도서 추천 시스템에 데이터 마이닝 기법의 적용

진승훈*, 김병익*, 김태균*, 김종완*, 김영순**
*대구대학교 컴퓨터정보공학부
**포항1대학 전산정보처리학과
e-mail : GLIDE77@hitel.net

Applying Data Mining Techniques for Book Recommendation System

Seung-Hoon Jin*, Byoung-Ic Kim*, Tae-Kyun Kim*, Jong-Wan Kim*, Young-Sn Kim **
*Department of Computer and Information Engineering, Taegu University
**Department of Computer Information Processing, Pohang College

요 약

도서 정보 추천 시스템에서 기존 사용자들의 정보를 이용하여 마이닝 기법중 군집 분석을 적용하여 사이트에 처음으로 접속하는 사용자와 접속률이 낮아 피드백 정보가 많이 없고 적절한 추천을 하지 못하는 사용자에게 비슷한 군집의 사용자들의 정보를 이용하여 적절한 정보를 추천한다. 본 논문에서는 기존의 멀티에이전트 추천 시스템에 데이터 마이닝 에이전트와 패턴 분석 에이전트를 접목하여 더 나은 추천 정보를 제공하기 위한 시스템을 제안한다.

1. 서론

데이터 마이닝은 대량의 실제 데이터로부터 이전에 잘 알려지지 않는 묵시적이고 잠재적으로 유용한 정보를 추출하는 작업이라고 정의된다.[1] 이러한 데이터 마이닝 기법을 본 연구실에서 구축한 도서 정보 검색 사이트에 적용해 보고자 한다. 단순히 온라인 상에서의 상품판매에만 그치고 있는 전자상거래 사이트에 고객정보와 고객 검색 패턴을 알아내고 이러한 정보를 에이전트가 자동관리하며 불특정 다수를 위한 도서 검색 사이트가 아닌 개개인의 취향과 패턴을 읽어냄으로서, 도서 검색 사이트의 고객에 대한 신뢰와 경비절감 그리고 도서 상거래의 활성화로 이어갈 수 있다.[7]

본 연구에서는 에이전트기반 도서검색사이트와 데이터 마이닝 기술을 접목함으로서 도서추천 시스템의 추천기능을 강화하는 방법을 제시한다.

2. 도서 검색 시스템

본 연구에서는 [2]에서 구축된 전자상거래에서 개인화된 제품 정보 추천을 위한 멀티에이전트 시스템의 워크플로우에 데이터 마이닝 기법을 적용하여 RA(Recommendation Agent)에서의 도서 추천을 보완하고 강화하려고 한다.

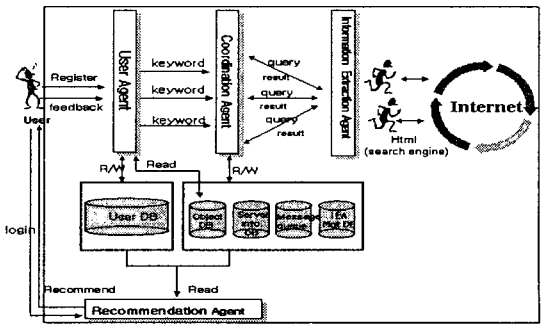

그림 1 제품 정보 추천을 위한 멀티 에이전트 시스템의 워크 플로우

그림 1은 본 연구진이 제안한 멀티에이전트 시스템 워크플로우의 구조이다.

· UA (User Agent)
사용자로부터 기본 정보(아이디, 비밀번호, 이름, 생년월일, 직업, 성별, E-mail)와 관심 분야에 대한 키워드를 입력받는다.

· CA (Coordination Agent)
UA로부터 넘겨받은 키워드를 Message Queue를 통해서 관리하고, IEA(Information Extraction Agent)에게는 URL과 키워드를 Query 형태로 전달하여 정보 검색을 요청한다. IEA로부터 검색 결과를 받으면, 이 결과로서 Server Info DB, Object DB, Message Queue, IEA Mgt DB를 수정하거나 내용을 추가하는 등 항상 최신의 정보로 유지한다.
- Server Info DB : 도서 종합 쇼핑몰 사이트와 출판사 사이트에 대한 정보들이 저장되어 있다. 초기의 정보 수집을 위하여 사용되며 실질적으로 IEA에 전달되어 사용된다. 현재 본 논문에서 교보문고, 삼성 인터넷 서점 크리센스, 정보문화사 사이트를 대상으로 검색이 이루어지고 있다.
- Object DB : 검색된 도서에 대한 도서명, 출판사, ISBN, 출판년도, 가격의 정보들이 저장되어 있으며, 새로운 정보가 발견됨에 따라 도서 정보가 계속 추가된다.
- Message Queue : UA로부터 넘겨받은 키워드와 정보 검색 진행 과정을 관리한다.
- IEA Mgt DB : IEA의 검색 진행 과정을 관리한다.

