2001년도 대한전기학회 하계학술대회 논문집 2001.7.18-20

개인화 상품 추천을 위한 해쉬테이블 기반 협력 필터링 에이전트

이은영* , 조동섭
이화여자대학교 컴퓨터학과

Hash Table based Collaborative Filtering Agent for personalized Item Recomme

Eunyoung Lee*, Dongsub Cho
Dept. of Computer Science and Engineering , Ewha Womans University

Abstract - 인터넷은 정보의 바다로 표현할 만큼 방대하며, 이러한 넘치는 정보 속에서 사용자에게 필요한 정보들을 추출하여 사용자들의 효율성과 만족도를 높이는 것이 개인화 정책이고, 결과적으로 전자상거래 사이트에서의 판매의 증가를 이루기 위해 필요한 것이다. 따라서 개개인의 특성에 맞춘 개인화 서비스가 현재의 인터넷에서 제공하는 효율성을 뛰어넘을 수 있는 새로운 해결점으로 주목받고 있다.

본 논문에서는 기존의 협력 필터링(Collaborative filtering) 방법을 개선하여 사용자의 선호도(preference)를 결정하고, 이를 토대로 알맞은 아이템 추천 서비스를 사용자에게 제공하는 해쉬테이블 기반 협력 필터링 에이전트(Hash Table based Collaborative Filtering Agent)를 제안하고자 한다. 이를 통하여 기존의 사용자 또는 처음 방문한 사용자에게도 사이트를 방문하는데 만족도와 효율성을 높이도록 하는 것이 목표이다.

1. 서 론

인터넷은 정보의 바다로 표현할 만큼 방대하며, 이러한 넘치는 정보 속에서 사용자들의 만족도를 높이기 위해 필요한 것이 바로 개인화 정책이다. 따라서 개개인의 특성에 맞춘 개인화 서비스가 현재의 인터넷에서 제공하는 효율성을 뛰어넘을 수 있는 새로운 해결점으로 주목받고 있다.

개인화를 위해서는 인터넷의 방대한 정보 내에서 사용자에게 필요한 정보를 추출해 내기 위한 필터링 작업이 필요하다. 정보의 필터링을 위해서는 보통 사용자 프로파일을 기반으로 하여 필터링 작업을 수행한다. 본 논문에서는 필터링에 에이전트 개념을 도입하여 사용자로부터의 피드백(Feedback)과 사용자의 관심(preference)도를 측정하고 이를 학습하여 사용자의 프로파일의 정교하게 만들게 한다. 이러한 개인화를 효율적으로 제공하기 위해 본 논문에서는 협력필터링(collaborative filtering) 방법을 이용하였다[1]. 즉, 협력적인 필터링 에이전트를 통해 사용자 개개인의 경향분석을 한다. 협력적인 필터링을 사용하는 이유는 다음과 같다.

- Autonomy
- Sociability
- Responsiveness
- Proactiveness

기존의 협력 필터링 에이전트에서는 사용자가 자발적으로 아이템에 대해 사용자의 선호도를 제공하여 사용자의 프로파일을 좀더 정교하게 만들어 필터링의 정확성도를 높인다. 그러면 이러한 사용자들의 프로파일들을 기반으로 정보를 공유하여 협력 필터링을 하도록 한다. 즉, E-Commerce에서 사용자가 아이템들의 일부에 등급(score)을 매겨주면, 그것과 기존의 사용자들이 등급을 매긴 기록을 기반으로 사용자가 등급을 매기지 않은 나머지 아이템들에 대한 등급을 예측하게 된다.

협력 필터링 알고리즘 중 사용자 기반 협력 필터링 알고리즘은 추천 에이전트 시스템에서 가장 많이 사용되는 기술로써 Amazone.com, CDnow 등 상업적으로 성공을 거두고 있는 여러 전자상 거래 사이트에서 적용하고 있다[5]. 하지만 사용자 기반 협력 필터링 알고리즘은 높은 지연시간을 가지는 단점이 있다. 따라서 방대한 데이터들을 실시간으로 빠르게 처리하기 어렵다[4]. 이에 기존 알고리즘의 단점을 해결하여 정확성과 빠른 속도 모두를 제공할 수 있는 협력 필터링 에이전트를 설계하는 것이 필요하다.

