SNS상에서 하이브리드 협업적 여과 기법을 이용한 전문가 추천 기법 설계

오영만*, 신영성**, 오병석**, 김형일**, 장재우**
*(주)유엠텍 기술개발실 부장
**전북대학교 전자정보공학부 컴퓨터공학
e-mail:yungman-oh@hanmail.net

An Expert Recommendation Technique Design using Hybrid Collaborative Filtering in SNS

Yung-Man Oh*, Young-Sung Shin**, Byeong-Seok Oh**, Hyeong-il Kim**, Jae-Woo Chang**
*Dept of Technology, u-MTec Co., Ltd.
**Dept of Computer System Engineering, Chonbuk National University

요 약

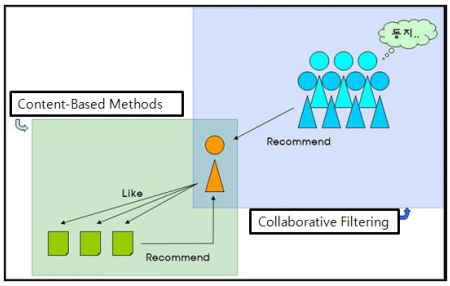
최근 다양한 직업을 가진 SNS 사용자가 증가함에 따라, SNS 사용자들은 전문가 간 협업 및 기술적 의사소통을 위한 전문가 추천 기능을 요구하고 있다. 하지만 기존 협업적 여과 기법은 전문가 추천 서비스를 효율적으로 제공하지 못한다. Content-boosted 협업적 여과 기법은 다양한 예측 알고리즘을 제시하여, 효과적인 추천을 수행할 수 있도록 지원한다. 그러나 명확한 계산 조건이 제시되지 못하는 경우 아이템 및 사용자 유사도 계산을 수행할 수 없는 단점이 존재한다. 따라서 본 논문에서는 Content-boosted 협업적 여과 기법의 단점을 해결하는 하이브리드 협업적 여과기법을 이용한 새로운 전문가 추천기법을 제안한다. 또한, 이를 이용하여 SNS에서의 전문가 추천 시스템을 설계한다.

1. 서론

SNS(Social Network Service) 사용자의 직업, 전문분야가 다양해지고 그 규모가 커짐에 따라, SNS의 사용자들은 팀별 협업 혹은 전문적 정보 공유 등 전문성을 지닌 커뮤니케이션을 위한 전문가 추천 기능을 요구하고 있다. SNS의 협업적 여과 추천 시스템(Collaborative Filtering Recommendation System)의 추천서비스는 주로 아이템 정보(영화, 뉴스정보, 웹 페이지 검색, 책, 음악 등)에 초점이 맞추어져 있으며, 어느 분야의 전문가를 추천해 주는 서비스는 미흡한 실정이다. 따라서 새로운 서비스 대상인 전문가 정보를 협업적 추천에 반영하기 위한 연구가 필요한데 현재의 SNS에서 공유되는 정보가 일상적이고 전문성이 결여된 내용이어서 전문가 추천을 위한 세밀한 전문가 선별 알고리즘 연구가 필요하다.

추천 시스템의 내용기반 기법(Content-based Method)은 사용자의 속성 분석을 통해 사용자가 추천한 아이템과 사용자와의 관련성을 분석하여, 각 아이템의 유사도를 반영한 점수 예측을 통해 상위 아이템을 추천하는 기법이고,.

