지능정보연구 제14권 제2호 2008년 6월(pp.157~177)

효과적인 추천 시스템을 위한 협업적 태그 기반의 여과 기법

연 철
인하대학교 컴퓨터 · 정보공학과
(entireboy@eslab.inha.ac.kr)

지애띠
인하대학교 컴퓨터 · 정보공학과
(aerry13@eslab.inha.ac.kr)

김흥남
인하대학교 컴퓨터 · 정보공학과
(shlee@eslab.inha.ac.kr)

조근식
인하대학교 컴퓨터 · 정보공학과 교수
(gsjo@inha.ac.kr)

최근 웹 2.0의 영향으로 태깅을 지원하는 인터넷 서비스들이 많아졌다. 태깅의 원래 목적은 컨텐츠를 분류하고 재검색을 용이하게 하는 것이지만, 컨텐츠에 태깅되어 있는 태그들을 분석하여 컨텐츠의 특성을 파악할 수 있다. 본 논문에서는 내용 파악이 힘든 컨텐츠들이 증가함에 따라 이러한 컨텐츠들의 효과적인 추천을 위해, 여러 사용자들에 의해 협업적으로 태깅된 정보를 이용한 여과 기법을 제시한다.

제안하는 방법은 사용자가 태깅한 정보들을 바탕으로 사용자의 관심을 파악하는 부분과 파악된 관심에 맞는 컨텐츠를 선별하는 부분으로 나뉘어진다. 사용자의 관심을 파악하는 부분은 사용자가 태깅한 정보들을 협업적 여과를 이용하고, 컨텐츠 선별은 확률적인 방법인 나이브 베이지안 분류자를 이용한다. 이를 통해 협업적 여과 방법의 문제점인 희박성 문제(sparsity problem)와 초기 사용자 문제(cold-start user probleam) 대해 기존의 방법들과 비교하여 그 효과를 보인다.

논문접수일 : 2008년 05월 게재확정일 : 2008년 06월 교신저자 : 지애띠

1. 서론

웹 기술의 발전으로 현재 웹 상에는 수많은 컨텐츠들이 존재한다. 그리고 최근 디지털 기기의 발전과 YouTube[1]나 다음TV팟[2] 등과 같은 서비스의 보편화로 사진이나 동영상, 사운드와 같이 그 내용 파악이 힘든 컨텐츠의 생산이 많아지고 있다. 하지만 많은 컨텐츠들 중에 사용자가 원하는 정보를 찾기란 쉽지 않기 때문에 사용자가 원하는 정보를 손쉽게 찾을 수 있도록 도와주는 추천 시스템의 사용이 증가하고 있다(Sarwar et al., 2000; Miller et al., 2004).

이러한 추천 시스템의 컨텐츠를 선별하는 방법은 컨텐츠의 내용을 파악하여 선별하는 내용 기반 여과(Content-based Filtering)와 컨텐츠의 과거 구매 기록이나 사용 기록을 이용하여 선별하는 협업적 여과(Collaborative Filtering)가 대표적이다(Breese et al., 1998). 이 중 협업적 여과는 컨텐츠의 선별에 컨텐츠의 내용을 사용하지 않기 때문에, 내용 파악이 힘든 컨텐츠를 선별하는데 유용하게 사용될 수 있다(Sarwar et al., 2000; Sarwar et al., 2001; Miller et al., 2004; Resnick et al., 1994). 협업적 여과는 Amazon.com[3]과 같은 상업적 시스템

1) YouTube http://www.youtube.com.
2) 다음TV팟 http://tvpot.daum.net.
3) Amazon.com http://www.amazon.com.

에서도 사용할 정도로 우수한 추천 성능을 보이지만, 희박성 문제(sparsity problem)나 초기 사용자 문제(cold-start user probleam)와 같은 문제점이 있다(O'Donovan and Smyth, 2005; Deshpande and Karypis, 2004).

이러한 문제를 해결하기 위해 본 논문에서는 여러 사용자들에 의해 협업적으로 태깅(Collaborative Tagging)된 정보를 이용한 여과 기법을 제안한다. 웹 2.0의 영향으로 컨텐츠의 분류와 재검색을 용이하게 하기 위한 목적으로 태그(tag)를 많이 사용하기 시작했다. 사용자가 기존에 사용한 태그를 이용하여 사용자의 관심(interest)을 알 수 있고, 컨텐츠에 태깅(tagging)된 태그를 이용하여 컨텐츠의 특성(character)을 알 수 있다. 이러한 태그 정보를 이용하여 사용자의 관심을 알아내고 사용자 관심에 맞는 특성을 가진 컨텐츠를 선별하여 추천하는 방법을 제안한다.

본 논문은 제 2장 관련연구에서 협업적 여과와 태깅에 대한 배경 지식과 관련 연구를 소개하고 기존의 협업적 여과의 문제점을 제시한다. 제 3장 협업적 태깅을 이용한 여과 기법에서는 문제점을 해결하기 위해 본 논문에서 제안하는 보다 효과적인 추천을 위한 협업적 태깅을 이용한 여과 기법에 대해 기술하고, 제 4장 실험 및 평가에서 기존의 방법들과의 실험을 통해 성능 분석을 한다. 그리고 제 5장 결론 및 향후 연구에서 본 논문의 결론과 향후 연구에 대해 언급한다.

2. 관련연구

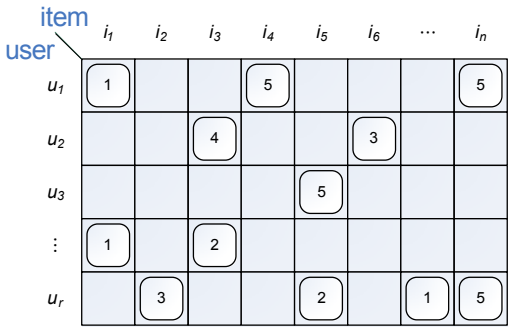
2.1 협업적 여과

컨텐츠를 선별하기 위해 컨텐츠의 내용을 파악하는 내용 기반 여과 방법과는 달리 협업적 여과는 컨텐츠의 내용이 아닌 컨텐츠의 구매 기록이나 사용 기록만을 통해 추천할 컨텐츠를 선별한다(Miller et al., 2004; O'Donovan and Smyth, 2005; Resnick et al., 1994).

어떠한 의사 결정을 하기 위해 사용자는 자신과 관심사가 비슷한 다른 사용자의 의견에 큰 영향을 받는다(Shardanand and Maes, 1995). 예를 들어, 어떠한 제품을 구매하고자 할 때 먼저 사용해 본 주변 사용자의 의견을 듣고 구매 여부를 결정하게 된다. 특히, 나와 구매 성향이 비슷한 사용자의 의견이라면 더욱 지대한 영향을 받을 수 있다. 협업적 여과는 이런 내용으로부터 출발한 컨텐츠 선별 방법이다.

협업적 여과의 한 방법인 사용자 기반의 협업적 여과(User-based Collaborative Filtering)는 나와 선호 성향(preference trend)이 비슷한 사용자의 의견에는 높은 가중치를 주고 그렇지 않은 사용자의 의견에는 낮은 가중치를 주어, 그 의견들을 종합하여 추천해 줄 아이템(컨텐츠)을 선별하는 방법이다.

협업적 여과를 사용하는 시스템은 일반적으로 <그림 1>과 같은 행렬을 사용한다. 행렬의 각 요소는 사용자 u가 아이템 i에 대해 자신의 선호(관심) 정도를 나타내는 점수로 채워진다. 사용자가 어떠한 아이템에 관심이 많으면 높은 선호 점수를 주고,



<그림 1> 사용자-아이템 행렬

그렇지 않은 경우 낮은 점수를 준다. <그림 1>의 사용자 u_1은 아이템 i_1에는 관심이 없어 낮은 점수를 줬지만, i_4는 관심이 많아서 높은 점수를 준 것을 볼 수 있다.

사용자 기반의 협업적 여과는 <그림 1>과 같은 사용자-아이템 행렬에 채워진 선호 점수를 추천에 이용한다. 추천 시스템으로부터 아이템을 추천 받을 추천 대상 사용자(target user)와 선호가 가장 비슷한 사용자 k명을 찾아 추천 대상 사용자의 이웃이라 칭한다. 이 k명으로 이루어진 이웃을 KNN (k nearest neighbor) 이라 하고, 아이템을 선별할 때 이 이웃들의 의견을 종합한다(Sarwar et al. 2000).

추천 대상 사용자와 선호가 비슷한 이웃을 찾기 위해 추천 대상 사용자와 다른 사용자가 얼마나 유사한지 유사도(similarity)를 계산한다. 두 사용자 간의 유사도 계산을 위해 식 (1)과 같이 코사인 유사도(Cosine Similarity)(Breese et al., 1998)와 피어슨 상관관계(Pearson Correlation)(Devore, 2007)가 대표적으로 사용된다.

