한국 퍼지 및 지능시스템 학회 2002, Vol. 12, No. 5, pp. 385-392

군집분석과 베이지안 학습을 이용한 웹 도서 동적 추천 시스템

Dynamic Recommendation System for a Web Library by Using Cluster Analysis and Bayesian Learning

최준혁* · 김대수** · 임기욱***
JunHyeog Choi*, DaeSu Kim**, KeeWook Rim***

*김포대학 컴퓨터계열
**한신대학교 컴퓨터학과
***선문대학교 산업공학과

요 약

기존의 동적 추천 시스템에서 사용하는 개인화 기법은 주로 협업 필터링 방식으로서 다른 사용자들에 대한 평가 정보를 이용하여 동적 링크를 제공하기 때문에 사용자가 고려하지 못한 아이템들을 추천한다는 장점을 갖고 있다. 그러나 협업 필터링 과정은 현재 사용자와 가장 유사한 패턴을 보이는 사용자를 선택하기 위해 전체 사용자와의 유사도를 재계산해야 한다는 계산의 복잡성과 사용자 프로화일의 정보가 현 사용자의 키워드 입력 시점에서 동적으로 갱신되지 않기 때문에 오류정보가 포함될 수 있다는 문제점이 있다.

본 논문에서는 유사한 선호도를 보이는 사용자를 대상으로 군집분석을 수행함으로서, 이웃 사용자를 선택하는 과정을 단순화할 수 있고, 또한 베이지안 학습을 이용하여 사용자의 선호도를 동적으로 갱신할 수 있는 알고리즘을 설계하고 구현하였다. 사용자의 키워드가 입력되는 순간 사전 데이터와 사후 데이터가 선호도 확률에 동적으로 반영됨으로써 오류정보를 최소화한다. 이렇게 설계된 시스템은 실험을 통해 웹 도서 추천시스템에 적용되어 사용자의 만족도를 증가시킬 수 있음을 보인다.

ABSTRACT

Collaborative filtering method for personalization can suggest new items and information which a user hasn't expected. But there are some problems. Not only the steps for calculating similarity value between each user is complex but also it doesn't reflect user's interest dynamically when a user input a query.

In this paper, classifying users by their interest makes calculating similarity simple. We propose the algorithm for readjusting user's interest dynamically using the profile and Bayesian learning. When a user input a keyword searching for a item, his new interest is readjusted. And the user's profile that consists of used key words and the presence frequency of key words is designed and used to reflect the recent interest of users.

Our methods of adjusting user's interest using the profile and Bayesian learning can improve the real satisfaction of users through the experiment with data set, collected in University's library. It recommends a user items which he would be interested in.

Key Words : Bayesian learning, cluster analysis, recommendation system

1. 서 론

대다수 웹 사이트는 기존 사용자들의 사전 프로화일 정보, 상품 검색 및 구입 관련 정보를 기반으로 사용자의 기호에 맞는 상품이나 기타 정보를 정적으로 추천하고 있다. 이런 추천 시스템의 방법으로 협업 필터링 방식을 사용하면 다른 사용자들의 평가 정보를 이용하여 동적으로 연관 링크를 제공할 수 있는 장점이 있다[3, 4]. 그러나 이는 현재 사용자와 유사한 선호도를 갖는 사용자를 선택하기 위해 전체 사용자와의 해당 아이템별 유사도 계산을 수행해야 하는 과정의 복잡성과 이로 인한 비효율성의 문제가 있다. 또한, 정적인 사용자 프로화일은 시간이 지남에 따라 그 효율성이 감소하여 사용자에게 오류 정보를 제공할 확률이 높다[4, 5]. 따라서, 프로화일 정보를 기반으로 형성된 사용자의 선호도 정보는 동적으로 갱신되어야 할 필요성이 있다. 이러한 문제를 해결

접수일자 : 2002년 7월 1일
완료일자 : 2002년 9월 30일
본 연구는 2002학년도 김포대학의 연구비 지원에 의하여 연구되었음.

하기 위해 본 논문에서는 [알고리즘 1]과 같은 사용자들의 군집화와 베이지안 학습에 의한 사용자 선호도 갱신 알고리즘을 제안한다. 이는 전체 사용자가 아닌 군집의 대표 사용자와의 유사도 측정을 통하여 유사도가 높은 군집내에서의 사용자 선택이 이루어지도록 함으로써 계산 과정을 단순화한다. 또한 베이지안 학습을 이용하여 각 사용자로부터 특정 키워드 입력시, 해당 사용자의 키워드별 선호도 정보가 동적으로 갱신되도록 설계한다.

