Collaborative Filtering기반 추천 시스템에 관한 연구

이 재 황**, 김용구*, 장정록*, 엄태광*
삼성전자 디지털미디어연구소

A Study on Recommendation System Using Collaborative Filtering

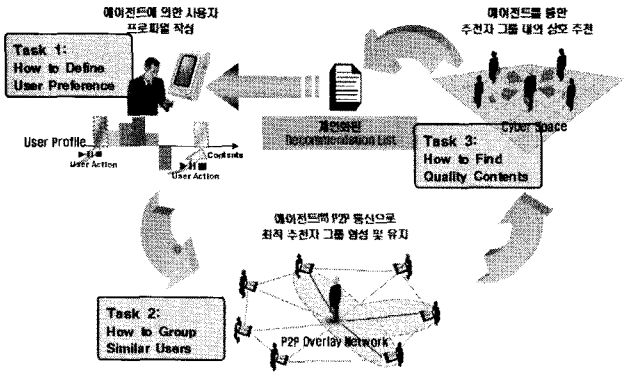
JaeHwang Lee**, Yongku Kim*, Jeongrok Jang*, Taekwang Um*
Digital Media R&D Center, Samsung Electronics

Abstract - 본 논문은 협업 필터링(Collaborative Filtering)기반의 추천시스템에 필요한 알고리즘을 제안한다. 제안한 알고리즘은 사용자의 선호도를 Implicit Feedback을 통해 예측하는 Implicit Rating과 사용자 선호도와 컨텐츠의 정보를 바탕으로 사용자의 프로파일을 형성하는 Tag 기반의 사용자 프로파일과 P2P망 내에서 자신과 유사한 사용자 그룹을 형성하는 알고리즘으로 구성되어 있다. 제안한 알고리즘을 적용하여 Web Text기반의 CF기반의 개인화 추천시스템을 구현하였으며 구현된 프로그램을 실제 사용자에게 배포하여 Feasibility를 검증하였다.

1. 서 론

최근 광대역 네트워크의 보급과 인터넷의 대중화로 웹의 규모는 폭발적으로 증가하여 전세계 도메인 개수가 1억개를 넘어섰다. 이로 인해 웹을 통해 얻을 수 있는 정보의 양은 과거에 비해 비교할 수 없을 정도로 증가했지만 오히려 많은 정보 중에 사용자가 원하는 것을 찾는데 많은 어려움을 겪고 있다. 이러한 정보의 홍수 속에 필요한 정보만을 쉽게 선별할 수 있는 정보 필터링 기법에 대한 요구가 증가하는 것은 검색관련 산업의 가파른 확대 및 발전을 통해서 쉽게 알 수 있다. 그러나 검색엔진과 같은 정보 필터링 기법은 사용자가 해당 검색엔진의 사용법과 적절한 키워드를 인지하고 있어야 하며 해당 검색결과가 많을 경우에 전체를 확인하기 어려운 문제점에 노출되어 있다. 이러한 문제점을 해결하여 정보접근의 근원적인 문제점을 해결할 수 있는 것이 추천시스템이다. 추천시스템을 통해 개인의 정보검색활동을 대행할 수 있고 개인의 선호나 요구를 통해 개인화된 정보 필터링이 가능하다. 추천알고리즘은 협업 필터링(Collaborative Filtering)과 컨텐츠기반 필터링(Contents Based Filtering)으로 구분할 수 있으며 서로 상호보완적인 관계를 가진다[1].

본 논문에서는 사용자가 접근한 컨텐츠에서 사용자의 기호(Preference)를 추출하여 사용자 프로파일을 생성하여 이를 협업 필터링을 통해 최종적으로 사용자에게 맞는 개인화된 추천컨텐츠를 제시하는 시스템을 연구하였다. 제안한 협업 필터링기반 추천 시스템의 전체적인 동작 흐름은 <그림1>과 같이 크게 세 부분으로 구성된다. 첫 번째는 사용자의 행동으로부터 기호를 추출하고 이를 사용자 프로파일로 생성하는 부분이고 두 번째는 생성된 사용자 프로파일과 유사한 사용자 그룹을 생성하고 유지하는 부분이고 세 번째는 최적추천자 그룹에서 상호 추천된 컨텐츠를 사용자의 기호에 맞게 우선순위화 하는 부분이다. 제안한 세부분을 Web에서 보는 Text를 대상으로 사용자 프로파일을 생성하고 P2P를 통해 사용자 그룹을 생성하여 상호 추천하는 P2P기반 Web Text 추천 시스템을 PC Application으로 구현하고 이를 사용자에게 배포하여 실제 사용하게 함으로써 전체 시스템의 Feasibility를 확인하였다.