코사인 유사도를 이용한 유사도 계산식은 두 사용자 u, v를 모든 아이템 I를 차원(dimension)으로 하는 벡터로 취급하여, 식 (1)과 같이 두 벡터의 코사인 내적으로 유사도를 계산한다. 여기서 $R_{u,i}$는 사용자 u가 아이템 i에 대한 선호 점수이다. 두 벡터가 비슷하면 즉, 사용자의 선호가 유사하면 코사인 유사도 값은 1에 가까운 값이 나오고, 그렇지 않으면 0에 가까운 값이 나오게 된다.

$$sim(u,v) = \cos(\vec{u},\vec{v}) = \frac{\sum_{i \in I} R_{u,i} \cdot R_{v,i}}{\sqrt{\sum_{i \in I}(R_{u,i})^2}\sqrt{\sum_{i \in I}(R_{v,i})^2}} \quad (1)$$

$$P_{u,i} = \overline{R_u} + \frac{\sum_{v \subset KNN(u)}(R_{v,i} - \overline{R_v}) \cdot sim(u,v)}{\sum_{v \subset KNN(u)} sim(u,v)} \quad (2)$$

사용자 u와 다른 모든 사용자와의 유사도를 계산하여 가장 유사도가 높은 사용자 k명을 추천 대상 사용자 u의 이웃으로 결정한다. 그리고 식 (2)와 같은 점수 예측 식을 이용하여 이웃 사용자들의 의견을 종합하게 된다.

$P_{u,i}$는 사용자 u가 아이템 i를 얼마나 선호할 것인가를 예측한 점수이고 $KNN(u)$는 사용자 u의 이웃 집합이다. 사용자 u와 이웃 v간의 유사도 $sim(u, v)$를 가중치로 하여 이웃 v의 아이템 i에 대한 선호 점수를 합산한다. 이웃과 유사도가 높다면 $sim(u, v)$값이 커서 그 사용자의 의견 $R_{v,i}$가 더 많이 반영이 되고, 유사도가 낮다면 적게 반영되게 된다. 이렇게 구해진 예측 점수가 가장 높은 아이템을 추천 대상 사용자에게 추천해 주게 된다.

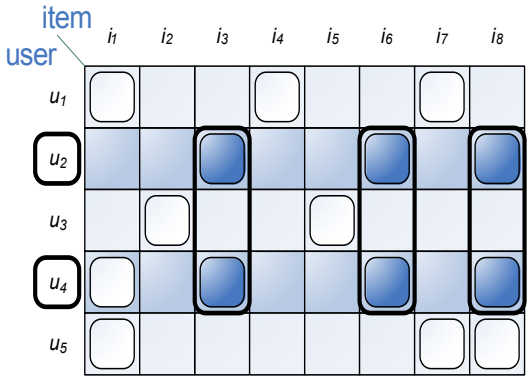
아이템 기반의 협업적 여과(Item-based Collaborative Filtering)는 사용자 기반의 협업적 여과와 비슷한 추천 과정을 가진 협업적 여과의 또 다른 종류이다(Deshpande and Karypis, 2004). 사용자 기반의 협업적 여과는 나와 비슷한 선호를 가진 사용자의 의견이 의사 결정에 높은 영향을 미친다는 것에 기반했다면, 아이템 기반의 협업적 여과는 내가 선호했었던 아이템과 비슷한 아이템을 선호할 것이라는 것에 기반한 아이템 선별 방법이다(Sarwar et al., 2001).

2.2 협업적 여과의 문제점

협업적 여과는 Amazon.com과 같은 상업 시스템에서도 사용되고 있을 정도로 추천 성능이 효과적이지만 몇 가지 문제점이 있다. 협업적 여과는 사용자-아이템 행렬에 채워진 사용자의 선호 점수를 바탕으로 추천해 줄 아이템을 선별하게 되는데, 실제 데이터는 행렬에 채워진 선호 점수가 너무 적어서 희박성 문제(Sparsity Problem)가 있다(O'Donovan and Smyth, 2005; Deshpande and Karypis,

2004). 실험을 위해 수집된 데이터 집합 역시 선호 점수가 0.1% 정도밖에 채워져 있지 않았다.

<그림 2>는 사용자 u_2와 사용자 u_4의 유사도를 계산할 때의 모습을 나타낸 행렬이다. 두 사용자 간의 유사도를 계산할 때 두 사용자가 함께 점수 표시(co-rating)를 한 아이템(i_3, i_6, i_8)의 점수를 이용한다. 하지만 점수 데이터가 희박하면, 동일한 아이템에 함께 점수 표시한 것을 찾기가 힘들어지고 추천의 정확성이 떨어질 수 있다(Sarwar et al., 2001; Shardanand and Maes, 1995).



<그림 2> 사용자 기반의 협업적 여과

협업적 여과의 또 다른 문제점으로 희박성 문제뿐만 아니라 초기 사용자(cold-start user) 문제도 있다(O'Donovan and Smyth, 2005; Deshpande and Karypis, 2004). 초기 사용자는 행렬에 새로 추가되어 선호 점수가 하나도 없거나 선호 점수가 있더라도 적은 수의 선호 점수만이 표시된 사용자를 뜻한다. 이런 초기 사용자들은 선호를 파악하기 위해 사용되는 선호 점수가 없거나 적기 때문에 사용자의 선호 파악이 쉽지 않으며, 그 파악된 선호의 정확성도 떨어진다. 아이템에 대해서도 초기 사용자 문제와 마찬가지로 선호 점수가 없거나 적은 초기 아이템 문제가 있다.

본 논문에서는 이런 협업적 여과의 문제점을 보완하기 위해 협업적 태깅을 이용한 아이템 선별 방법을 제시한다.

2.3 협업적 태깅(Collaborative Tagging)

태깅은 컨텐츠에 태그를 붙이는 행위를 뜻하며 기존의 컨텐츠에 키워드(keyword)를 붙이는 행위(annotation)와 같은 개념이다(Wikipedia, Last accessed 2008). 태그는 컨텐츠를 분류하거나 재검색하기 용이하게 하기 위해 주로 컨텐츠의 내용이나 컨텐츠 자체를 대표할 수 있는 단어들을 이용한다. 단어 사용에 제한 없이 사용자가 원하는 단어나 단어들의 조합으로 자유롭게 어떠한 형태로도 사용이 가능하며 여러 개의 태그를 태깅하는 것도 가능하다.

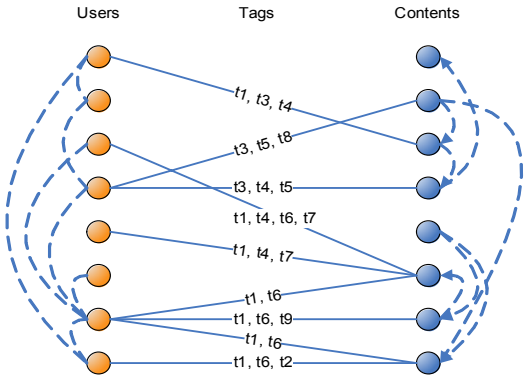
태깅은 기존에도 존재하던 개념이지만 최종 사용자(end-user)에게 직접적인 이득을 주는 서비스가 없어 적은 분야에서만 사용되었다. 하지만 최근에는 웹 2.0의 영향과 폭발적으로 증가하는 컨텐츠로 인해 Gmail[4]이나 Flickr[5], del.icio.us[6], 올블로그(allblog)[7], Technorati[8] 등과 같은 태깅을 지원하는 서비스가 점차 늘어나고 있다.

컨텐츠에 태깅된 각 태그들 간에는 상하관계가 아닌 모두 동일한 수평적인 관계를 가지기 때문에 기존의 카테고리(category)나 디렉토리(directory)와 같은 수직(계층 : hierarchical)적 분류와는 차이를 보인다(Golder and Huberman, 2005). 태그는 수직적인 분류의 보완재적인 역할을 하여 보다 효과적으로 컨텐츠를 분류하고 재검색할 수 있도록

4) Gmail http://www.gmail.com.
5) Flickr http://www.flickr.com.
6) del.icio.us http://del.icio.us.
7) allblog http://www.allblog.net.
8) Technorati http://www.technorati.com.

도와준다.

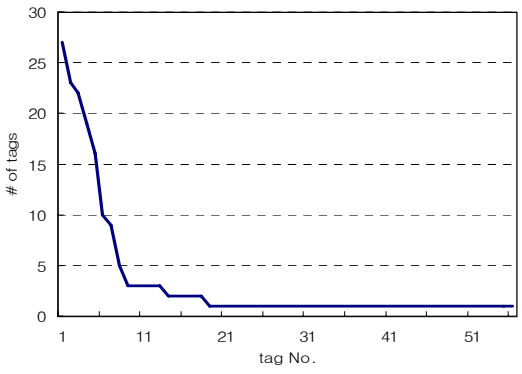
태깅은 그 특성에 따라 몇 가지 형태로 구분할 수 있다(Marlow et al. 2006). 우선, 컨텐츠에 태깅을 할 수 있는 권한에 따라 self-tagging과 permission-based, free-for-all로 구분할 수 있다. self-tagging 방식은 올블로그(allblog)나 YouTube, Technorati 등과 같은 서비스에서 사용하는 태깅 방식으로 컨텐츠 생성자만이 태깅이 가능하다. permission-based 방식은 컨텐츠 생성자에게서 태깅 권한을 부여받은 사용자만이 컨텐츠에 태깅할 수 있는 방식으로 Flickr와 같은 서비스에서 사용하고 있다. free-for-all 방식은 del.icio.us나 마가린[9], Last.fm[10]처럼 모든 사용자가 컨텐츠에 태그를 붙일 수 있는 권한을 가진 방식이다. 협업적 태깅은 <그림 3>과 같이 한 컨텐츠에 여러 사용자가 태깅을 할 수 있는 free-for-all 방식이나 permission-based 방식과 같은 태깅 방식에서 볼 수 있다.