협업적 여과 기법은 특정 사용자가 추천한 아이템을 추천한 사용자 클러스터 내의 점수 정보를 기반으로 타 아이템에 대한 점수를 예측하여, 상위 아이템을 추천하는 기법이다. 내용기반 추천 알고리즘은 개인적 취향에 대한 분석은 뛰어나지만, 특정 개개인의 정보 유사성은 고려하지 않는다. 협업적 여과 기법 알고리즘은 내용기반 기법의 부



<그림 1> 추천 시스템 알고리즘

수적인 정보가 불필요하지만, 결측된 정보가 있을 경우 해당 아이템에 대한 점수 예측이 힘들다. 이러한 두 알고리즘의 단점을 보완하기 위해 하이브리드 협업적 여과 기법이 제시되었으며, 대표적인 기법으로 Content-boosted Collaborative Filtering(CBCF)[1]이 존재한다. CBCF 기법은 기존 협업적 여과 기법의 희소성 문제를 내용기반 분석을 통해 해결하는 기법이지만 유사도 측정 시, 공통으로 점수가 매겨진 부분에 비해 해당 부분을 제외한 영역이 클 경우 신빙성이 떨어진다는 단점이 존재한다. 따라서 본 논문에서 제안하는 기법은, 첫째, 유사도 측정 시에 발생하는 문제점을 해결하기 위해 CBCF기법에서 사용한 피어슨 상관계수식을 보완한 새로운 유사도 측정 공식을 제안한다. 둘째, 사용자의 전문성을 판단할 수 있는 프로필 정보의 분석및반영을 통한, 예측 점수 테이블을 제안한다.

2장에서는 관련연구를 살펴보고, 3장에서는 전문가 추천 시스템을 위해 고려해야 할 속성들에 대한 설명과, 하이브

리드 협업적 여과 기법을 이용한 전문가 추천 시스템의 세부적인 알고리즘의 설계를 설명한다. 4장에서는 결론에 대해 설명하고, 향후 연구 방향을 제시한다.

2. 관련 연구

본 장에서는 기존 추천 시스템에서 가장 폭넓게 사용중인 협업적 여과 기법에 대해 설명한 후, 이를 활용한 전문가 추천 기법에 대해 기술한다.

협업적 여과 기법이란, 사용자들로부터 얻은 선호도 정보에 따라 특정 사용자의 관심사를 예측하는 방법을 의미한다. 예를 들어, 사용자 U_a가 A, B, C라는 아이템 추천 점수 데이터를 가지고 있다고 가정하자. 협업적 여과 기법은 U_a와 유사한 사용자 집합을 찾아 해당 사용자들의 정보를 기반으로 아이템을 추천한다. 만약, 사용자 U_b가 A, B, C, D에 대한 점수 데이터를 가지고 있고, U_a와 유사하다고 판단되면, U_a에게 아이템 D를 추천한다. 하지만 각 데이터를 반영하는 방법에는 세부적인 기법마다 차이점이 존재한다. 협업적 여과 기법은 크게 메모리 기반, 하이브리드 기반 기법으로 구분할 수 있으며 2.1~2.2 절에서 각각에 대해 기술한다.

2.1 메모리 기반 협업적 여과 기법

메모리 기반 협업적 여과 기법은 사용자가 아이템에 대해 점수를 부여한 데이터를 바탕으로, 유사한 패턴을 보이는 이웃 사용자 및 아이템을 찾아 사용자 본인과 이웃 사용자간의 연관성을 측정하고, 연관성이 높게 측정된 이웃 사용자가 점수를 준 아이템들의 선호도를 계산한다. 그 다음 사용자 본인의 이용 경험이 없는 아이템에 대한 선호도를 예측하기 위해 이웃 사용자들의 계산된 선호도 결과에 근거하여 높은 값을 보이는 아이템을 추천한다.

메모리 기반 협업적 여과 기법에서는 유사도를 계산하기 위해 주로 Pearson's Corr-elation Coefficient[2]를 사용한다. 식 (1)은 사용자 간의 유사도를 계산하기 위한 식을 나타낸다.

$$sim(a,u) = \frac{\sum_{i \in I}(R_{a,i} - \overline{R_a})(R_{u,i} - \overline{R_u})}{\sqrt{\sum_{i \in I}(R_{a,i} - \overline{R_a})^2}\sqrt{\sum_{i \in I}(R_{u,i} - \overline{R_u})^2}} \quad (1)$$

여기서, sim(a,u)는 사용자 U_a와 U_u와의 상관계수를, $R_{a,i}$와 $R_{u,i}$는 각각 U_a와 U_u의 아이템 i에 대한 점수를 의미하고, $\overline{R_a}$와 $\overline{R_u}$는 각각 I 집합에 있는 아이템에 대한 사용자 R_a와 R_u의 평균 점수를 의미한다.