따라서 본 논문에서는 기존의 사용자 기반 협력 필터링 알고리즘의 성능을 개선하기 위해 해쉬테이블 기반 협력 필터링 에이전트를 제안하였다. 해쉬테이블 기반 협력 필터링 에이전트는 기존의 협력 필터링 알고리즘에서 데이터를 필터링 하는 과정에 해쉬 모듈을 추가하였다. 기존의 사용자 기반 협력 필터링 알고리즘에 해쉬의 장점을 도입하여 해쉬를 사용함으로써 데이터의 검색 시간을 보장하고, 전체 계산 시간을 효율적으로 감소시킬 수 있도록 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 개인화를 위한 다른 기법들을 설명할 것이며, 3장에서는 본 논문에서 제안하는 해쉬테이블 기반 협력 필터링 에이전트에 대해 설명하고, 4장에서는 결론과 향후연구과제에 대해 논의 할 것이다.

2. 관련 연구

개인화의 방법으로는 규칙기반 필터링(Rules-based filtering), 학습 에이전트(Learning Agent), 협력 필터링(상호협동; Collaborative filtering)의 방법이 있다. 여기서는 본 논문에서 제안하는 협력필터링 외의 다른 개인화 정보를 주기 위한 기존의 다른 필터링 방법에 대해 설명하겠다.

2.1 규칙 기반 (Rule-based) 필터링

규칙기반 필터링은 사용자들에게 몇 가지 질문들을 한 이후에 이 질문들의 답에 적합한 내용들을 전달하는 것이다. 사용자들의 선택에 의한 내용 전달이라는 측면에서 보면 사용자들이 웹에서 제공하는 내용들을 편집하

여 자신의 웹 페이지를 구성할 수 있도록 만들어주는 웹 고객화(Customization)와 차이가 없는 것처럼 보인다. 그러나 규칙기반 필터링에서 제공하는 질문은 사용자들이 자신의 선택을 통해 내용을 스스로 구성하게 하기 위한 목적이 아니라 질문을 통해 사용자들을 구분하고 개개인을 구별하기 위한 목적으로 사용된다는 점에서 고객화와는 큰 차이가 있다. 사용자들을 구분하기 위해 사용되는 질문은 매우 다양한 형태를 가질 수 있다. 가령 사용자의 우편번호를 물어보고 사용자의 거주지를 구분할 수 있고 인적사항 정보와 어떤 사항들에 대한 선호도 등의 정보를 물어보고 그 정보들에 따라 사용자들을 구분할 수 있다[2].

2.2 학습 에이전트(Learning Agent)

학습 에이전트(Learning Agent)는 사용자의 웹 상에서의 활동을 관찰하고 사용자가 어떤 내용에 관심을 가지고 있는지 판단하여 사용자에게 알맞은 내용을 전달하도록 하는 것을 말한다. 사용자의 웹 내에서의 행동 중에서 중요하게 사용되는 것은 특정한 페이지를 보는 시간, 인쇄한 페이지, 전자 상거래 사이트의 경우에는 구매한 상품과 쇼핑 카트에 넣은 상품 등을 들 수가 있으며 이러한 사용자의 행동을 관찰하기 전에 선행되어야 할 것은 웹 페이지의 내용들에 대한 정보가 일정한 기준에 따라 분류되어 있어야 한다는 것이다. 즉 사용자에게 제공될 자료들에 대한 정보(meta-data)가 제공될 내용과 함께 데이터베이스 화 되어 있어야 한다는 것이다. 구축된 내용 데이터베이스와 관찰된 사용자의 웹 사용 습관 토대로 데이터 마이닝의 과정을 거쳐 사용자의 성향과 관심을 결정되고 사용자에게 알맞은 내용이 제공된다. 이러한 과정은 시스템이 사용자의 행동을 바탕으로 사용자의 성향을 학습한다는 점에서 학습 에이전트라고 불리고 있다[2].

3. 해쉬테이블 기반 협력 필터링 에이전트의 구현

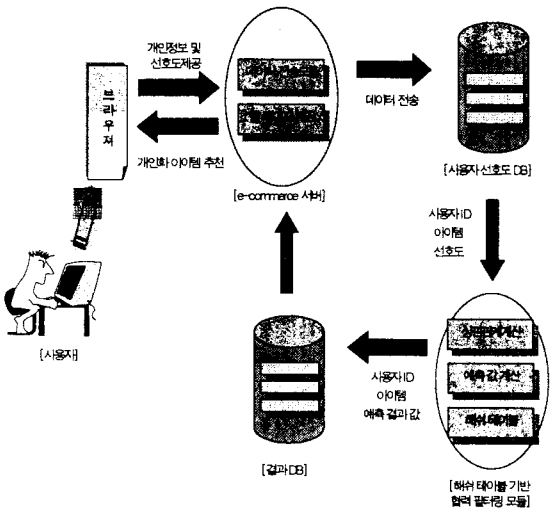
본 논문에서는 해쉬테이블기반 협력 필터링 에이전트를 제안하였다. 해쉬테이블 기반 협력 필터링 에이전트는 기존의 협력 필터링 알고리즘에서 데이터를 필터링하는 과정에 해쉬 모듈을 추가하였다. 따라서 기존의 사용자 기반 협력 필터링 알고리즘에 해쉬의 장점을 도입하여 해쉬를 사용함으로써 데이터의 검색 시간을 보장하고, 전체 계산 시간을 효율적으로 감소시킬 수 있도록 하였다.