<그림 3> 협업적 태깅

컨텐츠에 태깅된 태그들의 모음 방식에 따라 bag 형태와 set 형태로 구분할 수 있다(Marlow et al., 2006). set 형태의 태깅 방법은 중복을 허용하지 않아 컨텐츠에 태깅되어 있는 태그의 종류는 알 수 있지만 각 태그의 빈도수는 알 수 없다. 하지만 bag 형태의 태깅 방법은 중복을 허용하기 때문에 컨텐츠에 태깅되어 있는 태그의 종류와 빈도수를 모두 알 수 있다. del.icio.us나 Yahoo! My Web, 마가린과 같은 서비스는 bag 형태의 태깅 방법을, Flickr나 YouTube, Technorati 등의 서비스에서는 set 형태의 태깅 방법을 사용한다.

컨텐츠에 태깅된 태그의 빈도수를 알 수 있는 bag 형태의 태깅 방법을 사용하여 통계적인 접근이 가능하다. <그림 4>는 본 논문의 실험을 위해 수집된 실험 데이터 집합 중 한 컨텐츠에 태깅되어 있는 태그의 종류별 빈도수를 나타낸다. 일반적으로 한 컨텐츠에 태깅되어 있는 태그의 빈도수 분포를 보면 <그림 4>와 같이 특정 몇 종류의 태그가 높은 비중으로 태깅되어 있는 롱테일(The Long Tail) 형태의 곡선(Power law curve 또는 Power curve)인 것을 볼 수 있다.



<그림 4> 한 컨텐츠에 태깅된 태그별 빈도수

이러한 롱테일 현상은 브로드 폭소노미(Broad folksonoly)일수록 두드러진다(Wal, 2005). 폭소노미(folksonomy)는 folk와 taxonomy의 합성어로 집

9) 마가린 http://mar.gar.in.
10) Last.fm http://www.last.fm.

단에 의한 자연적인 분류를 뜻하는 말이다. 한 컨텐츠에 많은 사용자가 태깅을 하면 <그림 4>와 같은 롱테일 형태의 태그 분포를 띠는 것을 볼 수 있고, 이 중 높은 빈도수를 가지는 태그들로 이 컨텐츠를 나타낼 수 있다. 이로부터 한 컨텐츠에 태깅된 태그들 중 빈도수가 높은 태그들이 이 컨텐츠의 모든 특성을 나타낼 수는 없어도 이 컨텐츠를 대표할 수 있다고 볼 수 있다. 카테고리나 디렉토리와 같은 분류는 전문가에 의해 만들어지고 각 분류 간에는 계층을 가진 복잡한 형태를 띠지만, 폭소노미에 의한 분류는 다수의 사용자들에 의해 자연적으로 만들어지고 각각의 분류는 계층을 가지지 않는다는 차이가 있다.

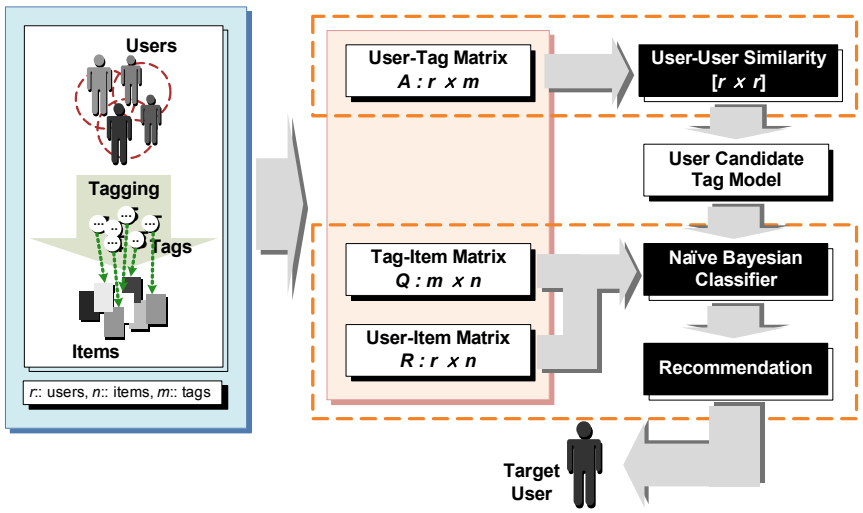
브로드 폭소노미는 이러한 폭소노미의 일종으로 del.icio.us와 같이 한 컨텐츠에 태깅할 수 있는 사용자가 많은 free-for-all 방식의 태깅을 사용하는 시스템에서 잘 나타난다. 반면에 네로우 폭소노미(Narrow folksonomy)는 Flickr와 같이 한 컨텐츠에 태깅할 수 있는 사용자가 적은 permission-based 방식의 태깅을 사용하는 시스템에서 볼 수 있다(Wal, 2005).

태그는 사용자가 컨텐츠를 분류하고 재검색을 용이하게 하기 위해 각자 나름의 의미를 가지고 붙이는 단어이기 때문에 그 사용자의 선호를 나타낸다고 할 수 있다. 또한, 각 컨텐츠에 태깅된 태그들 중 빈도수가 높은 태그들로 그 컨텐츠를 대표할 수 있듯이, 사용자가 많이 사용한 태그도 마찬가지로 그 사용자의 선호를 대표한다고 할 수 있다.

3. 협업적 태깅을 이용한 여과 기법

본 논문에서는 기존의 협업적 여과를 이용한 추천 시스템에서 발생할 수 있는 문제점인 희박성 문제와 초기 사용자/아이템에 대한 문제를 보완하기 위해 협업적 태깅을 이용한 여과 기법을 제안한다.

사용자가 아이템에 태깅을 함으로써 사용자가 많이 태깅한 태그로 사용자의 선호를 알 수 있고, 아이템에 많이 태깅된 태그로 아이템의 특성을 파악할 수 있다. 이를 바탕으로 본 논문에서 제안하는 추천 방법은 <그림 5>와 같이 크게 2부분으로



<그림 5> 협업적 태깅을 이용한 추천 시스템

나뉜다. 사용자가 태깅에 사용했던 태그 정보를 바탕으로 사용자의 선호를 파악하는 부분과 아이템에 태깅된 태그 정보를 바탕으로 첫 번째 부분에서 파악된 사용자의 선호에 맞는 아이템을 추천해 주는 부분이다.

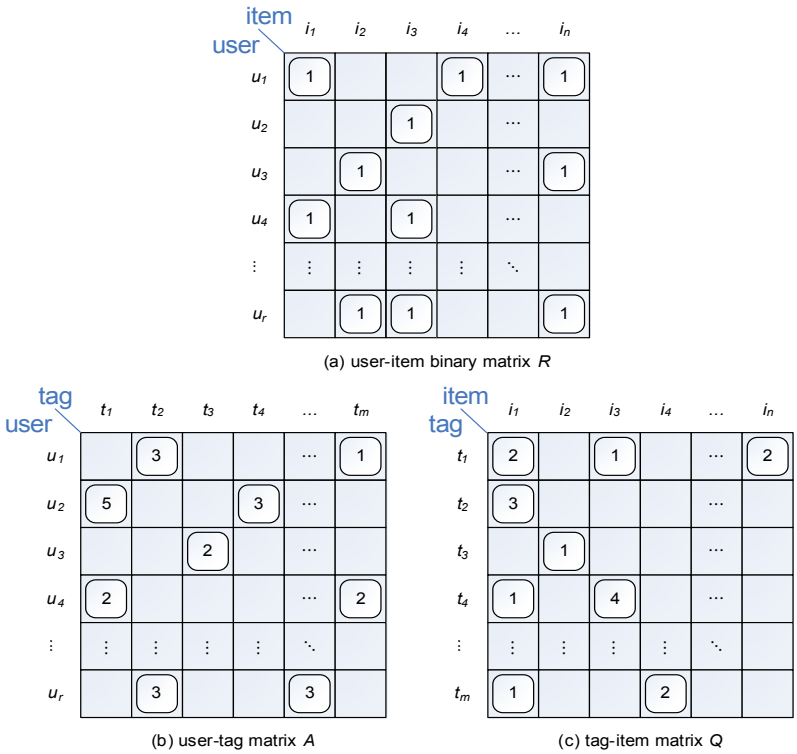
사용자의 선호는 후보 태그 집합(Candidate Tag Set, *CTS*)으로 표현된다. 후보 태그 집합 *CTS*는 사용자의 선호를 나타내는 태그들로 이루어진 태그의 집합으로, 사용자가 자주 사용하여 직접적으로 관심을 보인 태그들, 그리고 그 태그들과 유사한 태그들로 구성된다. 이 태그 집합을 사용자의 선호 성향으로 보고 태그들과 가장 특성이 맞는 아이템을 추천해 주게 된다.

제안하는 추천 시스템을 설명하기에 앞서 본 논문에서 사용되는 행렬들을 정의한다. 기존의 협업적 여과는 사용자가 아이템을 얼마나 선호하는가에 대한 점수 값, 혹은 사용자가 관심을 나타냈거나 그렇지 않았거나 하는 이진 값을 가진 사용자-아이템 행렬 R을 사용한다.