[알고리즘 1] 베이지안 학습에 의한 사용자 선호도 갱신 알고리즘

- 단계1 : 각 사용자들의 프로화일을 이용해 군집화를 수행한다.
- 단계2 : 키워드(도서명, 분류번호)를 이용해 각 군집내의 키워들에 대한 선호도를 계산하다.

$$P(C_j)=\frac{\sum_{i=1}^{n} f_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{k} f_{i,j}},$$

where i=1,2,...,n and j=1,2,...,k and $f_{i,j}$: 대출빈도수

- 단계3 : 각 군집내의 사용자들의 각 키워드별 선호도를 계산한다.

$$P(C_j|U_i)=\frac{P(C_j,U_i)}{P(U_i)}, \text{ where}$$

$$P(C_j,U_i)=\frac{f_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{k} f_{i,j}} \text{ and}$$

$$P(U_i)=\frac{\sum_{j=1}^{k} f_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{k} f_{i,j}} \cong P(C_{j,i})$$

즉, $C_{j,i}$ 는 다음과 같이 계산된다.

$$P(C_{j,i})=\frac{f_{i,j}}{\sum_{i=1}^{n} f_{i,j}}$$

- 단계4 : 특정 사용자의 키워드 입력으로 인한 해당 키워드 선호도를 베이지안 학습에 의해 갱신한다.

$$P(C_{j,i}|X_i)=\frac{P(C_{j,i},X_i)}{P(X_i)}=\frac{P(C_{j,i})P(X_i|C_{j,i})}{P(X_i)}$$

, where $P(X_i)=\sum_{j=1}^{k} P(C_{j,i})P(X_i|C_{j,i})$

- 단계5 : 키워드를 입력한 사용자와 군집내의 유사한 키워드 선호도를 갖는 사용자와의 협업 필터링을 통하여 동적으로 도서를 추천한다.

2. 군집분석과 베이지안 학습을 이용한 선호도 갱신

기존의 상품 추천 및 자료검색 시스템에서는 사용자 프로화일의 해당 정보에서 각 아이템에 대한 선호도가 현재 사용자의 키워드 검색 패턴을 반영하지 못한채 사전 데이터 값으로만 계산된 정보를 이용한다[6]. 이는 사용자의 특정 아이템에 대한 키워드 입력을 통한 선호도 행위가 발생한 시점 후의 해당 아이템에 대한 선호도, 즉 사후 데이터를 반영하지 못함으로써 많은 오류 발생률을 내포하고 있다. 이러한 문제를 해결하기 위해 본 논문에서는 사용자 프로화일의 사전 데이터로 구성된 선호도 테이블의 정보를 동적으로 갱신하기 위해 베이지안 학습 방법을 이용하여, 사후 데이터가 사용자의 선호도에 정확히 반영될 수 있도록 [알고리즘 1]과 같이 사용자 선호도 갱신 알고리즘을 설계하였다. 또한 유사한 선호도를 갖는 사용자들은 군집화를 수행하여 각각의 군집들이 갖는 특성을 용이하게 파악할 수 있도록 하였다.

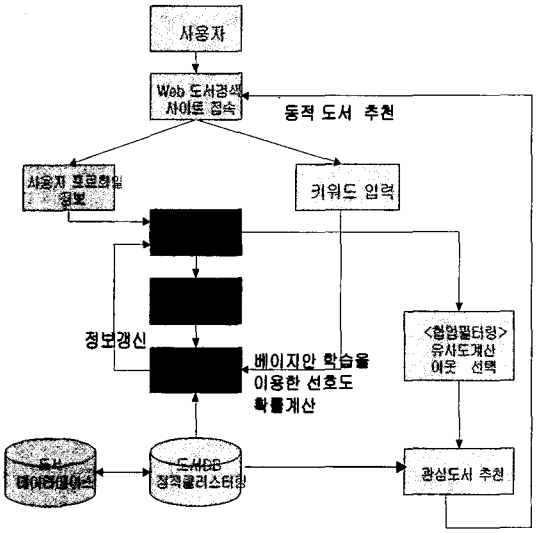
군집화를 수행하는 이유는 현재 사용자와 유사한 선호도를 갖는 다른 이웃한 사용자를 선택함으로서 관련 아이템을 예측하기 위해 전체 사용자를 탐색 대상으로 설정하기보다는 사용자가 소속되어 있는 군집만 탐색하도록 함으로써 검색의 소요 시간을 단축하는데 있다.

3. 웹 도서 동적 추천 시스템의 설계

[그림 1]은 웹 도서 동적 추천 시스템의 전체적인 구성도로서, 사용자 프로화일을 기반으로 베이지안 학습에 의한 협업 필터링 모듈로 구성된다. 여기서, 도서 데이터베이스는 범주별로 분류되어 있다. 이를 바탕으로 사용자의 입력 키워드에 의해 선호도 테이블을 대상으로 동적 갱신 알고리즘이 적용된다.

3.1 도서분류표와 매트릭스

본 논문에서의 웹 도서 동적 추천시스템은 컴퓨터관련 도서로 그 대상 범위를 제한하고 있다. 도서분류 방법은 듀이의 10진 분류 체계를 이용하며, 도서 분류표에 기록된 각 도서명들은 이를 대표할 수 있는 고유의 분류 번호를 갖게 된다. 분류 번호는 도서의 하위 범주와 상위 범주를 쉽게 구분할 수 있도록 위계성을 갖도록 표시된다. 따라서, 각 도서 범주별 사용자의 군집화가 수행되면 사용자가 소속되어 있는 도서의 상·하위범주를 쉽게 구별할 수 있으므로 도서 추천의 범위를 사용자의 선호에 맞게 제한할 수 있다.



