대량의 소셜 컨텐츠에서 의미 있는 정보의 필터링 연구

안득현
성균관대학교 정보통신공학부 컴퓨터공학과
e-mail : novum21@gmail.com

A Study on Filtering for Meaningful Information in the Massive Social Contents

Deuk-Hyeon Ahn
Dept. of Computer Engineering, School of Information & Communication Engineering,
SungKyunKwan University

요 약

무수히 많은 정보가 쏟아져 나오는 시대에 살고 있는 웹 사용자에게 유용한 정보를 제공하기 위한 여과기법의 연구는 큰 중요성을 갖는다. 이런 기법엔 크게 내용 기반 여과방식과 협업적 여과방식 두 가지로 나눌 수 있다. 이들 각각은 서로 장, 단점을 가지고 있으며 따라서 이를 병합한 기법의 연구는 필수적이다. DB 의 WAL 기법과 진화알고리즘을 이용하여 좀 더 사용자에게 최적화된 추천을 가능하게 할 수 있다. 또한 폭소노미에 기반한 태깅기법 및 패턴인식, 온톨로지(ontology) 기법의 연구를 통해 기존의 한계를 보완하여 향후 더욱 개선된 여과 기법을 기대할 수 있다.

1. 서론

웹 2.0 에 적응하기도 전에 웹 3.0 이라는 단어가 나올 정도로 인터넷산업은 빠르게 변화하고 있으며, 이에 부응하듯 수 많은 정보가 쏟아져 나오고 있는 이 정보화 사회에서 사람들 또한 서로 소통하고 원하는 정보를 얻기 위해 트위터와 같은 다양한 서비스를 이용하고 있다. 그러나 대량으로 생산되는 정보들 사이에 의미 없는 정보 또한 쏟아져 나와 사용자가 원하는 정보를 얻는 데에 적지 않은 방해가 되고 있다. 또한 움직이며 사용하는 스마트폰의 특성과 비싼 통신요금 등으로 인해 1 초가 중요한 현 상황에서, 사용자가 원하는 정보를 정확하고 신속하게 얻기 위한 컨텐츠 여과 및 추천 기법 등의 중요성은 갈수록 부각되고 있다. 이에 현재 존재하는 여과기법들을 특성에 따라 분류해보고 그 장, 단점의 분석을 통해 향후 이를 이용한 연구방향을 논의하고자 한다.

2. 관련 연구

2.1 여과 기법 분류

사용자의 선호도를 알기 위한 방법에는 여러 가지가 있지만 그 중 가장 대표적인 기법은 내용 기반 여과(Content-based Filtering), 협업적 여과(Collaborative Filtering)[1]가 있다. 먼저 내용 기반 여과기법은 아이템 즉, 컨텐츠의 속성과 사용자의 요구를 비교하여 가장 적합한 컨텐츠를 여과하는 방법이다. 이 기법에서의 컨텐츠는 사진, 동영상, 소리와 같은 내용의 특징과 속하는 분류로 표현된다.
협업적 여과기법은 다수의 사람들이 컨텐츠에 가중치를 부여하고 이를 통해 유사한 선호도를 가진 이웃들(neighbors)을 발견하여, 해당 컨텐츠에 대한 이웃들의 평가를 기반으로 추천 컨텐츠를 여과하는 방법이다. 그 종류에는 메모리기반 방식과 모델기반 방식 두 가지가 있다. 메모리 기반 방식은 사용자 선호도 전부를 데이터베이스에 저장하여 이를 통해 예측하며, 모델 기반 방식은 이런 데이터베이스가 만들어지게 된 히스토리를 가진 하나의 모델을 만들어 예측한다.

2.2 내용 기반 여과의 특징 및 장, 단점

컨텐츠엔 여러 종류가 있지만 가장 기본적이고 일반적으로 많이 이용하는 텍스트 검색이 있다. 내용 기반 여과기법에선 우선 텍스트의 검색과 범주화 과정이 필요하다. 색인을 추출하기 위해 불용어제거, 어근추출, 동의어사전, 역문서 빈도수 방법을 이용하여 인덱스화 하며, 범주화의 방법에는 높은 효율을 보이는 네이브 베이지안 분류자[2]를 주로 이용한다. 또한 사용자에게 맞춰진 정보를 위해 유저프로파일[3]을 통해 연관된 정보를 이용하여 콘텐츠의 특징을 추출하는 방법도 있다. 이들 방법엔 형태소 분석에 의해 명사를 추출하는 Aprior 알고리즘이 많이 이용된다. 그 결과물로 연관 단어 집합(bag-of-association)을 만들고 그 단어들의 빈도수를 보고 유저에 니즈에 가장 적합한 컨텐츠를 여과할 수 있다.
텍스트로 된 컨텐츠와 달리 멀티미디어 컨텐츠인 사진이나 동영상, 소리와 같은 컨텐츠는 단순히 이런 방법을 적용시키긴 힘들며 어떤 분류 방법을 이용하느냐에 따라 여과의 정확도가 변한다는 단점이 존재한다. 또 유저들이 이미 익숙한 컨텐츠에 제한될 한계도 존재한다. 하지만 내용기반 여과기법은 뒤에서 나올 초기 사용자 문제로 알려진 협업적 여과의 단점

이 없다는 장점도 있다.