본 논문에서는 태그를 이용한 사용자의 선호를 파악하기 위해 사용자-태그 행렬 A와 아이템에 태깅되어 있는 태그로 아이템의 특성을 파악하기 위해 태그-아이템 행렬 Q를 추가로 이용한다.

- **사용자-아이템 이진 행렬(user-item binary matrix) R**

  r명의 사용자 집합 $U = \{u_1, u_2, \cdots, u_r\}$의 각 사용자가 n개의 아이템 집합 $I = \{i_1, i_2, \cdots, i_n\}$



<그림 6> 협업적 태깅을 이용한 여과를 위한 행렬

의 각 아이템에 대한 선호 여부를 표시한 행렬이다. <그림 6>(a)와 같은 $r \times n$의 사용자-아이템 행렬에 표현할 수 있다. 선호한다고 표현했다면 1의 값으로 나타낸다. 사용자가 모든 아이템에 대해 선호 여부를 표현할 수 없기 때문에, 0 값이 사용자가 선호하지 않는 것을 뜻하는지 아직 선호 여부를 표시하지 않은 것을 뜻하는지는 알 수 없다. $R_{u,i} \in \{0, 1\}$

- **사용자-태그 행렬(user-tag matrix) A**
  사용자가 태깅할 때 사용한 m개의 태그 집합 $T = \{t_1, t_2, \cdots, t_m\}$의 태그에 대한 빈도수를 $r \times m$의 사용자-태그 행렬로 <그림 6> (b)와 같이 나타낸다. 행렬의 각 요소 $A_{u,t}$는 각 사용자 u가 태깅할 때 사용한 태그 t의 빈도수를 나타낸다.

- **태그-아이템 행렬(tag-item matrix) Q**
  아이템에 태깅된 태그의 빈도수를 $m \times n$의 태그-아이템 행렬로 <그림 6>(c)와 같이 나타낸다. $Q_{t,i}$는 사용자들이 아이템 i에 태깅한 태그 t의 빈도수를 나타낸다.

<그림 7>은 위와 같이 정의된 행렬들에 포함된 사용자들의 선호 정보를 바탕으로 추천 대상 사용자에게 아이템을 추천해 주는 방법을 도식화한 것이다.

3.1 후보 태그 집합 구성

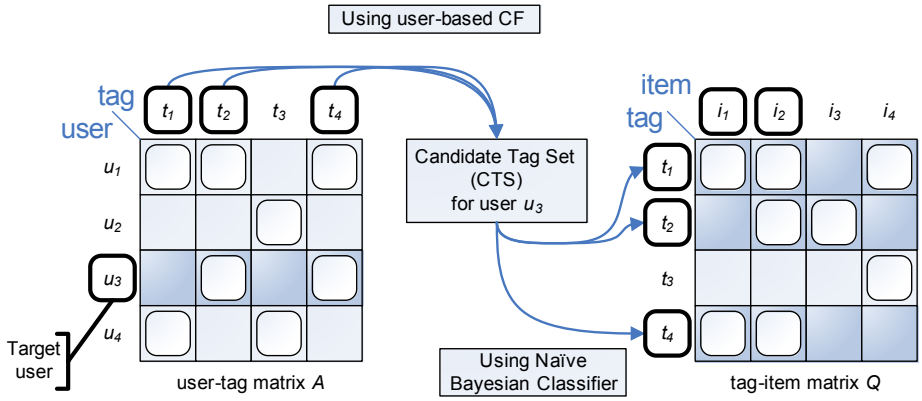
사용자 u의 선호를 파악하기 위해 사용자 u의 선호 성향을 나타내는 $CTS(u)$를 생성한다.

- **후보 태그 집합(Candidate Tag Set, CTS)**
  사용자 u의 후보 태그 집합 $CTS(u)$는 사용자 u의 선호 성향을 대표할 수 있는 태그들의 집합으로 정의한다. 사용자가 과거에 태깅했던 태그들 및 그 태그들과 유사한 태그들로 이루어진다.

$$CTS_w(u) = \{t_x \mid x = 1, 2, \cdots, w, t_x \in T\}$$

w는 후보 태그 집합을 구성하는 태그의 개수이고, T는 전체 태그 집합이다.

사용자의 선호 성향을 파악하고 그 성향에 맞는



<그림 7> 협업적 태깅을 이용한 여과를 위한 행렬

아이템을 추천해 주는 것이 기존의 협업적 여과 방법 보다 더 많은 아이템을 추천 대상으로 선정할 수 있어 희박성 문제나 초기 사용자/아이템 문제를 완화시킬 수 있다.

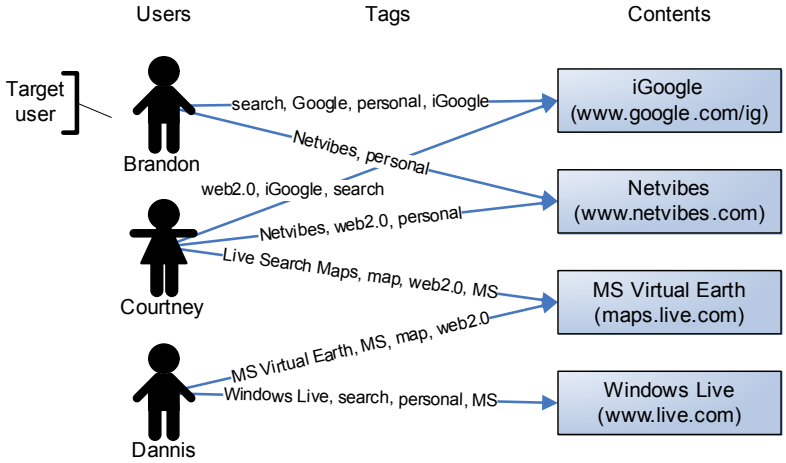
예를 들어, 웹 서핑 중인 사용자 Brandon과 Courtney, Dannis가 있다고 하자. 각자가 찾은 웹 컨텐츠들은 태깅을 이용하여 북마킹(bookmarking)한다. 어떠한 컨텐츠를 북마킹한다는 것은 사용자가 관심을 가질 때 하는 것이므로, 북마킹한다는 것은 사용자가 명시적으로 관심이 있다고 표시하지는 않았지만 암묵적으로 관심이 있다는 것을 의미한다(Nichols, 1998).

각 사용자들이 <그림 8>과 같이 북마킹 하였다. Brandon은 "iGoogle(www.google.com/ig)"과 "Netvibes(www.netvibes.com)"를, Courtney는 "iGoogle"과 "Netvibes", "MS Virtual Earth(maps.live.com)"를, Dannis는 "MS Virtual Earth"와 "Windows Live(www.live.com)"를 북마킹하였다.

이 때 Brandon이 추천을 받기 위해 추천 시스템을 이용한다면, 기존의 사용자 기반의 협업적 여과는 Brandon의 이웃을 찾고 그 이웃이 북마킹한 컨텐츠 중 예측 점수가 높은 컨텐츠들을 추천하게 된다. Courtney는 Brandon과 동일한 컨텐츠를 북마킹하였기 때문에 Brandon의 이웃이 될 수 있고, Brandon은 Courtney가 북마킹한 "MS Virtual Earth" 컨텐츠를 추천 받을 가능성이 있다. 하지만 Dannis는 Brandon과 동일한 컨텐츠를 북마킹한 것이 없기 때문에 Brandon의 이웃이 될 수 없고, Dannis가 알고 있는 "Windows Live" 컨텐츠는 Brandon이 선호하는 컨텐츠일지라도 Brandon에게 추천될 수가 없다. 또한, "MS Virtual Earth" 컨텐츠는 Brandon과 함께 북마킹한 컨텐츠를 가지고 있어 이웃이 될 수 있는 Courtney의 의견이 추천 결과에 반영될 수는 있지만, 함께 북마킹한 컨텐츠를 가지고 있지 않아 이웃이 될 수 없는 Dannis의 의견은 추천 결과에 반영되지 않는다.

하지만 북마킹 시에 사용한 태그 정보를 이용하여 추천 대상 사용자의 선호를 파악하고 그에 맞는 컨텐츠를 추천해 준다면, "Windows Live" 컨텐츠와 같이 Brandon의 이웃이 될 수 없는 사용자가 북마킹한 컨텐츠도 추천받을 수 있다.

추천 대상 사용자의 선호를 파악하여 *CTS*를 구



<그림 8> 태깅을 이용한 북마킹 예

성하기 위해 추천 대상 사용자가 북마킹 시에 사용한 태그 정보를 이용한다. Brandon은 북마킹을 하면서 여러 태그들을 사용하였지만, “personal”, “portal”이라는 태그를 많이 사용하였기 때문에 이 태그들에 관심이 있다고 할 수 있다. 그리고 Brandon이 직접적으로 사용하지는 않았지만, Brandon이 많이 사용했던 태그들과 유사한 태그들도 사용자 기반의 협업적 여과를 통하여 함께 *CTS*를 구성하게 된다. 이렇게 구성된 *CTS*를 추천 대상 사용자의 선호로 보고 이 선호에 맞는 컨텐츠를 추천한다. 컨텐츠에 태깅되어 있는 태그들과 추천 대상 사용자의 *CTS*로 구성된 태그들을 비교하여 추천 대상 사용자의 관심에 맞는 컨텐츠를 추천한다. Brandon이 태깅할 때 자주 사용한 태그들의 집합인 {“personal”, “portal”}으로 *CTS*(Brandon)이 구성되었다고 할 때, 이 태그들이 태깅되어 있는 컨텐츠들을 추천해 주어 “Windows Live” 컨텐츠도 추천받을 수 있다.