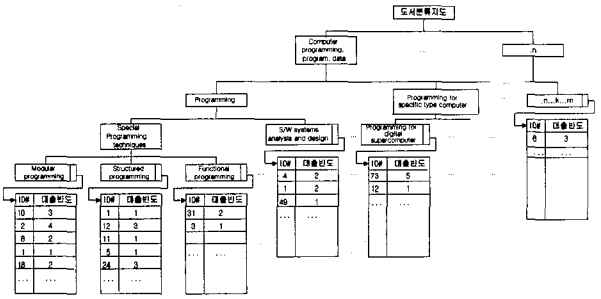
[그림 1] 전체 시스템 구성도

사용자 프로화일에는 대출한 도서의 분류 번호와 대출 빈도수, 검색에 사용된 키워드가 저장되어 있어 선호도 정보에 대한 자료로 이용된다. [그림 2]는 도서 데이터베이스를 도서

분류표로 나타낸 것이다. [그림 3]은 도서 분류별 해당 도서를 대출한 사용자의 학번과 대출 빈도수에 관한 정보를 매트릭스 형태로 구성한 것이다.

005 Computer programming, program, data
.1 Programming
.11 Special programming techniques
.111 Modular programming
.112 Structured programming
.113 Functional programming
.114 Object-oriented programming
.115 Visual programming
.12 Software systems analysis and design
.13 Programming languages
.131 Symbolic(Mathematical) logic
.132 Specific programming languages
.14 Verification, testing, measurement, debugging
.15 Preparation of program documentation
.2 Programming for specific types of computers, for specific operating systems
.21 Programming for digital supercomputers
.22 Programming for digital mainframe computers
<중 략>

[그림 2] 도서 분류표



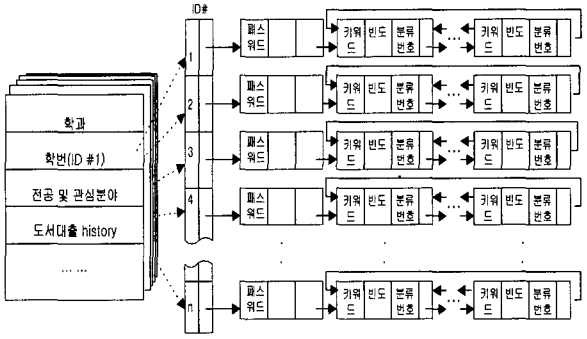
[그림 3] 도서 분류 메트릭스 구성도

[그림 4]는 본 논문에서 설계한 사용자 프로화일 구조로서, 가장 최근의 도서 검색에 사용되었던 질의어들 및 이에 대한 축적된 사용 빈도로 구성되었다.

초기의 사용자 프로화일은 사용자의 학과, 학번, 전공 및 관심분야 등을 사용자가 스스로 작성하도록 하여 사용자에 관한 초기 선호도 정보를 나타내도록 하였다. 그러나 이러한 단순 자료만으로는 유동성있는 사용자의 선호도를 동적으로 반영하지 못하므로, 사용자가 도서 검색 사이트에 접속하여 특정 도서에 대한 검색 및 대출 행위를 할 경우 그 도서에 대한 대출 빈도를 선호도 정보로 사용한다. 사용자 프로화일의 빈도정보는 항목에 대한 관심도를 반영하기 위해 우선 순위의 순서로 정렬된 연결 리스트의 형태로 구성하였다.

[그림 4]에서 사용 빈도가 높은 단어들은 최근 사용자의 검색 패턴을 잘 반영한다고 가정하며, 사용자 접속이 이루어지는 시점에서 사용 빈도에 따라 동적으로 내림차순에 따라 정렬된다. 분류 번호는 사용자가 검색한 도서를 나타내고, 도서 분류 매트릭스와 연결되어 사용자의 패턴을 발견할 수 있는 자료가 된다. 사용자 프로화일 정보는 각 도서 범주별 선호도 테이블 생성에 대한 입력 자료가 되어 군집화를 수행하는데 사용된다.

본 논문에서는 사용자 프로화일에 나타난 사용자의 명시적(explicit)인 데이터와 사용자의 선호를 나타내는 웹사이트에서의 클릭 패턴이나 도서 검색 패턴, 대출 도서 선택 등을 웹 로그나 도서대출 이력 데이터에서 발견하여 특정 도서에 대한 사용자의 선호도를 예측함으로써 신뢰도를 향상시켜 개인화 정보를 효율적으로 지원할 수 있도록 설계하였다.



[그림 4] 웹도서 추천시스템을 위한 사용자 프로화일

3.2 사용자 프로화일을 이용한 군집화 수행

본 논문에서는 도서 분류 번호를 이용한 코사인 기반의 유사도 계산 결과를 바탕으로 군집화를 수행한다. 동일 계층을 이루는 도서 분류 번호의 도서를 대출한 사용자들은 동일 군집의 원소가 된다.

[표 1]은 사용자의 도서 대출 현황을 도서 분류 번호를 이용해 나타낸 것이다.

[표 1]에서 사용자가 대출한 도서의 분류 번호를 이용해 분류 번호를 기준으로 군집화를 수행하면 [표 2]에서와 같이 G1, G2, .., G6의 6개의 군집이 형성된다.