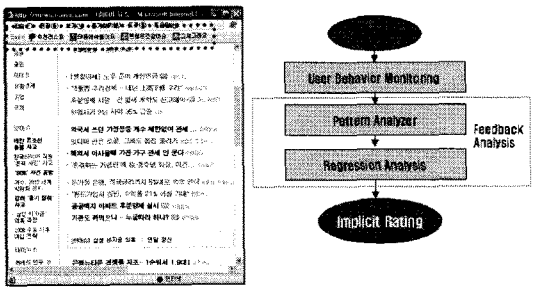


<그림 1> 제안한 협업필터링 기반 추천시스템 전체 동작 흐름

2. 본 론

2.1 사용자의 행동으로부터 기호 추출(Implicit Rating알고리즘 검증)

일반적으로 사용자의 선호도를 파악하는 방법에는 명시적인 방법(Explicit Rating)과 묵시적인 방법(Implicit Rating)이 있다[2]. 명시적인 방법은 사용자의 명시적이 입력이 없을 경우 해당 컨텐츠에 대한 기호파악이 불가능한 한계로 인해 본 논문에서는 Implicit Rating을 이용하여 사용자 선호도를 추출하였다[3]. Implicit Rating방법의 유효성 검증을 위해 Web에 있는 Text기반 컨텐츠에서 사용자 행동을 추출하기 위한 User Behavior Monitoring Tool을 그림<2>와 같이 구현하였다[4]. User Behavior Monitering Tool은 사용자가 Web에 있는 Text를 읽고 평가를 직접 입력(Explicit Rating)하는 Toolbar프로그램과 해당 Text내에서의 사용자 행동 패턴을 수집하는 프로그램으로 구성하였다. 구현한 Tool을 실제 50여명의 사용자에게 배포 후 1달 정도 사용하도록 하여 4000여개의 데이터를 수집하고 분석하여 사용자 행동패턴 중 기호와 가장 연관성이 높은 인자를 도출하여 수집된 Explicit Rating간의 상관관계를 회귀식으로 도출하였다. 도출된 회귀식으로 계산된 결과와 사용자가 직접 평가한 Web Text의 평가값을 RMSE(Root Mean Square Error)식을 이용하여 비교하였다.



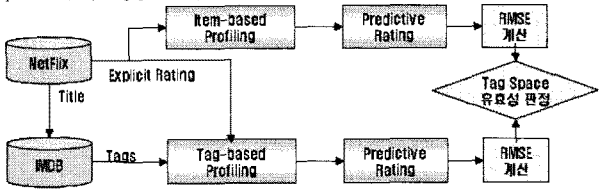
* Implicit Rating 검증용 Tool : 사용자가 봤던 Web Text에 대한 평가(1점~5점)을 직접 입력, Web Text내에서 행해졌던 사용자 행동(마우스 스크롤, Web Text에 머문 시간 등)을 저장

* Implicit Rating 알고리즘: 평가점수와 사용자 행동 데이터를 취합하여 연관관계를 분석하여 모델링

<그림 2> Implicit Rating 검증용 Tool 및 알고리즘 검증 Flow

2.2 Tag기반 사용자 프로파일 검증

본 절에서는 CF기반의 추천시스템에서 효율적으로 사용할 수 있는 사용자 프로파일을 제안하였다. 실제 Test를 위해 Netflix dataset으로 검증하였고 확장된 정보(Metadata)는 IMDB라는 미국영화정보제공사이트에서 추출하였다[5][6]. 일반적으로 Netflix dataset을 이용한 Item-base 추천방식은 컨텐츠에 대한 사용자 평가가 없는 Sparsity문제가 존재한다. 이러한 문제점을 해결하고자 Item space범위를 넓힐 수 있도록 각 Item의 제목이외의 확장된 정보(Metadata)를 이용하는 Tag기반 사용자 프로파일을 제안하였다. Tag기반의 프로파일의 유효성을 검증하기 위해 그림3과 같이 Item-based profile과 Tag-based profile을 이용하여 Predictive Rating값을 추출하여 이를 RMSE값으로 비교하였다. 추출된 dataset을 Training set과 Test set으로 나누어 Predictive Rating의 RMSE를 각각 구하여 Tag-based profile의 유효성을 표1과 같이 검증하였다.



