,0,0,0,0,1,0)
7	포플린	(0,0,0,0,0,0,0,1)

7.1.3 사이즈/색상

번호	특징 값	벡터 값
0	95	(1,0,0,0)
1	100	(0,1,0,0)
2	105	(0,0,1,0)
3	110	(0,0,0,1)
0	청색	(1,0,0,0)
1	회색	(0,1,0,0)
2	검정색	(0,0,1,0)
3	방색	(0,0,0,1)

7.1.4 트기

번호	특징 값	벡터 값
0	벤트리스	(1,0,0,0)
1	센터벤트	(0,1,0,0)
2	사이드벤트	(0,0,1,0)
3	타이트 후크벤트	(0,0,0,1)

7.1.5 포켓 모양

번호	특징 값	벡터 값
0	플랩	(1,0,0,0,0,0)
1	제티드	(0,1,0,0,0,0)
2	패치	(0,0,1,0,0,0)
3	플리티드	(0,0,0,1,0,0)
4	벨로우즈	(0,0,0,0,1,0)
5	웰트	(0,0,0,0,0,1)

7.2 셔츠

7.2.1 사이즈

번호	특징 값	벡터 값
0	95	(1,0,0,0)
1	100	(0,1,0,0)
2	105	(0,0,1,0)
3	110	(0,0,0,1)

7.2.2 칼라 모양

번호	특징 값	벡터 값
0	핀	(1,0,0,0,0)
1	버튼다운	(0,1,0,0,0)
2	윈저	(0,0,1,0,0)
3	레귤러	(0,0,0,1,0)
4	탭	(0,0,0,0,1)

7.2.3 색상

번호	특징 값	벡터 값
0	흰색	(1,0,0,0,0,0,0,0)
1	청색	(0,1,0,0,0,0,0,0)
2	하늘색	(0,0,1,0,0,0,0,0)
3	분홍색	(0,0,0,1,0,0,0,0)
4	연두색	(0,0,0,0,1,0,0,0)
5	녹색	(0,0,0,0,0,1,0,0)
6	회색	(0,0,0,0,0,0,1,0)
7	방색	(0,0,0,0,0,0,0,1)

7.3 바지

7.3.1 앞면/뒷면/옆면 스타일

번호	특징 값	벡터 값
0	앞면, 뒷면, 옆면 스타일 1	(1,0,0,0,0)
1	앞면, 뒷면, 옆면 스타일 2	(0,1,0,0,0)
2	앞면, 뒷면, 옆면 스타일 3	(0,0,1,0,0)
3	앞면, 뒷면, 옆면 스타일 4	(0,0,0,1,0)
4	앞면, 뒷면, 옆면 스타일 5	(0,0,0,0,1)

7.3.2 사이즈/색상

번호	특징 값	벡터 값
0	95	(1,0,0,0)
1	100	(0,1,0,0)
2	105	(0,0,1,0)
3	110	(0,0,0,1)

번호	특징 값	벡터 값
0	청색	(1,0,0,0)
1	회색	(0,1,0,0)
2	검정색	(0,0,1,0)
3	방색	(0,0,0,1)

7.4 넥타이

7.4.1 배경색

번호	특징 값	벡터 값
0	흰색	(1,0,0,0,0,0,0,0)
1	청색	(0,1,0,0,0,0,0,0)
2	하늘색	(0,0,1,0,0,0,0,0)
3	분홍색	(0,0,0,1,0,0,0,0)
4	연두색	(0,0,0,0,1,0,0,0)
5	녹색	(0,0,0,0,0,1,0,0)
6	회색	(0,0,0,0,0,0,1,0)
7	방색	(0,0,0,0,0,0,0,1)

7.4.2 무늬

번호	특징 값	벡터 값
0	로얄크레스트	(1,0,0,0,0,0,0,0,0,0,0,0)
1	솔리드	(0,1,0,0,0,0,0,0,0,0,0,0)
2	모티브	(0,0,1,0,0,0,0,0,0,0,0,0)
3	도트	(0,0,0,1,0,0,0,0,0,0,0,0)
4	크레스트	(0,0,0,0,1,0,0,0,0,0,0,0)
5	태터솔	(0,0,0,0,0,1,0,0,0,0,0,0)
6	니트	(0,0,0,0,0,0,1,0,0,0,0,0)
7	헤어라인 스트라이	(0,0,0,0,0,0,0,1,0,0,0,0)
8	블러 스트라이프	(0,0,0,0,0,0,0,0,1,0,0,0)
9	펜슬 스트라이프	(0,0,0,0,0,0,0,0,0,1,0,0)
10	더블바 스트라이프	(0,0,0,0,0,0,0,0,0,0,1,0)
11	쵸크 스트라이프	(0,0,0,0,0,0,0,0,0,0,0,1)
12	레지멘탈 스트라이프	(0,0,0,0,0,0,0,0,0,0,0,1)