[표 1] 사용자 대출 현황 테이블

사용자	도서분류번호	사용자	도서분류번호
U1	001.23	U11	003.6
U2	004.71	U12	005.216
U3	003.119	U13	004.397
U4	001.5	U14	001.236
U5	008.627	U15	005.17
U6	008.1	U16	003.62
U7	005.34	U17	001.47
U8	001.257	U18	001.534
U9	002.412	U19	003.1
U10	003.417	U20	004.32

[표 2]에서 G1은 도서 분류 번호 001.로 시작하는 동일 도서 범주에 포함되는 도서를 대출한 사용자들 {U1, U4, U8, U14, U17, U18}로 구성된다. 따라서, 동일 군집에 속한 사용자들은 유사한 선호도를 보이고 있음을 알 수 있다. 만약, 새

로운 사용자 U21에게 도서를 추천할 경우 군집화를 수행하지 않으면 새로운 사용자는 전체 20명의 사용자와 개별적으로 유사도를 계산하여 이중에서 유사도가 높은 사용자를 선택해야 한다. 여기서, 전체 사용자가 대단위일 경우는 유사도 계산 작업량이 매우 많아지게 된다. 그러나 기존의 사용자에 대한 군집화를 먼저 수행하고, 각 군집에 대한 평균 정보를 얻게 되면 전체 사용자가 아닌 기존의 생성된 군집과 새로운 사용자의 유사도를 계산함으로써 유사도 계산 과정이 단순화된다.

[표 2] 도서 분류 번호를 이용한 사용자 군집 테이블

도서분류번호	사용자	군집
001.~	{U1, U4, U8, U14, U17, U18}	G1
002.~	{U9}	G2
003.~	{U3, U10, U11, U16, U19}	G3
004.~	{U2, U13}	G4
005.~	{U7, U12, U15}	G5
008.~	{U5, U6, U20}	G6

사용자 프로화일 정보로부터 선호도 테이블을 생성한 후 이를 이용한 군집화를 수행하는데, [표 3]은 사용자 선호도 테이블의 한 예시를 나타낸다.

[표 3] 사용자 선호도 테이블

도서범주 / 사용자	C1	C2	C3	C4	TOT
U1	1	2	1	3	7
U2	2	1	1	2	6
U3	2	1	1	2	6
U4	1	2	1	2	6
TOT	6	6	4	9	25

[표 3]은 사용자 U1, U2, U3, U4가 도서범주 C1, C2, C3, C4를 대출한 대출 빈도수, 즉 선호도에 대한 정보를 나타낸다. 예를들어, 사용자 U3는 도서범주 C1, C2, C3, C4에 대해 각각 2, 1, 1, 2의 대출빈도 정보를 보이고 있다는 것을 의미한다. 사용자 프로화일의 대출빈도를 기반으로 생성된 선호도 테이블은 현 사용시점 이전의 축적된 과거 정보를 이용한 것이다.

베이지안 학습에 의한 개인의 선호도 갱신시 빈도가 없는 곳의 확률값은 0이 되기 때문에 최초의 선호도 테이블을 작성할 때 대출빈도는 모든 사용자가 모든 분야에 관해 동일한 관심을 가지고 있다는 전제 조건하에 1로 설정하였다.

선호도 테이블을 이용하여 각 범주(C1, C2, C3, C4)에 대해 코사인 기반 유사도 기법을 이용하여 각각의 사용자에 대한 유사도 계산을 수행하며, 기준치 이상의 높은 유사도를 갖는 사용자들로 군집을 형성한다. 예를 들어, 기준치가 0.9일 경우 그 이상의 유사도 값 1을 갖는 {U2, U3}가 C1에 대해 하나의 군집으로 형성된다. {U2, U3}는 C2에 대해, {U2, U3, U4}는 C4에 대해 형성된 군집이다. 여기서 대출 빈도수의 하한치를 정함으로써 정보의 가치성을 높일 수 있다.

3.3 군집내 선호도 계산

각 군집내의 키워드별 선호도를 계산하는 목적은 군집의 특성을 파악하기 위함이다. 식 (1)은 특정키워드에 대한 군집내의 선호도를 구하는 식으로써, 총 개수는 (총군집 * 총키워드수)로 표현될 수 있다.

$$P(\text{특정키워드선호도}) = \frac{\text{특정키워드의대출빈도수}}{(\text{해당군집내의대출빈도수} \times \text{키워드수})} \quad (1)$$

[표 1]에서 임의의 군집이 사용자 {U1, U2, U3, U4}로 구성되었다고 가정할 경우에, 각 키워드별 선호도 확률 테이블은 [표 4]와 같다.

[표 4] 군집의 선호도 확률 테이블

도서분류 / 군집	C1	C2	C3	C4
{U1, U2, U3, U4}	P(C1)=6/25 =0.24	P(C2)=6/25 =0.24	P(C3)=4/25 =0.16	P(C4)=9/25 =0.36

사용자 {U1, U2, U3, U4}로 구성된 군집은 [표 3]의 선호도 테이블 정보를 통해 도서범주 C1, C2, C3, C4에 대한 총 대출 빈도수가 25이고, 각각의 도서범주(키워드)에 대한 총빈도수는 각각 6, 6, 4, 9임을 알 수 있다. 따라서, 선호도 확률값은 각각 0.24, 0.24, 0.16, 0.36이다. 이는 C4에 대한 선호도 확률값이 가장 높음으로 군집{U1, U2, U3, U4}는 도서범주 C4에 대한 선호도가 높고, 또한 군집의 원소를 이루는 사용자들도 동일하게 C4에 대한 선호도가 높음을 예측할 수 있다.