<그림 3> Tag기반 프로파일 유효성 검증 Flow

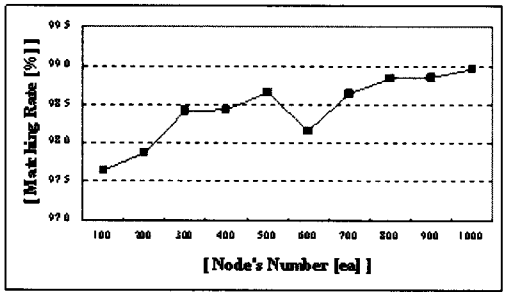
Tag's best result (option 1)	0.9438	0.9685
Item's Best result (option 2)	0.9305	0.9768

Option1: 추천자그룹100명 Option2:추천자그룹 300명

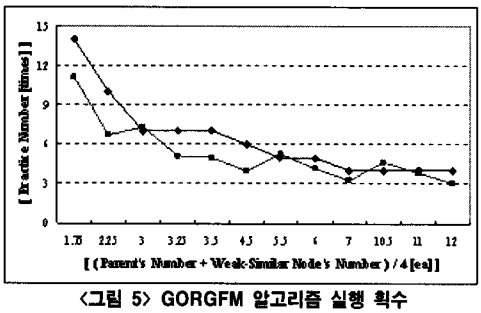
<표 1> Tag/Item 기반 RMSE비교

2.3 P2P 기반 유사 사용자 그룹 형성 알고리즘 검증

대부분의 추천시스템이나 추천서비스는 서버에서 모든 데이터를 가지고 이를 이용한다. 대표적인 상용 추천 서비스인 도서 추천서비스인 아마존도 서버기반의 추천시스템이다. 이러한 서버기반 추천시스템의 단점은 서버가 동작하지 않을 경우 서비스를 받을 수 없고 컨텐츠나 사용자의 수가 급증에 대응할 수 있도록 서버의 유지 및 관리하기 위한 운용비용도 증가하게 된다. 특히 사용자 정보와 같은 개인정보 유출이나 서버 운영자에 의해 특정 컨텐츠만 추천될 수 있는 것과 같은 신뢰성의 문제점에 노출되어 있다. 반면에 P2P환경에서는 서버에 집중화된 컨텐츠가 분산되고 사용자 정보가 각 개인마다 분산되어 있으므로 개인정보 유출로부터 비교적 안전하며 서버 운영자에 의한 정보노출의 신뢰성 문제를 해결할 수 있다. 그러한 이러한 P2P의 장점에도 불구하고 컨텐츠 자체가 각 Peer들에게 분산되어 있으므로 특정 컨텐츠를 찾고자 할 때 전체 Peer에게 모두 접속해야 한다는 문제점이 발생한다. 따라서 CF기반의 필터링 기술을 적용하기 위해 Peer로 구성된 분산망 내에서 효과적으로 컨텐츠를 찾는 방법이 필요하다. 본 절에서는 P2P네트워크 환경에서 특정노드와 유사한 특성을 가진 노드를 찾아 추천자 그룹을 형성하고 유지하는 알고리즘(GORGFM : Global Optimal Recommender Group Formation&Maintenance, 이하 GORGFM)을 제안하였다[7][8]. 제안한 알고리즘의 검증을 위해 특정 노드가 P2P 네트워크 환경에 처음 접속했을 때 제안한 알고리즘에 의해 형성된 추천자 리스트(Recommender List)와 실제 최적 추천자 리스트를 비교하는 Matching Rate를 그림 4과 같이 확인하였다. 아울러 추천자 리스트를 얼마나 빠르게 형성되는지도 상관관계를 분석하여 상관관계식을 도출하여 그림 5와 같이 2000여개의 노드에 대해서 관계식에 의해 계산된 값(파란색)과 실제 시뮬레이션 하여 나온 결과(빨간색)을 비교하였다. 그래프를 통해 관계식과 시뮬레이션을 통해 나온 결과값이 거의 동일함을 확인할 수 있다.



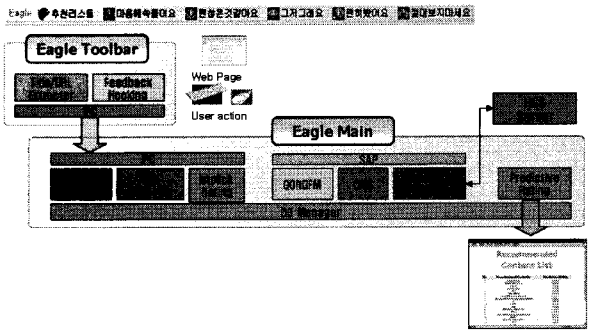
<그림 4> Node(Peer)의 개수에 따른 Matching Rate 결과



<그림 5> GORGFM 알고리즘 실행 획수

2.4 CF기반 개인화 Web Text추천 시스템 구현

본 절에서는 2.1~2.3절에서 검증한 알고리즘을 기반으로 CF기반의 Web Text추천 시스템을 구현한 결과를 기술하였다. 구현된 시스템은 그림 5와 같이 Web Text에서 사용자 행동을 추출하여 사용자 프로파일을 생성하는 IE Plug-In형태로 구현된 Toolbar application과 사용자 프로파일을 가지고 GORGFM을 통해 P2P망 내에서 최적추천자그룹을 형성하고 Web Text 추천리스트를 보여주는 Application으로 구성되어 있다.