기존의 협업적 여과에서는 추천 대상 사용자와 동일한 컨텐츠에 함께 북마킹하지 않은 사용자의 의견은 추천 결과에 반영이 되지 않는다. 그러므로 Brandon과 동일한 컨텐츠에 북마킹한 정보가 없는 Dannis의 의견은 반영되지 않아 “Windows Live” 컨텐츠는 추천이 될 수 없다. 하지만, *CTS*를 이용하여 추천 대상 사용자의 관심을 표현하고 그에 맞는 컨텐츠를 추천해 줌으로써 Brandon에게 “Windows Live” 컨텐츠도 추천이 가능해져서, 협업적 여과의 희박성 문제로 인해 추천 대상 사용자에게 추천이 힘들었던 점을 보완할 수 있다.

또한, 초기 사용자는 선호를 표시한 수가 적기 때문에 동일한 컨텐츠에 함께 선호를 표시한 사용자를 찾아 이웃을 형성하는데 어려움이 있다. 하지만 *CTS*를 구성할 때 컨텐츠에 대한 선호 정보를 이용하는 것이 아닌 태그에 대한 선호 정보를 이용하기 때문에 이웃을 형성하는데 도움을 줄 수 있다. 일반적으로 사용자는 하나의 컨텐츠에 태깅을 할 때 여러 개의 태그를 이용한다. 실험을 위해 수집된 데이터 집합(dataset)의 한 번 태깅 당 사용된 평균 태그 수는 2.98개였다. 사용자가 컨텐츠를 태깅할 때 사용한 태그의 수가 태깅한 컨텐츠 수의 3배가 된다는 것이다. 이것은 협업적 여과에서 이용하는 행렬에 채워지는 사용자의 선호 정보가 많아져 희박성이 줄어들게 되고 희박성에 따른 문제점도 완화될 수 있다는 것을 의미한다.

3.1.1 태그를 이용한 사용자 이웃 집단 구성

추천 대상 사용자 u의 후보 태그 집합 $CTS(u)$는 추천 대상 사용자가 북마킹을 하면서 사용한 태그 정보를 이용하여 사용자 기반의 협업적 여과를 통해 구성된다(Breese et al. 1998). 추천 대상 사용자 u와 비슷한 태그를 사용한 이웃 집단 $KNN(u)$를 구하기 위해 식 (3)을 이용하여 추천 대상 사용자와 각 사용자들 간의 유사도를 계산한다.

$$sim(u,v) = \cos(\vec{u},\vec{v}) = \frac{\sum_{t\in T} A_{u,t} \cdot A_{v,t}}{\sqrt{\sum_{t\in T}(A_{u,t})^2}\sqrt{\sum_{t\in T}(A_{v,t})^2}} \quad (3)$$

$sim(u, v)$는 사용자 u와 v의 유사도를 나타내고, 전체 태그 집합 T의 모든 태그 t에 대하여 각 사용자가 태깅한 빈도수 $A_{u,t}$, $A_{v,t}$를 이용한다. $A_{u,t}$는 앞서 정의한 행렬 A의 요소 값을 뜻한다. 두 사용자 간의 태그 선호 유사도는 0에서 1 사이의 실수 값으로 계산되고, 유사도 값이 높을수록 두 사용자의 선호도는 비슷한 것을 나타낸다. 추천 대상 사용자와 모든 사용자와의 유사도를 계산하여 유사도가 가장 높은 k명의 사용자를 선별하여 $KNN(u)$로 결정한다. 이웃 집단의 크기 k가 너무 작으면 올

바른 예측이 어려우며, 크기가 커질수록 예측 값이 정확해지지만 계산량이 늘어나 시간이 오래걸린다. 따라서 적당한 이웃의 크기를 결정해야 한다(Sarwar et al., 2001; Herlocker et al., 1999).

3.1.2 후보 태그 집합 구성

이웃 집단이 정해졌으면 이웃 집단으로부터 의견을 받아 추천 대상 사용자가 해당 태그를 얼마나 선호할 것인가 하는 선호도를 예측한다. 선호도를 예측하는 식은 식 (4)와 같다(Resnick et al., 1994).

$S_{u,t}$는 사용자 u가 태그 t를 얼마나 선호할 것인가를 예측한 값이다. 사용자 u의 이웃 $KNN(u)$에 포함된 모든 사용자 o로부터 사용자 o가 태그 t를 선호한 정도 $A_{o,t}$를 추천 대상 사용자 u와 사용자 o 와의 유사도 $sim(u, o)$를 가중치로 하여 계산한다. 유사도 $sim(u, o)$는 두 사용자 u와 o간의 선호 성향이 비슷하면 높은 값을 가지는 가중치로 이용할 수 있다.

$$S_{u,t} = \sum_{o \in KNN(u)} (A_{o,t}) \cdot sim(u,o) \qquad (4)$$

태그 선호 예측 값 $S_{u,t}$가 가장 높은 w개의 태그를 추천 대상 사용자 u의 후보 태그 집합 $CTS_w(u)$로 정의하고, 이 태그들이 추천 대상 사용자의 선호 성향을 뜻하게 된다.

<알고리즘 1>은 태그 선호도를 예측하는 과정을 나타낸 것이다. 사용자 간의 유사도를 계산하여 $r \times r$ 유사도 행렬 D를 구축하고, 사용자의 각 태그에 대한 선호도를 예측한 값으로 $r \times m$ 선호도 예측 행렬 S를 구축한다. 사용자 간의 유사도를 계산해 행렬 D와 같이 미리 구축해 이용하면, 다른 사용자의 태그 선호를 예측할 때 유사도를 다시 사용할 수 있다.

<알고리즘 1> 태그 선호도 예측 알고리즘

```
ComputeUserTagPreferenceMatrix(U, k, A, D, S)

input
    U : total user list
    k : size of KNN
    A : user-tag matrix
    D : user-user similarity matrix : r × r
    S : (empty) user-tag preference matrix : r × m

01  set all elements in matrix S with 0
02  for each u ∈ U
03      // get KNN of each user
04      for i ← 1 to r // r is row count of matrix U
05          add D_{u,i} to itemset KNN
06      for each o ∈ KNN
07          if o ≠ among the k largest values in KNN
08              remove o from KNN
09      // compute user-tag preference matrix S
10      for each t ∈ T
11          for each o ∈ KNN
12              S_{u,t} ← S_{u,t} + (A_{o,t} × D_{u,o})
```

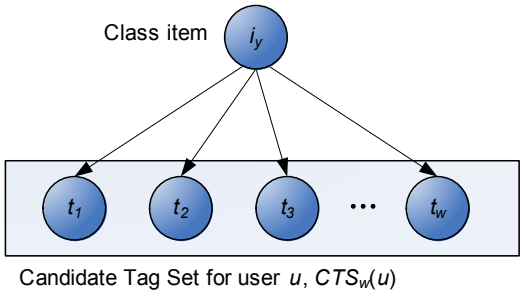
3.2 나이브 베이지안 분류자를 이용한 아이템 추천

구성된 후보 태그 집합 CTS를 나이브 베이지안 분류자(Naïve Bayesian Classifier)의 확률적인 방법으로 아이템에 대한 선호를 예측(Sahami et al. 1998)하고 예측한 값이 가장 높은 상위 N개의 아이템을 상위-N 추천(Top-N Recommendation)을 통해 추천(Deshpande and Karypis, 2004)한다.

3.2.1 아이템 선호도 예측

나이브 베이지안 분류자의 클래스 집합으로는 아이템 집합 $I = \{i_1, i_2, \cdots, i_n\}$를 이용하고, 특징(feature)으로는 추천 대상 사용자의 후보 태그 집합인 $CTS(u)$의 태그들을 이용한다. 추천 대상 사용자의 관심을 나타낸 후보 태그 집합의 태그들이

해당 아이템 클래스에 포함될 확률은 추천 대상 사용자의 해당 아이템에 대한 선호 예측치로 이용한다. 이렇게 구성된 베이지안 네트워크는 <그림 9>와 같이 나타낼 수 있다.



<그림 9> 아이템 선호도 예측을 위한 베이지안 네트워크

<그림 9>와 같이 구성된 베이지안 네트워크는 식 (5)와 같이 나타낼 수 있다. P_{u,i_y}는 전체 아이템 I의 y번째 아이템 i_y에 대한 사용자 u의 선호 예측 값을 나타낸다.

$$P_{u,i_y} = P(I=i_y)\prod_{j=1}^{w} P(t_j|I=i_y) \quad (5)$$

사전 확률 $P(I = i_y)$와 아이템 클래스 i_y 일 때 후보 태그 집합의 j번째 태그에 대한 확률 $P(t_j \mid I = i_y)$은 다음과 같이 계산할 수 있다.