· RA (Recommendation Agent)
User DB와 Server Info DB, Object DB를 참조하여 각 사용자에게 개인화 된 정보를 추천한다. 추천 정보들은 키워드를 중심으로 가격, 출판년도, 출판사의 조건들을 분석하여 선호도가 높은 순서대로 도서 정보가 제공될 뿐만 아니라 사용자가 직접 다른 키워드를 이용하여 검색할 수 있도록 지원한다.

3. 시스템 제안

그림 2와 같이 기존의 멀티에이전트 시스템 기반에 2개의 새로운 에이전트 모듈이 추가되며 User Agent에서 사용자 패턴 감지 모듈이 추가되어 사용자가 검색을 하거나 관련된 서적의 사이트에서의 행동을 체크하여 사용자 패턴 분석 에이전트로 정보를 넘겨준다. User Agent 와 User DB 의 경우 기존의 시스템에 있는 것을 수정해야되며 나머지 모듈들은 새롭게 제안되는 모듈이다.

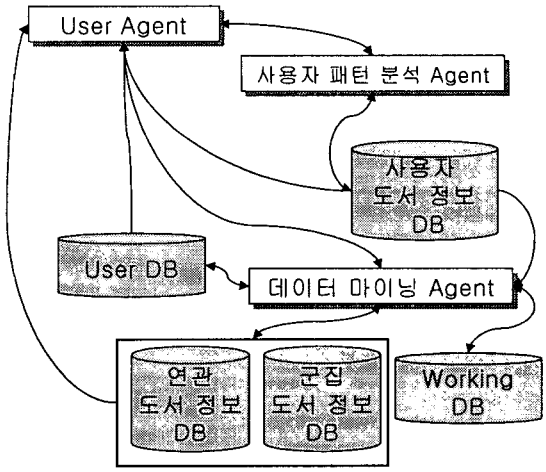

그림 2 제안된 시스템 모듈

· 사용자 패턴 분석 Agent
: UA에서 사용자의 행동 패턴을 감지하여 사용자가 읽은 웹 페이지나 검색한 색인어 등을 본 Agent에 보내면 본 Agent는 정보를 분석하여 적절한 카테고리로 분류하여 사용자 도서 정보 DB에 적절한 값을 적용한다. 예를 들어 사용자가 "자바" 카테고리의 책을 검색하였을 경우 "자바" 카테고리의 가중치를 1 증가시키고, 구입을 하기 위해 주문 페이지로 이동하였을 경우 가중치를 5 증가시키는 것이다.

· 사용자 도서 정보 DB
: 각 사용자별로 본 시스템에서 분류한 카테고리를 가지고 적절한 값을 변화해가며 사용자의 도서 정보를 관리하는 DB이다.

문학 : 영문학, 일문학, 프랑스문학, 독문학, 국문학 …
컴퓨터 : 자바, 인터넷, 오피스, C++, 리눅스, asp …
어 학 : 영어, 일본어, 프랑스어, 독일어, 한국어 …
⋮

위와 같이 카테고리를 분류하고 순서대로 순번을 부여한다. "알파벳" + "숫자" 형으로 이루어지며 알파벳의 경우는 대분류 카테고리를 뜻하며 숫자의 경

우 소분류 카테고리를 뜻한다.

<표 1> 사용자 도서 정보 DB

	A1	A2	A3	.....	D1	D2	.....	J5
진승훈	1	1	2	.....	2		.....	
김혜영	2		1	.....			.....	
김병익		3	3	.....		8	.....	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

표 1은 사용자 패턴 분석 Agent가 사용자가 검색하거나 클릭한 도서 카테고리의 가중치를 분석 저장한 정보이다.