키워드별 유사도가 높은 사용자들로 구성된 군집에서 각 사용자별 선호도 확률값을 나타내는 테이블은 [표 5]와 같다.

[표 5] 군집내의 사용자별 선호도 확률값 테이블

도서범주 / 사용자	C1	C2	C3	C4
U1	P(C1|U1)= 1/7 =0.41	P(C2|U1)= 2/7 =0.29	P(C3|U1)= 1/7 =0.41	P(C4|U1)= 3/7 =0.43
U2	P(C1|U2)= 2/6 =0.33	P(C2|U2)= 1/6 =0.17	P(C3|U2)= 1/6 =0.17	P(C4|U2)= 2/6 0.33
U3	P(C1|U3)= 2/6 =0.33	P(C2|U3)= 1/6 =0.17	P(C3|U3)= 1/6 =0.17	P(C4|U3)= 2/6 =0.33
U4	P(C1|U4)= 1/6 =0.17	P(C2|U4)= 2/6 =0.33	P(C3|U4)= 1/6 =0.17	P(C4|U4)= 2/6 =0.33

군집내의 사용자들에 대한 키워드별 선호도 계산은 식 (2)와 같이 나타낼 수 있다.

$$P(\text{키워드선호도}|\text{특정사용자}) = \frac{\text{특정사용자의 특정키워드 대출빈도수}}{\text{특정사용자의 총대출빈도수}} \quad (2)$$

개인별 선호도 확률 테이블의 값을 통해 특정 사용자의 특정 키워드에 대한 선호도와 다른 사용자들과의 선호도를 상대적으로 비교해 볼 수 있으며, 임의의 기준치값을 두어 기준치 이상의 확률 정보를 가지고 있는 동일 군집에서의 다른 사용자의 도서 선호 패턴을 통해 현 사용자에게 선호도가 높을 것으로 예상되는 도서가 추천되는 것이다.

3.4 베이지안 학습에 의한 사용자 선호도 갱신

각 사용자로부터 특정 키워드(X)가 입력되었을 때, 이 키워드에 대한 사용자 선호도가 동적으로 갱신된다면 최신 정보가 반영된 보다 정확한 선호도 확률을 얻을 수 있다. 이는 베이지안 학습 기법을 통해 사전 확률뿐만 아니라 사후확률 정보를 참조하여 선호도를 계산함으로서 수행될 수 있다.

식 (3)은 특정 키워드가 입력되었을 경우 해당 키워드 선호도의 확률 정보를 계산하기 위한 식이다.

$$P(\text{키워드선호도}|\text{입력키워드}) = \frac{P(\text{사용자의 키워드선호도}) \times P(\text{입력키워드로 인한 비율})}{\sum_{\text{키워드}=1}^{k} P(\text{사용자의 키워드선호도}) \times P(\text{입력키워드로 인한 비율})} \quad (3)$$

베이즈 공식에 의하면 식 (3)은 식 (4)와 같이 표현할 수 있다.

$$P(C_{j,i}|X_i) = \frac{P(C_{j,i}, X_i)}{P(X_i)} = \frac{P(C_{j,i})P(X_i|C_{j,i})}{P(X_i)},$$
$$\text{where } P(X_i) = \sum_{j=1}^{k} P(C_{j,i})P(X_i|C_{j,i}) \quad (4)$$

식 (4)를 이용해서, [표 3]의 사용자 U3가 도서범주 C4에 대한 대출 횟수가 4회 증가하여 총 6회의 대출 빈도값을 갖는다고 하면, 사전 데이타는 2, 사후 데이터는 6이고, 전체 대출 빈도수는 10이 된다.

베이지안 학습은 사용자의 행위에 의해 변화된 사후 데이터를 반영하여 사용자의 선호도 확률을 계산한다. 사용자(U3)가 키워드(X=C4)를 입력하였을 때 베이지안 학습에 의해 갱신된 키워드 C1에 대한 선호도 확률은 식 (5)와 같이 계산된다.

$$P(C_1|X=C_4) = \frac{P(C_1)P(X=C_4|P(C_1))}{\sum_{j=1}^{4} P(C_j)P(X=C_4|C_j)} \quad (5)$$
$$= \frac{0.33 \times \frac{2}{10}}{0.33 \times \frac{2}{10} + 0.17 \times \frac{1}{10} + 0.17 \times \frac{1}{10} + 0.33 \times \frac{6}{10}}$$
$$\approx 0.22$$

식 (5)와 같은 방법으로 전체 키워드에 대한 사용자 선호도 확률을 구할 수 있다. 사용자 U3에 대해 특정 키워드 입력(C4)에 의한 갱신된 선호도 확률값과 갱신되기 전의 확률값을 비교하면 [표 6]과 같다.