$R_{u,y}$는 u번째 사용자의 y번째 아이템에 대한 선호 값을 의미하고 사용자-아이템 행렬의 값을 사용한다. $Q_{j,y}$는 후보 태그 집합의 j번째 태그가 y번째 아이템에 태깅된 빈도수를 의미하며 태그-아이템 행렬의 값을 사용한다. 그리고 식 (7)은 y 번째 아이템에 해당 태그가 태깅되지 않은 경우 분모의 $\sum_{t=1}^{m} Q_{i,y}$ 값이 0이 되는 것을 방지하기 위해 Laplace Correction을 사용하였다(Han and Kamber, 2006).

$$P(I=i_y) = \frac{\sum_{u=1}^{r} R_{u,y}}{\sum_{g=1}^{r}\sum_{u=1}^{r} R_{u,g}} \quad (6)$$

$$P(t_j|I=i_y) = \frac{1+Q_{j,y}}{m+\sum_{t=1}^{m} Q_{t,y}} \quad (7)$$

3.2.2 상위 N개의 아이템 추천

나이브 베이지안 분류자를 이용한 확률적인 방법으로 추천 대상 사용자의 모든 아이템에 대한 선호 예측치를 계산한 후, 선호 예측치가 높은 아이템을 추천해 주게 된다(Deshpande and Karypis, 2004). 이 때 상위-N 추천을 이용하게 되며, 상위-N 추천은 추천 대상 사용자가 선호를 표시하지 않은 아이템 중 선호 예측치가 가장 높은 아이템 순으로 정렬하여 상위 N개의 아이템을 추천해 주는 것이다.

- **상위-N 추천(Top-N Recommendation)**
  모든 아이템 집합 I에 대해서 I_u는 사용자 u가 선호를 표시한 아이템이라 하고, L_u는 사용자 u가 아직 선호를 표시하지 않은 아이템이라 하자.

  $$I_u \cup L_u = I,\ I_u \cap L_u = \varnothing$$

  사용자 u에 대한 상위-N 추천 $TopN_u$은 선호 예측 점수가 높은 아이템 순으로 다음과 같은 조건을 만족하는 아이템의 집합을 추천하는 것이다.

  $$|TopN_u| \le N,\ TopN_u \cap I_u = \varnothing,$$
  $$TopN_u \subseteq L_u$$

<알고리즘 2> 후보 태그 집합을 이용한 추천 알고리즘

```
Recommend(u, w, N, L_u, S)

input
    u : target user
    w : size of CTS
    N : size of Top-N
    L_u : items not rated by user u
    S : user-tag preference matrix

output
    TopN_u : recommended itemset to user u

01  // get CTS of user u from user-tag matrix S
02  for i ← 1 to m // m is column count of matrix S;
    same as one of matrix A
03      add S_u,i to itemset CTS_w(u)
04  for each x ∈ CTS_w(u)
05      if x ≠ among the w largest values in CTS_w(u)
06          remove x from CTS_w(u)
07
08  for each i_y ∈ L_u
09      // calculated by equation(5)
10      add NaiveBayesClassifier(u, CTS_w(u), i_y, Q) to
        itemset TopN_u
11
12  // recommend Top-N items to user u
13  for each z ∈ TopN_u
14      if P_u,z = 0 ∨ P_u,z ≠ among the N largest values
        in TopN_u then
15          remove z from TopN_u
16
17  return TopN_u
```

<알고리즘 2>는 후보 태그 집합으로부터 아이템을 추천하는 알고리즘을 나타낸다. 계산된 사용자의 태그에 대한 선호 예측 값을 가지는 행렬 S에서 사용자 u의 높은 예측 값을 가지는 w개의 태그를 선택하여 $CTS_w(u)$로 구성한다. $CTS_w(u)$의 각 태그들을 특징으로 나이브 베이지안 분류자를 이용하여 사용자가 아직 선호를 표시하지 않은 아이템 L_u에 대한 확률을 계산하고 그 확률이 높은 N개의 아이템을 추천한다.

본 연구에서 제안하는 추천 시스템의 전체 알고리즘은 <알고리즘 3>과 같다. 전체 사용자 집합 U의 모든 사용자 간의 유사도를 계산해 사용자 간 유사도 행렬 D를 구축한다. 이 사용자 간의 유사도를 이용하여 각 태그에 대한 사용자의 선호도를 예측하여 행렬 S를 구축한다. 이 구축된 태그 선호 예측 값이 있는 행렬 S의 태그 중 예측 값이 높은 w개의 태그를 각 사용자 u의 $CTS(u)$로 결정하고 나이브 베이지안 분류자를 이용하여 모든 아이템에 대한 적합 확률을 구한다. 구해진 확률이 높은 N개의 아이템을 상위-N 추천 방법으로 추천한다.

<알고리즘 3> 전체 추천 시스템 알고리즘

```
RecommenderSystemUsingCollaborativTagging
01  // generate user-user similarity matrix D
02  for each u ∈ U
03      for each v ∈ U
04          if v ≠ u
05              // calculated by equation (3)
06              D_u,v ← sim(u, v)
07
08  // generate user-tag preference matrix S
09  ComputeUserTagPreferenceMatrix(U, k, A, D, S)
10
11  // recommend items to each user
12  for each u ∈ U
13      Recommend(u, w, N, L_u, S)
```

4. 실험 및 평가

본 장에서는 제안하는 협업적 태깅을 이용한 여과 기법, 기존의 사용자 기반의 협업적 여과 기법(Sarwar et al. 2000), 아이템 기반의 협업적 여과 기법(Deshpande and Karypis, 2004)을 실험하여 각 추천 시스템의 여과 기법들의 성능을 비교하고 제안하는 기법의 성능에 대해 논한다.

4.1 실험 데이터 집합 및 평가 방법

실험을 위한 데이터 집합(dataset)은 소셜 북마킹(social bookmarking) 서비스인 del.icio.us를 크롤링(crawling)하여 수집하였다. del.icio.us는 웹 페이지의 북마킹 서비스를 제공하는 사이트로, 북마킹 할 때 1개 이상의 태그를 태깅하도록 되어 있으며 태깅의 중복이 가능한 bag 형태의 태깅을 지원한다(Marlow et al. 2006). 수집한 데이터 집합은 <표 1>과 같이 1,544명의 사용자로부터 얻은 17,390개의 웹 사이트와 10,077개의 태그로 구성되며, 27,066개의 북마킹 정보와 44,681개의 태깅 정보를 가지고 있다.

<표 1> 실험에 사용된 데이터 집합

users	items	tags	book markings	taggings
1,544	17,390	10,077	27,066	44,681

◦ 1,544×17,390 사용자-아이템 이진 행렬 R
◦ 1,544×10,077 사용자-태그 행렬 A
◦ 10,077×17,390 태그-아이템 행렬 Q

행렬 내에 선호도 예측에 이용할 수 있는 선호 정보가 얼마나 있는가 판단할 수 있는 희박성 수준(Sparsity Level)은 식 (8)과 같이 계산할 수 있다(Sarwar et al., 2001). 실험에 사용된 사용자-아이템 행렬의 희박성 수준은 0.9989로 매우 희박하다.

$$\text{SparsityLavel} = 1 - \frac{\text{nonzeroelemontsin user} - \text{item matrix}}{\text{totalelementsin user} - \text{item matrix}} \quad (8)$$

추천의 성능 평가를 위해 총 북마킹 데이터를 80%의 트레이닝 데이터(21,653 북마킹)와 20%의 테스팅 데이터(5,413 북마킹)로 나누어 실험하였다. 성능 평가 방법은 트레이닝 데이터로 학습하여 추천한 결과가 테스팅 데이터와 얼마나 일치하는지의 재현율(recall)을 비교하였다(Sarwar et al., 2000; Deshpande and Karypis, 2004). 재현율의 hit율은 다음과 같이 정의될 수 있다.

$$hit-ratio(u) = \frac{|Test_u \cap TopN_u|}{|Test_u|} \quad (9)$$

$Test_u$는 테스팅 데이터에 있는 사용자 u의 아이템 집합이고, $TopN_u$는 추천 시스템에 의해 추천된 사용자 u의 아이템 집합이다. 사용자 별로 추천된 아이템 집합 중 테스팅 데이터와 얼마나 일치하는가의 비율을 평균으로 계산하여 전체 재현율을 식 (9)와 같이 측정한다.

$$recall = \frac{\sum_{u=1}^{r} hit-ratio(u)}{r} \times 100 \quad (10)$$

4.2 실험 데이터 집합 및 평가 방법

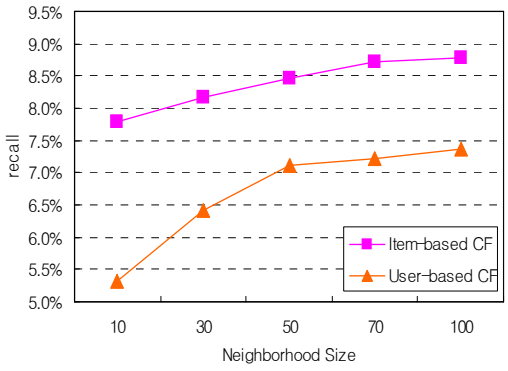
4.2.1 기존 여과 방식 실험

협업적 여과는 이웃 집단 KNN 의 크기에 따라 추천 성능이 크게 변한다(Herlocker et al. 1999). 보다 정확한 실험을 하기 위해 이웃 집단의 크기 k를 10, 30, 50, 70, 100으로 변경하면서 추천 성능을 측정하였다. 상위-N 추천의 추천 아이템 개수 N 은 10으로 고정하였다.