피어슨 상관계수 등을 이용하여 이웃 사용자들 간의 유사도를 계산한 후, 실제 아이템들에 대한 점수 예측[3,4]은 식 (2)을 통해 계산한다. 값의 예측을 위해서는 K명의 가장 가까운 이웃 집단(KNN : K-Nearest Neighbor)이 필요하며, 식을 통해 계산된 값을 이용하여 사용자 U_u에게 예측 선호도 점수가 높은 아이템을 추천한다.

$$\check{R}_{a,j} = \overline{R_a} + \frac{\sum_{u \in KNN}(R_{u,j} - \overline{R_u}) \times sim(a,u)}{\sum_{u \in KNN}|sim(a,u)|} \quad (2)$$

여기서, $\widetilde{R_{a,j}}$는 U_a의 아이템 j에 대한 예측 선호도 점수를 의미한다. $R_{u,j}$는 예측하려는 아이템 j에 대해 이웃 집단의 U_u가 준 점수를 의미하고, $\overline{R_a}$는 U_a가 점수를 준 아이템들에 대한 평균 점수를 의미한다. 예측하려는 아이템 점수는 U_a와 U_u와의 상관계수(sim(a,u))를 반영하여 계산한다.

메모리 기반 협업적 여과 기법은 비교적 높은 예측 성능을 보이지만, 새로 확장된 데이터에 대한 Cold-Start 문제와 사용자-아이템 데이터의 희소성에 관련하여 심각한 제약을 갖는다. Cold-Start 문제란, 시스템에 새로 가입한 사용자와 같이 유사성을 판단하기 위한 아이템을 전혀 가지고 있지 않는 사용자의 경우 상관관계를 계산할 수 없는 것이고, 희소성 문제란 서로 동시에 점수를 매긴 아이템의 수가 적어 추천의 정확도가 낮아지는 문제를 말한다.

2.2 하이브리드 협업적 여과 기법

하이브리드 협업적 여과 기법은 기존 협업적 여과 기법의 단점을 보안하고자 둘 또는 그 이상의 협업적 여과 기법을 결합하여 보다 정확한 예측을 제공하는 추천 기법이다. 하이브리드 협업적 여과 기법의 대표적인 연구로는 협업적 여과 기법과 내용기반 여과 기법을 결합한 기법인 Content-Boosted Collaborative Filtering(CBCF)[1]이 존재한다. CBCF는 기존 협업적 여과 기법의 희소성 문제 및 Cold-start 문제를 내용기반 여과 기법의 사용자 속성 분석을 통해 해결한다.

CBCF 연구는 내용기반 여과와 함께 피어슨 상관계수를 이용하여 사용자에게 영화를 추천한다. 이를 위해 영화간 거리 계산을 위한 속성을 정의하고, 이를 통해 내용기반 분석을 수행한다. 또한, 도출된 각 속성에 회귀분석을 통한 가중치 값을 설정하여, 식 (3)을 통해 내용기반 유사도를 측정한다.

$$ContentSim(I_i, I_j) = \sum_{n=1}^{m} w_n \times f(A_{ni}, A_{nj}) \quad (3)$$

여기서, ContentSim(I_i, I_j)는 두 아이템 I_i, I_j 간의 내용기반 유사도를, w_n은 가중치 값을 의미하고, $f(A_{ni}, A_{nj})$는 두 사용자 사이의 한 속성(A_n)에 대한 유사도를 계산하는 함수이다.

측정된 내용기반 유사도는 협업적 여과 기법을 통해 계산된 유사도와 결합하여, 추후 CBCF의 점수 예측 알고리즘에 활용된다.