<그림 6> CF기반 개인화 Web Text 추천 시스템 전체 구성도

아울러 P2P 네트워크 내에서 최적추천자그룹을 형성하고 Web Text추천리스트를 도출하기 위한 사용자 프로파일 및 Content metadata 및 Peer정보 등록을 위한 P2P프로토콜을 설계 구현하였다. 기존 표준과의 호환성 및 방화벽을 통과하기 위해 HTTP over TCP를 이용하였고 내부 Message는 XML형태로 구현하였다. Web Text 추천 시스템을 50명의 사용자에게 배포 후 사용한 결과 추천된 Web Text에 대한 만족도보다는 타인이 본 Web Text 중 우연하게 본인에게 꼭 필요한 정보가 될 수 있는 Serendipity효과와 같은 의외성에 높은 평가를 주었다. 이는 Web Text라는 컨텐츠 자체의 개인별 선호도 차이가 크기 않고 Web Text로부터 추출한 사용자 프로파일이 선호도를 정확히 반영하기 못하기 때문이다.

3. 결 론

본 논문에서는 무한컨텐츠 시대에 새로운 컨텐츠 접근 방법을 해결하고자 CF기반 추천시스템에 필요한 새로운 알고리즘을 제안하고 검증하였다. Implicit Feedback알고리즘은 Web Text내에서의 사용자의 기호를 나타내는 유효 Feedback 선정하고 유효 Feedback과 사용자의 Rating점수를 통해 회귀식을 추출하여 통계학적으로 사용 가능한 결과를 도출하였다. Tag기반 사용자 프로파일 알고리즘은 비록 Item기반의 프로파일보다는 결과 측면에서는 좋지 않았지만 Data Sparsity문제를 해결할 수 있는 시도로써 Tag의 추출과 후처리(예:Tag Normalization)를 통해 프로파일의 성능을 향상시킬 수 있는 가능성을 확인할 수 있었다. GORGFM알고리즘은 P2P내 특정 Profile을 가진 Peer를 찾을 경우에 Convergence속도나 계산량에서 좋은 성능을 볼 수 있었으며 다양한 활용이 가능할 것으로 기대된다. 마지막으로 추천 Prototype구현 및 배포를 통해 제안한 알고리즘의 Feasibility확인 및 향후 성능개선의 토대를 마련하였다. 향후 현재의 Tag기반 사용자 프로파일 같은 통계적 접근 이외에 지식이나 추론(Ontology) 등을 이용하는 새로운 기술을 추가하여 프로파일의 정확도를 향상시킬 예정이다. 아울러 컨텐츠에 대한 Implicit Rating의 정확도 향상을 위해 컨텐츠에 대한 사용자의 기호를 쉽게 추출할 수 있는 방안을 다각도로 고려할 것이다. 마지막으로 추천시스템 성능평가를 위해서 기존의 통계적 방법을 이용한 추천시스템 성능 측정 Metric이외에 사용자의 추천시스템 만족도를 직접적으로 반영할 수 있는 새로운 Metric에 대한 연구도 병행되어야 할 것이다.

[참 고 문 헌]

[1] Badrul Sarwar, George Karypis, Joseph Konstan, John Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms", Proceedings of the 10th International World Wide Web Conference, pages 285-295, 2001

[2] Mark Claypool, David Brown, Phong Le, and Makoto Waseda, "Inferring User Interest", Internet Computing, IEEE, Vol 5, Issue 6, Nov-Dec 2001, pp. 32-39.

[3] D.M Nichols, "Implicit Rating and Filtering", Proc. 5th DELOS Workshop on Filtering and Collaborative Filtering, Nov. 1997, ERCIM, Sophia Antipolis, France, pp. 31-36 [2] D.M Nichols, "Implicit Rating and Filtering", Proc. 5th DELOS Workshop on Filtering and Collaborative Filtering, Nov. 1997, ERCIM, Sophia Antipolis, France, pp. 31-36

[4] Choochart Haruechaiyasak, Mei-Ling Shyu, and Shu-Ching Chen, " A data mining framework for building a Web-page recommender system", Proceeding of the 2004 IEEE International Conference on Information Reuse and Integration, 8-10 Nov 2004, pp. 357-162

[5] Netflix Prize, http://www.netflixprize.com

[6] IMDB, http://www.imdb.com

[7] Goldberg, D.E., Genetic algorithms in search, optimization, and machine learning, Reading, MA, Addison-Wesley, 1989.

[8] Adeli, H. and Hung, S., Machine learning : neural networks, genetic algorithms, and fuzzy systems, New York, Wiley, 1995.