전체적인 해쉬테이블 기반 협력필터링 에이전트의 과정은 우선 사용자가 전자상거래 사이트를 방문하여 관심있는 아이템에 선호도를 표시하면 그것을 웹 서버에서 받아 데이터베이스에 저장하고, 이렇게 데이터베이스에 기록된 내용을 기반으로 먼저 데이터의 정제과정을 거치고, 필터링 과정을 거쳐 최종 결과를 다시 데이터베이스에 저장하여 사용자에게 추천할 수 있도록 하는 것이다.

이와 같은 내용은 오른쪽의 [그림1] 과 같이 통합되어 해쉬테이블 기반 협력필터링 에이전트를 구성한다.

3.1 협력 필터링 알고리즘

사용자 프로파일은 사용자로부터 직접 피드백과 사용자의 관심도를 측정하고 이를 학습하여 사용자 프로파일



[그림 1]. Hash Table based Collaborative Filtering Agent

을 더욱 정교하게 만들 수 있게된다. 그러나 사용자 프로파일이 사용자에 대해 개인화 되어 있지 않다면 단순한 cognitive 필터링에 의존할 수 밖 에 없다.

따라서 이러한 사용자 프로파일을 기반으로 정보를 공유한다면 이러한 문제가 어느 정도 해결 될 수 있을 것이다[1].

이러한 협력 필터링은 사용자가 자발적으로 제공한 정보를 사용하여 사용자를 비슷한 선호도를 가진 집단으로 나누어 그 집단 내에서 서로에게 추천하는 방식을 사용한다[1]. E-Commerce에서 사용자가 아이템들의 일부에 등급(score)을 매겨주면, 그것과 기존의 사용자들이 등급을 매긴 기록을 기반으로 사용자가 등급을 매기지 않은 나머지 아이템들에 대한 등급을 예측하는 것이다.

3.2 상관 관계

상관분석(Correlation Analysis)이란 두 변수간에 얼마나 밀접한 관계를 가지고 있는가를 분석하는 방법으로 두 변인간의 상호관련성에 대해 측정하는 통계적 방법이다. 확률변수 X의 평균을 $\overline{X}$ 라고 하면, X의 분포가 중심위치의 측도인 $\overline{X}$ 로부터 떨어진 정도를 나타내는 양으로서 이를 확률변수 X의 표준편차(standard deviation)이라 부른다. 또한, 공분산은 확률변수 X의 증감에 따른 확률변수 Y의 증감의 경향에 대하여 측정한 것으로 공분산 값이 양수이면 두 변수는 같은 방향으로 움직이고, 공분산 값이 음수이면 두 변수는 반대 방향으로 움직이고 있음을 나타낸다. 만약 공분산 값이 0이라면 두 변수간에는 아무런 선형관계가 없으며 두 변수관계가 독립적이라는 뜻이다[6]. 따라서 상관 계수를 구하는 식은 분자는 두 변수간의 공분산이며 분모는 두 변수의 표준편차의 곱인, 즉 상관계수는 공분산을 두 변수의 표준편차로 나눈 표준화된 공분산을 의미한다.

아이템의 등급을 예측하기 위해서는 Correlation Coefficient를 사용한다. 이 값은 1 과 -1 사이의 값을 가지게 되는데 pearson correlation은 다음과 같이 구하게된다[4].

$$W_{a,u}=\frac{\sum_{i\in I_a\cap I_u}(r_{ai}-r_{a*})(r_{ui}-r_{u*})}{\sigma_a\sigma_u} \quad (1)$$

위의 식에서 표현된 각 표기는 다음의 [표 1] 같다.