CBCF의 또 다른 특징 중 하나는 로컬 사용자 유사도 및 글로벌 사용자 유사도 적용이다. 로컬 사용자 유사도는 사용자들 간의 직접적인 유사도 관계를 나타내는 것으로, 로컬 사용자 유사도를 측정하기 위한 식은 다음과 같다.

$$ColabSim(a,u) = \frac{Min(|I_a \cap I_u|, \gamma)}{\gamma} sim(a,u) \quad (4)$$

여기서, ColabSim(a,u)는 사용자 U_a와 U_u의 로컬 사용자 유사도를, γ는 가중치를 위한 임계값, sim(a,u)는 사용자

U_a와 U_u사이의 피어슨 상관계수를 의미한다.

식 (4)을 통해 사용자들 간의 로컬 사용자 유사도를 구한 후, 다른 경로를 통해 사용자 U_a, U_u에 대한 글로벌 사용자 유사도 [5]를 계산한다. 만약, U_a에서 U_u로 이어지는 어느 경로상의 각 사용자 U_k 간의 유사도 값이 U_a와 U_u 간의 로컬 사용자 유사도 값보다 크다면, 해당 값을 U_a와 U_u의 글로벌 유사도로 결정한다. 이를 통해, 사용자에게 추천할 수 있는 이웃 그룹을 보다 확장하여, 희소성 문제를 해결하고 추천의 질을 높일 수 있다. 글로벌 사용자 유사도는 U_a와 U_u 사이에 직접적 연결성이 없을 때 주로 사용된다.

결과 테이블들을 토대로 실제 점수 예측은 Effective Missing Data Prediction (EMDP)[6] 알고리즘을 통해 수행한다. EMDP 알고리즘은 사용자 유사도 정보, 아이템 유사도 정보의 유무에 따라 총 네 가지의 결과를 통해 결측 정보에 대한 점수 데이터를 예측 및 반영하고, 실제 사용자-아이템간의 점수를 통해 상위 점수를 가지는 아이템을 추천한다.

하지만 CBCF는 식 (4)에서 사용된 실험적 변수 γ는 임의로 설정된 값으로 최적의 값을 설정하기 어려우며, 또한 만약 γ가 실제 사용자의 수보다 아주 작은 값을 가질 경우(e.g. 만 개의 아이템 중 20), 즉, 공통으로 점수가 매겨진 부분에 비해 해당 부분을 제외한 영역이 클 경우 결과의 신빙성이 떨어지는 단점이 존재한다.

2.3 SNS 상에서 전문가 추천 기법

전문가 추천에 대한 기존의 연구들 중 SNS를 기반으로 한 전문가 추천 연구는 초기단계에 있고, 협업적 여과 기법을 사용하여 전문가를 추천한 연구는 존재하지 않는다. 국내의 연구중 박상원[7] 등이 제안한 온톨로지 기반 소셜 네트워크 분석을 이용한 전문가 추천 시스템은 온톨로지의 추론을 통해 기존에 없던 결론을 예측, 도출해낼 수 있는 장점이 있지만, 결측 데이터가 존재할 경우 온톨로지 구축이 어렵다는 문제점이 있다.

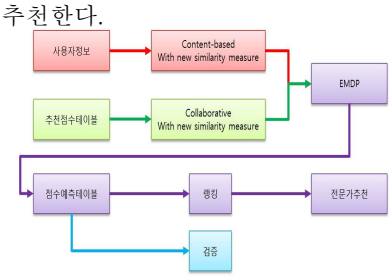
그리고 한승민[8] 등이 제안한 그리드 상의 소셜-네트워크를 이용한 전문가 검색 시스템은 빠른 검색이 가능하다는 장점이 있지만, 연구 방향이 그리드 환경과 크롤링 머신에 맞추어 있어, 전문가를 추천하기 위한 점수 데이터를 활용하지 못하는 문제점이 존재한다.