<그림 10>은 이웃 집단 크기 k의 변화에 따른 재현율의 변화를 나타낸 것이다. 대체적으로 이웃 집단의 크기가 커질수록 추천 성능은 좋아졌고, 이웃 집단의 크기가 50 부근에서 서서히 성능 증가율이 감소하는 것을 볼 수 있었다.

사용자 기반의 협업적 여과 보다 아이템 기반의

협업적 여과가 더 좋은 성능을 보였으며, 이는 데이터 집합의 희박성 수준이 높기 때문으로 분석된다(Miler et al., 2004). 아이템 기반의 협업적 여과는 사용자 기반의 협업적 여과보다 작은 모델 사이즈를 이용해도 보다 좋은 추천 성능을 보인다. 하지만, 전체 사용자 수 r보다 전체 아이템 수 n이 크기 때문에 $r \times r$ 사용자 간의 유사도 행렬 보다 $n \times n$ 아이템 간의 유사도 행렬이 상당히 커져서 아이템 기반의 협업적 여과의 유사도 계산이 상당히 오래 걸렸다.



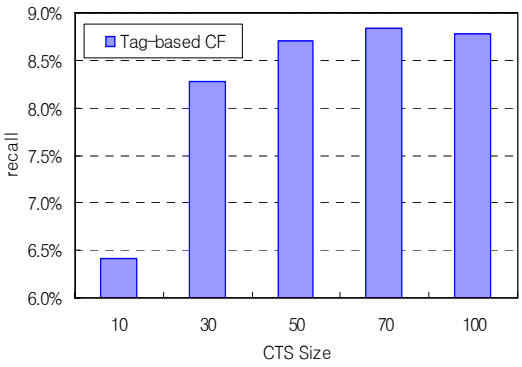
<그림 10> 이웃 집단 크기에 다른 재현율

4.2.2 후보 태그 집합 크기에 따른 실험

기존의 협업적 여과 방법이 이웃 집단의 크기에 따라 추천 성능의 차이가 있듯이, 본 논문에서 제안하는 여과 방법도 사용자의 관심을 몇 개의 태그로 나타내는가에 따라 추천 성능이 차이를 보인다. 따라서 후보 태그 집합의 크기 w의 변화에 따른 추천 성능을 측정하였다. 기존의 협업적 여과 방법의 이웃 집단 크기에 따른 실험과 동일하게 학습 후에 테스팅 데이터의 재현율을 측정하였다.

후보 태그 집합을 구할 때 사용된 사용자 기반의 협업적 여과에 사용된 이웃 집단 크기는 50으로 하였으며, 사용자에게 추천한 아이템의 수 N은 10으로 하여 실험하였다.

후보 태그 집합의 크기에 따른 재현율 실험도 이웃 집단의 크기에 따른 실험과 마찬가지로 그 크기가 증가할수록 추천 성능이 좋아지는 것을 볼 수 있었다. 후보 태그 집합의 크기가 70일 때 가장 좋은 추천 성능을 볼 수 있었지만 그 크기가 더 커지면 추천 성능이 낮아지는 것을 보았다. 이는 후보 태그 집합의 크기가 커짐에 따라 사용자의 선호에 맞지 않는 태그들이 함께 포함된 것으로 분석된다. 즉, 너무 많은 후보 태그는 사용자의 정확한 선호 성향을 파악하는데 좋지 않은 영향을 미칠 수 있고 불필요한 계산량의 증가로 성능이 저하될 수 있다. 좋은 추천 성능을 위해 적당한 크기의 후보 태그 집합을 선택하는 것이 중요하다.



<그림 11> 후보 태그 집합 크기에 다른 재현율

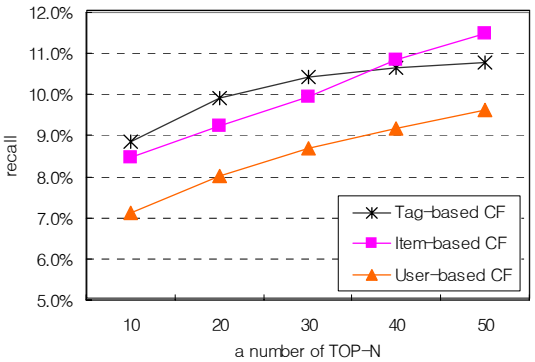
4.2.3 성능 비교 평가

본 논문에서 제안하는 여과 방법과 기존의 협업적 여과의 성능을 비교하기 위해 상위-N 추천의 추천하는 아이템 개수 N의 변화에 따른 성능을 비교하였다. 제안하는 방법의 후보 태그 집합의 크기 w는 이전 실험에서 가장 좋은 성능을 보인 70으로

하고, 비교하는 기존의 협업적 여과 방법의 이웃 집단의 크기 k는 성능 향상이 감소하기 시작하는 50으로 하여 비교 실험을 하였다.

일반적으로 추천하는 아에템의 개수 N이 증가함에 따라 추천 성능이 향상되었다. 하지만 실험을 위해 수집한 데이터의 희박성 수준(0.9989)이 너무 높고, 전체 사용자 수(1,544)에 비해 천체 아이템의 수(17,390)가 너무 많아 모든 방법이 낮은 추천 성능을 보였다.

본 논문에서 제안하는 여과 방법이 기존의 협업적 여과 방법들 보다 대체적으로 좋은 추천 성능을 보였다. 추천하는 아이템의 개수 N이 커짐에 따라 추천 성능이 떨어지는 것을 볼 수 있었는데, 이것은 후보 태그 집합을 결정하는 실험의 상위-N 추천의 추천 아이템 개수 N을 10으로 정하여 얻어진 후보 태그 집합의 개수(w = 70)를 비교 실험에 이용했기 때문으로 분석된다. 추천 아이템 개수 N이 50과 같은 부분에서 좋은 추천 성능을 보이기 위해서는 그에 맞는 후보 태그 집합을 결정해야 할 것으로 보인다. 하지만, 추천 받은 아이템이 작을 때 추천 성능이 좋다는 것은 추천 대상 사용자가 선호하는 아이템이 추천 받은 아이템 목록의 상위에 포함될 확률이 높다는 것을 뜻한다.

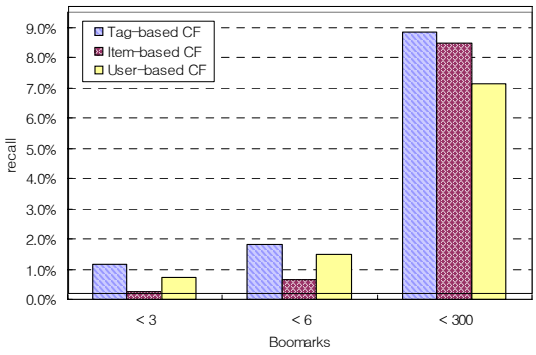


<그림 12> 추천 아이템 수에 따른 재현율 비교

4.2.4 초기 사용자에 대한 비교 평가

선호 정보가 적어 협업적 여과를 이용해서 사용자의 성향을 정확하게 파악하기 힘든 초기 사용자에 대해 성능을 비교하였다. 초기 사용자의 기준을 북마킹이 5개 이하인 사용자로 정의하여 실험을 하였다(Massa and Bhattacharjee, 2004).

제안하는 여과 방법과 기존의 여과 방법을 통해서 초기 사용자라 판단되는 사용자들만의 재현율을 비교하였다. 초기 사용자는 각각 2개 이하, 5개 이하로 북마킹한 사용자들에 대해 실험을 하였고, 300개 이하로 북마킹을 한 사용자(트레이닝 데이터 전체 사용자의 99.6%)와 재현율을 비교하였다. 초기 사용자에 대한 추천 성능은 대부분의 사용자(북마킹 300개 이하인 사용자)에 대한 추천 성능에 비해 상당히 떨어지는 것을 볼 수 있다. 이는 협업적 여과의 문제점 중의 하나이며, 초기 사용자의 선호 정보가 적어 선호 파악이 힘들기 때문이다.



<그림 13> 초기 사용자에 대한 비교 평가

각 여과 방법간의 재현율을 비교해 보면, 2개 이하로 북마킹한 사용자들의 재현율이 제안하는 여과 방법은 1.161%이고, 아이템 기반의 여과 방법은 0.254%, 사용자 기반의 여과 방법은 0.724%로 나타나 본 논문에서 제안하는 여과 방법이 기존의

여과 방법 보다 좋은 결과를 보이는 것을 볼 수 있었다. 이는 북마킹한 아이템 수 보다 태깅한 태그의 수가 더 많아 사용자의 선호를 파악하기 쉽기 때문으로 분석된다.

또한, 사용자 기반의 협업적 여과가 아이템 기반의 협업적 여과에 비해 전체 사용자에 대한 추천 성능은 떨어지지만 초기 사용자에게는 좋은 추천 성능을 보이는 것을 볼 수 있었다. 이는 실험에 이용된 데이터 집합의 전체 사용자의 수에 대한 전체 아이템의 수의 비율이 너무 커서 사용자-아이템 행렬의 아이템 축이 커진 것 때문으로 분석된다. 아이템 축이 커지면 아이템 기반의 협업적 여과를 이용하여 추천을 할 때 아이템 간의 유사도 계산 시에 두 아이템에 함께 선호를 표시한 사용자를 찾기가 어려워 추천 성능이 떨어지기 때문이다(Sarwar et al., 2001; Shardanand and Maes, 1995).