· 데이터 마이닝 Agent

: 본 Agent 에서는 2개의 계층을 가지고 있다. 먼저, 하위계층의 경우 새로운 사용자가 들어왔을 때 즉각적인 반응을 하는 것으로 새로운 사용자의 경우 군집이 나누어져 있지 않다. 그렇다고 사용자가 추가 될 때마다 군집분석을 시행하기에는 무리가 따른다. 이 부분을 해결하기 위해 Working DB 에서는 군집 분석을 하는 중간 단계의 사용자간의 유사도 정보와 각 그룹의 중심값을 가지고 있다. 군집 분석을 완료하면 각 그룹의 도서 정보를 군집 도서정보 DB에 저장하고 각 카테고리의 연관 관계를 분석하여 연관 도서 정보 DB에 저장하여 두었다가 User Agent가 사용자에게 추천 할 때 이 정보를 활용한다.[10]

4. 제안된 시스템의 데이터 마이닝 기법 적용

데이터 마이닝의 여러 기법들 중에서 본 시스템에서는 각 도서의 연관관계를 분석하기 위하여 연관 규칙을 사용하고 사용자들의 효율적 관리 및 도서 추천을 위하여 군집분석기법을 이용하여 사용자 도서 검색 패턴을 분석하여 군집을 나눈다.

4.1 카테고리간의 관계 파악을 위한 연관 규칙

연관 규칙을 통하여 각 카테고리간의 관계를 판단하여 사용자들이 관심 있는 카테고리를 검색할 때 검색한 카테고리와 연관된 카테고리를 추천하여 준다.

이를 위해 본 연구에서는 사용자 정보 DB의 값을 이용한다. 하지만 시간이 지날수록 각 사용자의 모든 카테고리의 가중치가 1 이상이 될 확률이 높다. 이럴 경우 지지도와 신뢰도가 항상 1 이 나오게 되므로 연관 규칙의 의미가 없어진다. 그러므로, 가장 선호하는 카테고리부터 정렬하여 상위 5개만이 값을 유지하고 나머지는 0 으로 초기화한다. 표 1의 도서 정보 DB의 카테고리들 중에 A1과 A2 간의 연관 규칙이 적용되는 개념을 아래에 설명한다.

1. 먼저 지지도(Support)를 구한다. 전체 사용자 중 항목 A1과 항목 A2를 동시에 포함하는 사용자가 어느 정도인가를 나타내주며 전체 사용자 관심도에 대한 경향을 파악할 수 있다

$$\text{지지도} = \frac{A1과\,A2를\,포함하는\,사용자수}{\text{전체사용자}} = P(A1 \cap A2)$$

2. 신뢰도(Confidence)를 구한다. 항목 A1를 포함하는 사용자 중에서 항목 A2가 포함될 확률은 어느 정도인가를 나타내주며 연관성의 정도를 파악할 수 있다.

$$\text{신뢰도} = \frac{A1과\,A2를\,포함한\,사용자수}{A1을\,포함한\,고객수} = P(A2|A1) = \frac{P(A1 \cap A2)}{P(A1)}$$

3. 리프트(Lift / Improvement)값을 구한다. 항목 A1를 포함한 경우 그 사용자가 항목 A2를 포함하는 경우와 항목 A2가 임의로 포함되는 경우의 비를 나타내 준다.

$$L = \frac{P(A2|A1)}{P(A2)} = \frac{P(A1 \cap A2)}{P(A1)P(A2)}$$

로 나타낼 수 있으며 리프트 값이 1 이상이면 유용한 정보로 판단할 수 있다. 이와 같은 방식으로 A1부터 J5까지 모든 카테고리에 대하여 쌍을 지어(pairwise) 이 작업을 반복한다. 결과 값이 1 이상인 관계에 대하여 연관 도서 정보 DB에 저장한다.

4.2 사용자 군집 분석

새로운 사용자가 본 사이트에 접속하였을 경우 사용자와 그룹과의 거리를 계산하여 가장 가까운 그룹으로 군집한다. 하지만 여기에서는 그룹의 중심값은 이동하지 않는다. 본 에이전트의 상위 계층은 시스템을 주시하며 새로운 사용자의 증가와 시스템 부하 정도를 측정하여 데이터 마이닝 계획을 세운다. 사용자가 증가할수록 그룹의 중심값이 많이 변할 것이며, 새롭게 군집이 편성될 것이다. 군집 분석 방법은 표 1의 사용자 도서 정보 DB는 각 카테고리별의 사용자 관심도이다. 이것을 비율로 변환한다. 이 비율은 각 사용자의 가중치 합으로 나눈값이다. 예를 들면, 표 1의 진승훈 사용자 경우에 가중치의 합이 25

이기 때문에 A1의 비율 $\frac{1}{25}$ = 0.04 로 계산된다.