3. 전문가 추천을 위한 하이브리드 협업적 여과 기법

본 연구에서는 코사인 유사도를 적용한 유사도 계산 및 전문가 가중치를 이용한 추천 기법을 사용하여 설계한다. 본 장에서는 설계하고자 하는 기법을 위한 시스템 구조에 대해 설명하고, 전문가 추천을 위해 고려해야 할 사용자 속성, 코사인 유사도 기반 전문가 가중치 추천에 대해 기술한다. 또한, 이를 바탕으로 하이브리드 협업적 여과 기법을 이용한 전문가 추천 기법 알고리즘을 설계한다.

3.1 시스템 구조

본 시스템에서는 사용자가 입력한 사용정보를 활용하여 내용기반 유사도를 측정하고 추천점수테이블을 활용하여 협업적 여과 유사도를 측정한다. 계산된 결과 값을 기반으로 결측된 데이터에 대하여 EMDP[6]을 활용하여 결측 데이터에 대한 값을 설정하여 점수예측 테이블을 구성하고, 구성된 테이블을 이용한 검증을 통해 랭킹을 설정하여 전문가를 추천한다.



<그림 2> 전문가 추천 시스템

3.2 전문가 추천을 위한 사용자 속성

내용기반 분석을 위해서는 전문가 간 거리계산을 위한 사용자 속성이 필요하다. 이를 위해 <표 1>과 같이 전문가 속성 정보를 구성한다.

Dimension	Type	Domain	Distance measure
전문분야	String / String list	Computer, Graphic, etc.	1, 0 / $\lvert C_1 \cap C_2 \rvert / C_1$
관심분야	String list	Computer, Graphic, etc.	$\lvert C_1 \cap C_2 \rvert / C_1$
출신지역	string	Seoul, Pusan, etc	1, 0
종사지역(회사)	string	Seoul, Pusan, etc	1, 0
직장명	string	Samsung, LG, etc	1, 0
경력	Integer	[1,50]	1/abs or linear function
학력	String list	Ph.D, Master, Under, high, etc	1, 0
인맥의 수	Integer	[1,)	
나이	Integer	[1, 120]	1/abs or linear function
게시 글 수	Integer	[1,)	
답변 글 수	Integer	[1,)	
논문 수	Integer	[1,)	
rating	Integer	[1, 5]	
국적	String	Korea, USA, etc	1, 0

<표 1> 전문가 속성 정보

전문가를 추천하기 위해서는 각 사용자가 작성한 전문분야 및 관심분야 정보가 필수적이고, 지역성을 하기 위해서는 출신지역, 종사지역, 직장명과 같은 정보가 필요하다. 그 외 경력은 전문가의 전문분야 종사 기간을 위한 속성, 학력은 전문가의 최종학력을 위한 속성, 인맥의 수는 각 분야에서 활동적인 전문가를 선별하기 위한 속성, 나이는 연령대별 전문가 추천을 위한 속성으로 활용한다. 한편, 분야에 대한 게시 글 수, 다른 사용자들이 분야에 대해 질문한 글에 답변한 글의 수, 분야에 관련된 논문 작성 수는 전문성을 판단하기 위한 가중치 속성으로 활용한다. 마지막으로 점수(Rating) 테이블 데이터는 각각의 사용자들이 해당 전문가에게 점수를 부여한 속성으로 활용된다.

전문분야와 관심분야 속성은 비교대상이 되는 사용자와 같은 분야를 얼마나 공유하고 있는지에 따라 유사도를 결정한다. 한편, 출신지역, 종사지역, 직장명은 사용자 간에 서로 같은 값을 가지면 1, 다른 값을 가지면 0을 거리 값으로 설정한다. 다음으로, 해당 값들을 식(3)에 대입하여 내용기반 유사도(Content Similarity)를 측정한다. 또한 논문 수, 게시 글/답변 글 수, 인맥 수, 경력, 학력은 전문성을 판단하기 위해 사용되며 weight를 통해 계산된다. 마지막으로, 전문가 간의 내용기반 거리는 식 (5)을 통해 계