[표 6] 갱신되기 전 · 후의 사용자 선호도 확률 비교표

도서범주 / 사용자U3	C1	C2	C3	C4
갱신전 선호도 확률	0.33	0.17	0.17	0.33
갱신후 선호도 확률	0.22	0.11	0.11	0.67

[표 6]을 보면 사용자가 특정 키워드를 입력했을 때 특정 키워드의 선호도 확률값은 사전 확률보다 상향 조정되었고, 그외 다른 키워드들은 상대적으로 하향 조정되어 사용자의 전체 데이터에 대해 현재의 선호도 패턴이 반영되었음을 알 수 있다.

사용자가 소속된 군집의 선호도 확률도 갱신되었음을 [표 7]을 통해 알 수 있다. 사용자의 선호도 변경값은 그 사용자가 속해있는 해당 군집의 선호도 확률에도 반영된다.

[표 7] 군집의 갱신 선호도 확률 테이블

도서분류 / 군집 {U1,U2,U3,U4}	C1	C2	C3	C4
갱신전 선호도 확률	0.24	0.24	0.16	0.36
갱신후 선호도 확률	0.21	0.21	0.14	0.45

군집 벡터내의 특정 도서에 대한 선호도를 재계산하기 위해서는 군집내에서의 특정 도서에 대해 선호도를 보이고 있는 사용자들의 선호도와 사용자가 군집에 속할 확률을 곱하여 합계한 후, 특정 도서에 대한 선호도를 보이는 사용자들이 해당 군집에 속할 확률의 합으로 값을 나누어 새로운 도서에 대한 선호도를 구한다. 이렇게 계산된 군집의 대표 벡터를 새로운 복합 사용자로 간주하여 특정 사용자와 복합 사용자간의 코사인 기반 유사도 측정을 통해 이웃을 구한다. 최종적으로 선택된 군집내에서, 현재 사용자의 확률 이상의 값을 갖는 사용자보다 높은 선호도를 보인 도서 목록을 추천한다.

4. 실험 및 평가

4.1 실험

본 논문의 실험을 위해 인하대학교 중앙도서관에서 학생들의 도서 대출 현황 중, 2002년 3월1일부터 6월30일까지의 컴퓨터 관련 도서에 관한 총 4,644건에 대해 10회 이상의 대출 정보를 보이는 98명에 대한 941건의 자료를 추출하였다. [표 8]은 941개의 사용자 도서 대출 현황의 일부를 나타낸다.

[표 8] 실험에 사용된 사용자 대출현황 데이터

학 번	성 명	도 서 명	저 자	대출일
22011678	이윤	C프로그래밍 700제	정병건	20010802
22001282	박칙	포토숍 채널 작업	비드니, 데이빗	20010820
21991333	유도영	Java for engineers&scientists	Chapman, Stephen J	20010821
11940871	김수학	프로그래밍언어론	세베스타, 로버트더블	20011005
11980079	김용훈	SQL server bible ver 7.	권병희	20011005
11990070	김은석	구조적 C언어 프로그래밍	김종교	20011005
22011360	김정윤	윈도두 98바이블&비밀	심슨, 알렌	20011005
12001025	김진언	실습중심 UNIX시스템 개론	이동호	20011005
·	·	·	·	·
·	·	·	·	·

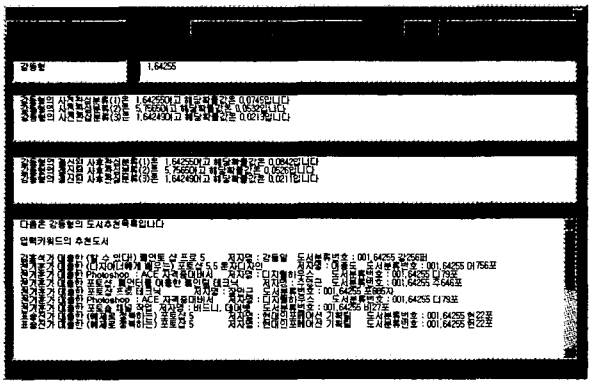
웹도서 동적 추천을 위한 알고리즘은 Pentium III 450MHz, 256MB RAM 환경에서 Visual C++ 6.0으로 구현하였다.

먼저, 코사인 기반 유사도 측정을 이용한 사용자 군집을 수행한 결과, 각 군집에서의 사용자들의 도서 대출 현황은 [표 9]와 같다.

[표 9] 사용자 군집 수행후 각 군집별 대출현황

<군집1>				
학 번	성 명	도 서 명	저 자	대 출 일
12010008	강민형	(알기쉬운)C언어	고작의명	20011030
12010008	강민형	볼랜드 C++라이브러리	강성국	20011011
11951698	김민수	(초보자용)리눅스 프로그래밍	매튜,네일	20011102
11980523	김형석	Turbo-C언어	김광희	20010924
...	...	...		
<군집17>				
학 번	성 명	도 서 명	저 자	대 출 일
11940362	김현철	OS제작의 정석	오재준	20010927
11981590	박소현	UNIX/LINUX 커널의 설계및구현	이형봉	20010920
...	...	...	...	...