5. 결론 및 향후 연구

기존의 텍스트 위주였던 웹 컨텐츠들이 사진, 동영상, 사운드 등으로 다변화하고 있는 가운데 컨텐츠의 분류와 재검색을 용이하게 하기 위해 태깅을 제공하는 서비스들이 많아졌다. 이에, 본 논문에서는 내용 파악이 쉽지 않은 컨텐츠의 추천을 위해 보다 효과적인 추천을 위해 협업적 태깅을 이용한 여과 기법을 제안하였고 기존의 협업적 여과 기법들과의 비교를 통해 협업적 태깅의 효과를 살펴 보았다.

상위-N 추천 방법을 이용하여 제안하는 여과 기법과 기존의 협업적 여과 기법의 추천 성능을 비교하였으며, 추천 받은 아이템의 개수 N이 작을 때는 보다 좋은 추천 성능을 볼 수 있었다. 추천 받은 아이템이 작을 때 추천 성능이 좋다는 것은 추천 대상 사용자가 선호하는 아이템이 추천 받은 아이템 목록의 상위에 포함될 확률이 높다는 것을 뜻한다. 사용자들이 검색 엔진을 이용할 때 원하는 결과를 얻지 못 하면 일반적으로 결과 페이지의 3 페이지를 넘기지 않고 검색어를 바꿔 재검색하는 사용자 행동 습관이 있다(iProspect, 2006). 이러한 습관에 기반하여 보면, 추천 목록의 상위에 사용자가 선호하는 아이템을 추천할 확률이 높다는 것은 추천 시스템에서 유용하게 이용될 수 있는 부분이다.

하지만, 사용자가 북마킹을 하면서 사용했던 개인적이거나 감정적인 “bad”나 “me”, “to read”, “my work”과 같은 태그는 사용자의 선호 성향을 파악하는데 있어서 추천 성능을 떨어뜨리는 요인이 되었다. 또한, 사용자가 태깅을 하면서 사용하는 태그들이 사용자마다 다른 의미로 사용할 수 있기 때문에 동음이의어나 이음동의어와 같이 단어의 형태 분석으로 그 구분이 불가능한 단어를 사용하여 발생하는 문제점은 의미 기반의 태깅(semantic tagging)에 대한 연구 등으로 보완되어야 할 문제이다.

참고문헌

Breese, J. S., D. Heckerman, and C. Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering”, *In Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence*, (1998), 43～52.

Cano, P., M. Koppenberger, and N. Wack, “Content-based Music Audio Recommendation”, *In Proceedings of the 13th Annual ACM International Conference on Multimedia*, (2005), 211～212.

Deshpande, M. and G. Karypis, “Item-based Top-N Recommendation Algorithms”, *ACM*

Transactions on Information Systems (TOIS), Vol.22, No.1(2004), 143～177.

Devore, J. L., "*Probability and Statistics for Engineering and the Sciences*", 7th Ed., Duxbury Press, 2007.

Golder, S. A. and B. A. Huberman, "The Structure of Collaborative Tagging Systems", Last accessed 3rd January, 2008 from http://arxiv.org/abs/cs/0508082, 2005.

Golder, S. A. and B. A. Huberman, "Usage Patterns of Collaborative Tagging Systems", *Journal of Information Science*, Vol.32, No.2 (2006), 198～208.

Han, J. and M. Kamber, "*Data Mining Concepts and Techniques*", 2nd Ed., Morgan Kaufmann Publishers, 2006.

Herlocker, J. L., J. A. Konstan, A. Borchers, and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering", *In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (1999), 230～237.

iProspect, "iProspect Search Engine User Behavior Study", White Paper, Last accessed 3rd January, 2008 from http://www.iprospect.com/premiumPDFs/hitePaper_2006_SearchEngineUserBehavior.pdf, 2006.

Marlow, C., M. Naaman, D. Boyd, and M. Davis, "HT06, tagging paper, taxonomy, Flickr, academic article, to read", *In Proceedings of the 17th ACM Conference on Hypertext and Hypermedia*, (2006), 31～40.

Massa, P. and B. Bhattacharjee, "Using Trust in Recommender Systems : an Experimental Analysis", *In Proceedings of 2nd International Conference on Trust Management*, (2004), 221～235.

Miller, B. N., J. A. Konstan, and J. Riedl, "PocketLens : Toward a Personal Recommender System", *ACM Transactions on Information Systems (TOIS)*, Vol.22, No.3(2004), 437～476.

Nichols, D. M., "Implicit Rating and Filtering", *In Proceedings of 5th DELOS Workshop on Filtering and Collaborative Filtering*, (1998), 31～36.

O'Donovan, J. and B. Smyth, "Trust in Recommender Systems", *In Proceedings of the 10th International Conference on Intelligent User Interface*, (2005), 167～174.

Resnick, P., N. Iacovou, M. Su-chak, P. Bergstrom and J. Riedl, "GroupLens : An Open Architecture for Collaborative Filtering of Netnews", *In Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, (1994), 175～186.

Sahami, M., S. Dumais, D. Heckerman and E. Horvitz, "A Bayesian Approach to Filtering Junk E-Mail", *In AAAI-98 Workshop on Learning for Text Ca-tegorization*, (1998), 55～62.

Sarwar, B., G. Karypis, J. A. Konstan, and J. Riedl, "Analysis of Recommendation Algorithms for E-Commerce", *In Proceedings of the 2nd ACM Conference on Electronic Commerce*, (2000), 158～167.

Sarwar, B., G. Karypis, J. A. Konstan and J. Riedl, "Item-based Collaborative Filtering Recommendation Algorithms", *In Proceedings of the 10th International World Wide Web Conference*, (2001), 285～295.

Shardanand, U. and P. Maes, "Social Information Filtering : Algorithms for Au-tomating 'Word of Mouth'", *In Proceeding of the SIGCHI Conference on Human Factors in Computing Systems*, (1995), 210～217.

Voss, J., “Collaborative Thesaurus Tagging the Wikipedia Way”, Last accessed 3rd January, 2008 from http://arxiv.org/abs/cs/0604036, 2006.

Wal, T. V., “Explaining and Showing Broad and Narrow Folksonomies”, Last accessed 3rd January, 2008 from http://www. personalinfocloud.com/2005/02/explaining_and_.html, 2005.

Wikipedia. Tag (metadata), Last accessed 3rd January, 2008 from http://en.wikipedia.org/wiki/Tag_%28 metadata%29, 2008.

Abstract

Collaborative Tag-based Filtering for Recommender Systems

Cheol Yeon* · Ae-Ttie Ji* · Heung-Nam Kim** · Geun-Sik Jo***

Even in a single day, an enormous amount of content including digital videos, posts, photographs, and wikis are generated on the web. It's getting more difficult to recommend to a user what he/she prefers among these contents because of the difficulty of automatically grasping of content's meanings. CF (Collaborative Filtering) is one of useful methods to recommend proper content to a user under these situations because the filtering process is only based on historical information about whether or not a target user has preferred an item before. Collaborative Tagging is the process that allows many users to annotate content with descriptive tags. Recommendation using tags can partially improve, such as the limitations of CF, the sparsity and cold-start problem.

In this research, a CF method with user-created tags is proposed. Collaborative tagging is employed to grasp and filter users' preferences for items. Empirical demonstrations using real dataset from del.icio.us show that our algorithm obtains improved performance, compared with existing works.

Key Words : Recommender System, Collaborative Tagging, Collaborative Filtering, Recommendation, del.icio.us

* Department of Computer & Information Engineering, Inha University

저 자 소 개

연 철

2005년에 건국대학교 정보시스템 학사 학위를 취득하였고, 현재 인하대학교 컴퓨터 정보공학과 석사 과정을 졸업 하였다. 주요 관심 분야는 Personalization, Web 2.0, Recommender System, Java 등이다.

지애띠

인하대학교 컴퓨터공학부 학사(2005), 동 대학원의 컴퓨터정보공학과 석사(2007)를 졸업하였고, 현재 인하대학교 컴퓨터·정보공학과 박사과정에 재학 중이다. 주요 관심분야는 E-Commerce, Semantic Web 등이다.

김흥남

인하대학교 컴퓨터공학부 학사(2002), 동 대학원의 컴퓨터정보공학과 석사(2004)를 졸업하였고, 현재 인하대학교 컴퓨터·정보공학과 박사과정에 재학 중이다. 주요 관심분야는 Recommender System, Web Intelligence, Data Mining, Semantic Web 등이다.

조근식

인하대학교 전자계산학과 학사(1982), 미국 뉴욕 CUNY(City University of New York)대에서 전자계산 석사(1985), CUNY에서 전자계산 박사를 취득하였다(1991). 한국지능정보시스템학회 편집장(1998), 인하대학교 창업지원센터 소장(2000), 동대학 전자계산소 소장(2005) 등을 역임하였고, 현재 인하대학교 컴퓨터·정보공학과 교수로 재직 및 BK21 정보기술 사업 단장을 역임 중이다. 주요 연구분야는 CSP, Intelligent E-Commerce System, 인공 지능 등이며 AI Magazine, Expert System with Application, Journal of Organizational Computing and Electronic Commerce 등의 학술지에 논문을 게재하였다.