<표 2> 카테고리별 관심 비율

	A1	A2	A3	.....	D1	D2	.....	J5
진승훈	0.04	0.04	0.08	.....	0.08		.....	
김혜영	0.063		0.031	.....			.....	
김병익		0.073	0.073	.....		0.195	.....	0.048
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

표 2와 같이 각 사용자들의 관심도를 비율로 표시하면 이것으로 카테고리 수만큼의 차원으로 사용자들의 위치를 표시 할 수 있다.

이것을 바탕으로 사용자들 간의 거리를 구한다. 군집 분석의 방법으로는 K-means 군집 분석을 사용한다. K-means 군집 알고리즘은 다음과 같이 동작한다.

1. K 개의 중심값을 임의로 생성한다.
2. 각 관찰치를 기준 값과 가까운 곳으로 군집한다.
3. 각 그룹의 평균을 구하여 중심값을 다시 구한다.
4. 2와 3을 반복하며 안정화 될 때까지 행한다.

이렇게 나누어진 각 군집들에 속한 사용자들의 도서 정보를 바탕으로 그들이 가장 선호하는 카테고리 순으로 정렬을 하여 저장하고 이를 바탕으로 사용자에게 필요한 정보를 추천한다.

5. 결론 및 향후 과제

본 연구진이 제안한 멀티에이전트 시스템에 도서 추천의 질을 향상시키기 위해 마이닝 기법을 응용하여 보았다. 웹 마이닝의 경우 빠른 결과를 필요로 하는 부분과 그렇지 않은 부분으로 나누어 생각해 보아야 될 것이다. 본 시스템에서는 데이터 마이닝 에이전트의 경우 두 개의 계층으로 분류하였다. 즉, 반응층 과 계획층으로 나누었다. 사용자가 본 사이트에 처음 접속하였을 때 군집 분석을 다시 한다면 상당한 시간이 걸려 적절한 정보를 사용자에게 제공하지 못할 것이다. 빠른 정보제공이 필요한 부분은 반응층에서 제어하고 시간이 많이 걸리는 작업의 경우 사용자 증가도와 서버 시스템 부하 여부를 판단하여 계획을 세우고 실행에 옮긴다.

향후 과제로는 본 시스템에 사용된 데이터 마이닝 기법의 알고리즘은 범용적인 것을 사용하였으나 본 시스템에 더욱 알맞은 알고리즘을 찾아보아야 할 것이며, K-means 군집 분석에서 군집의 수 K의 값을 적절히 발견하는 방법도 연구해야 한다. 본 논문에서는 자세히 언급하지 않았으나 사용자 패턴 감지 부분도 함께 연구되어야 한다.

참고문헌

[1] 강현철외 4인, SAS Enterprise Miner 4.0을 이용한 데이터마이닝 - 방법론 및 활용, 자유아카데미, 2001.

[2] 김영순, 김종완, 이승아, 진승훈, "멀티 에이전트 시스템의 연동 워크플로우 구축", 정보과학회 2001 봄 학술발표회 논문집(B), pp.280-282, 2001.

[3] 이선교, 이도현, "웹마이닝 개념 및 기술동향", 데이터베이스연구지, Vol.16, NO.1, pp.41~46, 2000. 8.

[4] 오병우, 박지웅, 한기준, "비연계 DB테이블상에서의 데이터 추출을 위한 규칙기반의 데이터마이닝 기법", 퍼지 및 지능시스템학회 2000년 추계학술대회, Vol,10 NO.2, pp.220~240, 2000. 11.

[5] Steven H. Kim, "Intelligent Services On The Internet through DataMining Agents", 정보처리학회 논문지, 제7권 제1호, pp.56~62, 2001.

[6] Gediminas Adomavicius, Alexander Tuzhilin, "Using DataMining Methods to Build Customer Profiles", IEEE Computer Magazine, Vol.34, No.2, pp.74~82, 2001.2

[7] 데이터마이닝의 개념 http://www.data-mining.co.kr/

[8] SAS EnterpriseMiner(DataMining Tool) http://www.sas.com/products/miner/index.html

[9] 개인화의 개념http://www.personalization.org/personalization.html

[10] 고수정, 임기욱, 이정현, "협력적 여과 시스템을 위한 효과적인 사용자 군집 알고리즘," 정보처리학회 논문지, 제8-B권 2호, pp144~154, 2001. 4