로그인한 임의의 사용자가 특정 키워드를 입력하였을 때, 로그인한 사용자의 특정 키워드 입력전과 입력 후의 갱신된 선호도 확률값은 [그림 5]와 같이 계산되어 해당 추천도서 목록을 보여준다.



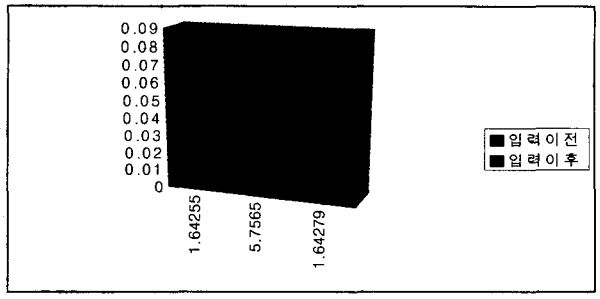
[그림 5] 특정 키워드 입력 전, 후의 갱신된 선호도 확률값

[그림 5]의 예로, 시스템에 로그인한 사용자(강동형)가 특정 키워드(포토샵)를 입력했을 때 선호도 갱신 알고리즘에 의해 사후 입력 데이터가 반영된 상위 3개의 선호도 확률값을 갱신하기 전의 값과 비교하면 [표 10]과 같다.

[표 10] 사용자 선호도 확률값 갱신 전,후 비교표

도서분류 번호	...	1.64249	...	1.64255	...	5.7565	...
갱신전 선호도 확률	...	0.02123	...	0.0745	...	0.0532	...
갱신후 선호도 확률	...	0.02110	...	0.0842	...	0.0526	...

상위 3개의 선호도 분야에 대한 확률값 갱신 전, 후를 그래프로 나타내면 [그림 6]과 같다.



[그림 6] 상위 3개의 선호도별 확률값 비교 그래프

[표 11]은 사용자(강동형)의 입력 키워드(포토샵)를 기반으로 동일 군집내에서의 유사한 사용자들이 선호하는 도서를 추천한 결과를 나타낸다.

[표 11] 입력 키워드에 대한 추천 도서

No.	서명	저자	청구기호
1	(할수있다!) 페이트 샵 프로5	강동일	001.6125 김256퍼
2	(디자이너에게 배우는) 포토샵5.5 문자디자인	이종도	001.64255 이756포
3	Photoshop:ACE 자격증대비서	디지털하우스	001.64255 디79포
4	포토샵,페인터를 이용한 페인팅 테크닉	주영근	001.64255 주646포
5	포토샵 프로 테크닉	장민근	001.64255 포885자
6	포토숍 채널작업	비드니, 데이빗	001.64255 비27포
7	(예제로 정복하는)포토샵5	현대인포 메이션기획팀	001.64255 현22포

또한, [표 12]는 로그인한 사용자와 유사한 선호도를 갖는 이웃 사용자가 로그인한 사용자의 입력한 특정 키워드 외의 관심 도서를 대출한 도서 목록을 나타낸다.

[표 12] 입력키워드 외의 추천 도서

No.	서명	저자	청구기호
1	포토샵 아트갤러리	이경훈	006.6869 이146포
2	포토숍5/5.5 와우!북	데이톤,리니아	006.6869 데69포
3	(깜찍한 캐릭터예제로 배우는)포토샵6.0디자인북	윤여민	006.6869 포885우
4	(웹디자이너에게 배우는) 포토샵6 그대로 따라하기	정민철	006.6869 정3931포
5	3D Studio MAX3.1&hint	성승욱	006.693 서586스

4.2 성능 평가

본 논문의 웹 도서 추천 시스템은 현재 웹 정보 추천 시스템에서 많이 사용되고 있는 메모리 기반 알고리즘인 피어슨의 상관 계수 알고리즘과 비교하였다[1]. 실험에 사용된 98명의 학생들에게 웹 도서 추천시스템에 의한 방법을 적용하여 추천한 도서 서비스에 대한 피드백을 보인 65명의 학생을 대

상으로 만족도 정보를 이용하여 성능을 평가하였다.

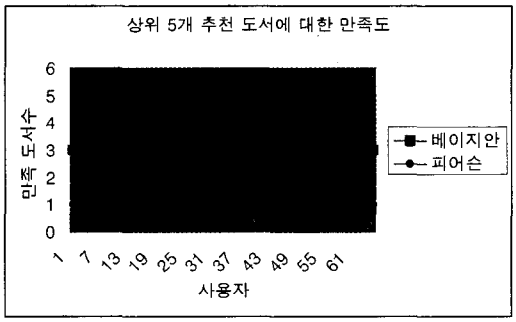
두 방법간의 만족도의 차이는 [표 13]과 같다. 이 만족도는 각 사용자가 가장 필요할 것으로 기대되는 최상위 1권의 책을 추천하였을 때 각 사용자가 추천 도서에 대한 만족 여부를 피드백한 결과이다.

[표 13] 사용자 만족도 평가

	베이지안 학습의 선호도 갱신 알고리즘	피어슨의 상관 계수 알고리즘
만족도	86.15 %	61.54%

피드백을 보인 65명의 사용자 중에서 베이지안 학습의 선호도 갱신 알고리즘에 의한 웹 도서 추천에 의한 정보 서비스에 만족을 보인 사용자는 56명(86.15%)이었고, 피어슨의 상관 계수 알고리즘에 의한 추천 도서에 만족을 보인 사용자는 40명(61.54%)이었다. 따라서 본 논문에서 제안하는 웹 도서 추천 시스템에 의한 만족도가 더 높은 것으로 나타났다.