2.3 협업적 여과의 특징 및 장, 단점

협업적 여과기법은 사용자들이 특정 컨텐츠에 대해 어느정도 선호를 하는지 알기 위해 사용자-컨텐츠 행렬[4]을 이용하여 예측한다. 이 방법은 사용자와 선호 성향이 유사한 사용자(Nearest Neighbor)의 의견에는 높은 가중치를 주고 반대의 경우에는 낮은 가중치를 주는 방법으로, 행렬의 각 요소는 컨텐츠에 대한 사용자의 선호 가중치로 채워진다. 그 후 컨텐츠를 원하는 유저와 가장 비슷한 성향의 사용자들을 코사인 유사도와 피어슨 상관관계를 이용하여 결정한다. 결정된 사용자들의 유사도를 반영하여 가장 높은 예측 점수를 받은 컨텐츠를 여과하게 된다. 이 방식이 유저기반(User-based)여과방식이며 아이템기반(Item-based)여과 방식은 유저가 이전에 선호했었던 컨텐츠와 비슷한 컨텐츠를 선호할 것이라는 것에 기반한 여과 방법이다.

협업적 여과에는 희박성(Sparsity) 문제가 존재하는데 사용자-컨텐츠 행렬의 선호 가중치를 바탕으로 추천 컨텐츠를 선별할 때, 실제 행렬에 채워진 데이터의 선호 가중치 점수가 너무 적어 추천의 정확성이 떨어질 우려가 있다는 점이다. 또한 초기 사용자 문제(cold-start problem)도 존재한다. 이는 사용자가 행렬에 처음 추가되어 선호 가중치점수가 아예 없거나 점수가 있더라도 미미한 수의 선호 점수만이 형성되는 문제이다. 그렇지만 컨텐츠에 종류에 관계없이 여과를 하기 쉽다는 장점[5]이 있고 과거에 경험하지 않은 새로운 컨텐츠에 노출되기 좋다는 장점이 존재한다.

3. 비교 및 분석

각각의 기법을 위에서 설명한 내용을 바탕으로 컨첸츠의 종류, 컨텐츠 수의 많고 적음, 정보를 원하는 사용자에게 있어서의 컨텐츠의 질, 컴퓨터 리소스의 측면에서 상대적인 비교를 아래 표와 같이 정리 해보았다.

	내용기반여과	협업적기반여과
컨텐츠의 종류	열세	우세
컨텐츠의 양	열세	열세
컨텐츠의 질	열세	우세
리소스측면	우세	열세

<표 1> 내용기반여과와 협업적기반여과의 상대적 비교

내용기반여과방식은 앞서 말했다시피 그 내용을 알기 힘든 컨텐츠들은 분석이 힘들기 때문에 상대적으로 협업적기반 여과방식에 비해 한계가 많이 있다. 그리고 컨텐츠의 양이 많을수록 협업적기반 여과방식은 가중치를 매겨야 할 아이템이 많아지기 때문에 희박성이 발생할 가능성이 높아 열세로 하였으며, 내용기반여과방식은 분석해야 하는 데에 그만큼 오버헤드가 생기기 때문에 열세로 하였다. 컨텐츠 질에 대해선 새로운 컨텐츠를 추천 받을 확률이 높은 협업적기반 여과방식에 우세를 주었으며, 컴퓨팅 리소스 측면에서 사용자의 선호 데이터가 많아 질수록 많은 자원을 필요로 하기 때문에 열세로 평가하였다.

4. 결론 및 향후연구

위와 같이 가장 대표적인 여과 기법에 대해 각각 특징을 살펴보고 장, 단점들을 비교해 보았다. 이러한 특성으로 인해 이들의 단점은 보완하고 장점은 살리기 위한 하이브리드 형식의 기법연구가 필수적이라는 것을 알 수 있었다. 따라서 다른 CS 분야에서 쓰이는 기법을 융합해보았는데, DB 기법 중 WAL(Write Ahead Log)은 히스토리를 구축할 수 있게 한다. 이 Log 파일을 분석하여 사용자의 패턴을 마이닝하고, 진화알고리즘을 이용하여 추천 패턴을 자동적으로 업데이트 시켜나가는 방법을 사용한다면 좀 더 사용자에게 맞춰진 정확한 추천이 가능하리라 여긴다. 최근 폭소노미[6]에 기반한 태그이용 여과기법의 연구도 진행되고 있는데, 여과성능의 향상은 있지만 이 또한 사용자들의 개인적이거나 감정적인 단어태깅으로 인해 여과성능을 떨어뜨리는 원인이 되고 있다. 따라서 이들을 개선하기 위해 패턴 인식과 관련된 인공지능 분야 및 다양한 사용자 참여 유도방법 고안, 온톨로지(ontology)기술 등을 연구하고, 다양한 분야의 융합을 이용한 여과 기법의 형태로 생성하는 연구를 한다면 향후 더욱 유용한 기법들이 생성 되리라 기대한다.

참고문헌

[1] 고수정, “전자상거래에서 협력적 여과와 내용 기반 여과를 병합한 사용자 선호도 마이닝”, 공학박사학회 2002

[2] Jae Moon Lee, “An Efficient Algorithm for NaiveBayes with Matrix Transposition”, KIPS 2004

[3] 정경용, 조선문, “내용 기반 필터링을 위한 프로파일 학습에 의한 선호도 발견”, 한국콘텐츠학회 2008

[4] Heung-Nam Kim, Ae-Ttie Ji, Inay Ha, Geun-Sik Jo “Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation”, Electronic Commerce Research and Applications 2008

[5] Ana Belén Barragáns-Martínez, Enrique Costa-Montenegro , Juan C. Burguillo , Marta Rey-López , Fernando A. Mikic-Fonte , Ana Peleteiro, “A hybrid content-based and item-based collaborative filtering approach”, Information Sciences 2010

[6] Adam Mathes, “Folksonomies – Cooperative Classification and Communication Through Shared Metadata”, Computer Mediated Communication 2004