U	Users
I	Items
R	matrix with element r_{ui}
a	Active user
r_{u*}	average score for each user
r_{i*}	average score for each item
σ	standart deviations of score calculated over Ia∩Iu

[표 1]. Notation

위의 내용을 쉽게 표현하면 다음의 [표 2]와 같다.

user \ item	item 1	item 2	...	item i-1	item i	평균
user 1		2	...	5	3	r_{1*}
user 2	4		...	2	4	r_{2*}
:	:	:	...	:	:	:
user u	2	1	...	1		r_{u*}
평균	r_{*1}	r_{*2}	...	r_{*i-1}	r_{*i}	r_{**}

[표 2]. 표기의 의미

여기서 Correlation coefficient 값이 1이면 perfect positive relationship 이라고 하며, 값이 -1이면 perfect negative relationship 이라고 한다. 만약 값이 0이면 relationship 이 존재할 수도 있고 존재하지 않을 수도 있는 경우가 발생하게 된다. 이것은 사용자와 사용자의 상호관계를 나타내게 된다.

3.3 예측 값의 계산

평가의 예측에는 위의 식(1)에서 구한 상관관계를 포함한 평균값을 이용한다 [1][4]. 예측 값은 다음과 같은 방법으로 구하게 된다.

$$P_{ai}=r_{a*}+\frac{\sum_{u\in U_i}W_{a,u}(r_{ui}-r_{u*})}{\sum_{u\in U_i}W_{a,u}} \quad (2)$$

3.4 해쉬테이블 기반 알고리즘

일반적으로 데이터를 검색하는 방법들은 자료 수 N이 커짐에 따라 평균 검색 길이도 길어진다. 하지만 해쉬 검색법은 자료 수 N에 관계없이 상수의 검색 길이를 가지는 매우 빠른 검색 방법이다. 해쉬 검색법은 메모리 내부에서 검색하는 내부 검색에도 사용되지만 더욱더 강력하게 디스크 상에서도 아주 빠른 속도로 검색해 주는 외부 검색에도 사용되는 알고리즘이다[7].

본 논문에서는 이러한 해쉬 알고리즘의 장점을 기존의 사용자 기반 협력 필터링 알고리즘에 추가하여 기존의 사용자 기반 협력 필터링 알고리즘의 계산 속도를 더 빠르게 했다. 이러한 알고리즘을 해쉬테이블 기반 협력 필터링 알고리즘이라 한다. 해쉬테이블 기반 알고리즘의 설계는 다음과 같다.

1. 해쉬테이블(hash table)을 미리 할당해 둔다
2. 처음 데이터를 읽을 때 hashing을 사용하여 hash table에 저장함.
3. 적당한 영역에다 데이터를 저장하여, 검색할 때에 데이터베이스의 앞에서부터 순차적으로 찾을 필요 없이 단번에 그 데이터의 위치를 알 수 있게 된다.

즉, 해쉬테이블을 사용함으로써 데이터의 검색 시간을 보장하고, 전체 계산 시간을 효율적으로 감소시킬 수 있다.

4. 결 론

본 논문에서 제안한 개인화 추천을 위한 협력 필터링 에이전트는 해쉬테이블 기반 협력 필터링 에이전트를 통하여 사용자 프로파일을 정확하게 필터링 하여 사용자의 만족도를 높일 수 있게 했다. 또한 해쉬알고리즘을 첨가하여 기존의 알고리즘 보다 예측 값의 계산 속도를 좀더 빠르게 했다.

향후 본 시스템을 실제 e-Commerce에 적용하여 실시간 추천을 가능하게 하는 것이 필요하며, 이 해쉬테이블 기반 협력 에이전트를 평가할 수 있는 시스템이 필요하다. 즉, 협력 필터링 알고리즘의 검증 및 분석 결과 검증 단계가 필요할 것이다.

[참 고 문 헌]

[1] 양재영, 최중민: "협동 에이전트 시스템", 전자공학회지 26권 1호, pp 25-33, 1999.

[2] http://www.personalization.co.kr

[3] M. Perkowitz, and O. Etzioni, "Towards adaptive Web Sites : Conceptual Framework and Case Study", Proc. of the The Eighth International World Wide Web Conference,(WWW8), Toronto Canada, May 1999.

[4] S.Vucetic, Z.Obradovic: "A Regression-Based Ap proach for scaling-up personalized recommender systems in E-Commerce ,ACM SIGKDD Interna--tional Conference on Knowledge Discovery and Data Mining Workshop pp.13-21, August 2000

[5] 아이비즈넷, " One-to-One 추천 서비스의 현주소 ", http://www.crmpark.com/, 2000.

[6] 이외숙, 임용빈, 성내경, 소병수, " 통계학 ", 경문사, 199

[7] 이재규, " C로 배우는 알고리즘 ", 세화출판사, 1994.