다음으로는 각 사용자별로 두 개의 비교 알고리즘을 통해 가장 만족할 것으로 기대되는 상위 5권의 도서를 추천하였다. 이 5권의 추천 도서 중에서 각 사용자가 만족한 도서 수를 피드백하여 얻은 결과는 [그림 7]과 같다.



[그림 7] 상위 5개 추천 도서에 대한 사용자 만족도

[그림 7]의 X축은 65명의 각 사용자이고, Y축은 5권의 추천 도서 중 자신에게 필요하다고 만족을 보인 도서의 수이다. [그림 7]에서 보면 대부분의 사용자들은 본 논문에서 제안하는 웹 도서 추천 시스템을 통해 만족한 결과를 얻고 있음을 알 수 있다.

5. 결 론

본 논문에서는 개선된 협업필터링 방법으로 군집화 기법과 베이지안 학습을 이용하여 사용자 프로화일을 동적으로 갱신하였다. 군집화를 통해 유사도가 비슷한 사용자를 군집으로 형성하여 군집별 특성을 나타내는 대표 사용자로 간주하였다. 따라서 사용자는 전체 사용자가 아닌 군집의 대표 사용자와 유사도 계산이 이루어지게 함으로써 수행 과정을 단순화하였다. 또한 베이지안 학습을 이용하여 사용자의 특정 키워드에 대한 입력 시점을 기준으로 사전 확률과 사후 확률을 사용자의 선호도 확률에 반영하여 사용자의 선호도 정도가 동적으로 반영되어 만족도가 증가된 정보를 제공할 수 있도록 하였다.

향후 연구 과제로는, 사용자의 입력 키워드간의 연관성이 고려되고, 군집을 분류하는데 있어서 신경망이나 유전자 알고리즘 등 보다 효율적인 분류 알고리즘을 적용하는 것이다. 또한 추천한 도서에 대한 사용자의 만족도에 대한 피드백 정보를 시스템에 적용하면 더욱 정교한 추천시스템을 만들 수 있을 것으로 기대한다.

참 고 문 헌

[1] Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2001.

[2] Bamshad Mobasher, Hoghua Dai, Tao Luo - Yuqing Sun, Jiang Zhu, "Integrating Web Usage and Content Mining for More Effective Personalization," EC-Web 2000.

[3] M. Pazzani, D. Billsus, Leraning and Revising User Profiles:The Identification of Interesting Web sites, Machine Learning 27, Kluwer Academic Publishers, pp.313-331, 1997.

[4] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," Accepted for publication at the WWW10 Conference. May, 2001.

[5] 김종민, 박영배, "Push기술을 이용한 쇼핑몰 개인 맞춤정보 제공시스템의 설계," *정보과학회* Vol.27 No.2(2) , pp.63~65, 2000.

[6] 이준규, 인터넷 개인화 아이템 추천 알고리즘에 대한 연구, 연세대학교, 석사학위 논문, 2001.

저 자 소 개

최준혁(Choi, JunHyeog)

1990년 경기대학교 전자계산학과 졸업(이학사)

1995년 인하대학교 대학원 전자계산공학과 졸업(공학석사)

2000년 인하대학교 대학원 전자계산공학과 졸업(공학박사)

1997년-현재 김포대학 컴퓨터계열 조교수

관심분야 : 정보검색, 데이터마이닝, 신경망, 유전자 알고리듬 등

E-mail : jhchoi@kimpo.ac.kr

김대수(Kim, DaeSu)

1977년 : 서울대학교 사대수학과 학사

1986년 : 미국 Univ. of Mississippi, Computer Science, M. S.

1990년 : 미국 Univ. of South Carolina, Computer Science, Ph. D.

1991년 - 1993년 : 한국전자통신연구원 컴퓨터연구단 선임연구원

1993년 - 현재 한신대학교 컴퓨터학과 교수
관심분야 : 신경망, 퍼지, 인공지능, 지능시스템, 에이전트, 융합 모델링
Phone : +82-031-370-6784
Fax : +82-031-372-3343
E-mail : daekim@hanshin.ac.kr

임기욱(Rim, KeeWook)
1977. 2. 인하대학교 공과대학 전자공학과 졸업
1987. 2. 한양대학교 전자계산학 석사
1994. 8. 인하대학교 전자계산학 박사
1977-1983 한국전자기술연구소 선임연구원
1983-1988 한국전자통신연구소 시스템소프트웨어 연구실장
1988 - 1989 미 캘리포니아 주립대학(Irvine) 방문연구원
1989 - 1997 한국전자통신연구원 시스템연구부장
주전산기(타이컴) III, IV 개발 사업책임자
1997 - 2000 정보통신연구진흥원 정보기술전문위원
2000 - 현재 선문대학교 산업공학과 교수

관심분야 : 실시간 데이터베이스시스템, 운영체제, 시스템구조 등
E-mail : rim@omega.sunmoon.ac.kr