산된다.

$$result_{u,i} = EMDP \times weight \quad (5)$$

3.3 코사인 유사도 기반 전문가 가중치 추천

기존 CBCF에서 로컬 사용자 유사도 계산 시 사용한 식(4)에서는 임의로 설정된 γ를 사용한다. 사용자 a와 u가 공통으로 부여한 아이템 I에 대한 수가 γ 값보다 크면 $\frac{\gamma}{\gamma}$, 즉, 1을 반환하고, 작으면 $\frac{|I_a \cap I_u|}{\gamma}$ 을 반환하여 피어슨 상관계수 값을 줄인다. 하지만 앞서 언급한 바와 같이 γ는 임의의 값으로 최적의 값을 선택하기 어렵다. 따라서 본 연구에서는 이러한 CBCF의 문제점을 해결하기 위해 코사인 유사도를 통해 피어슨 상관계수에 가중치를 부여한다. 식(6)은 사용자간 코사인 유사도를 계산하기 위한 식을 나타낸다.(아이템 간 코사인 유사도 계산식은 이와 유사하기 때문에 생략한다.)

$$cosweight(a,u) = \frac{\sum_{i \in I_a \cap I_u} R_{i,a} \times R_{i,u}}{\sqrt{\sum_{i \in I_a \cap I_u} (R_{i,a})^2 (R_{i,u})^2}} \quad (6)$$

여기서, $R_{i,a}$는 사용자 a가 부여한 아이템 I에 대한 수이고, $R_{i,u}$는 사용자 u가 부여한 아이템 I에 대한 수이다.

위 식으로 코사인 유사도가 계산되면 식(7) 및 식 (8)을 통해 로컬 유사도를 계산한다.

$$addsim(a,u) = cosweight(a,u) \times sim(a,u) \quad (7)$$

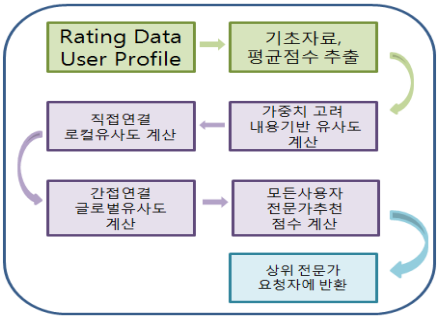
$$addsim(i,j) = cosweight(i,j) \times sim(i,j) \quad (8)$$

여기서, 식 (7)의 addsim(a,u)는 코사인 유사도를 적용한 사용자 a와 u의 로컬 사용자 유사도를 나타내고, sim(a,u)는 사용자 a와 u사이의 피어슨 상관계수를 의미한다.

제안하는 기법은 코사인 유사도를 사용함으로써 다음과 같은 장점을 보인다. 기존 CBCF의 단점인 γ 값에 따라 유사도가 바뀌는 문제점을 해결할 뿐만 아니라, 코사인 유사도를 사용함으로써 유사도의 편차를 줄여 기존 CBCF에서 제안한 최소반영 가중치보다 정확도를 향상시킬수있다.

3.4 하이브리드 협업적 여과 기법을 이용한 전문가 추천 기법 알고리즘 설계

앞서 기술한 3.2 전문가 추천을 위한 사용자 속성 및 3.3 코사인 유사도 기반 전문가 가중치 추천을 기반으로 SNS 상에서 전문가 추천을 수행하기 위한 알고리즘은 <그림 4>와 같다.



<그림 4> 전문가 추천 알고리즘

4. 결론 및 향후 연구

본 연구에서는 기존의 CBCF 알고리즘이 가진 가중치 식에 대하여, 실험적 변수로서 명확하게 쓰일 수 없는 점과 특정 상황에서 가중치의 역할을 제대로 수행하지 못하는 문제점을 제시한다. 이에 따라 로컬 사용자 유사도 및 아이템 유사도에서 사용하는 CBCF의 기존 식을 수정하여 코사인 유사도를 가중치로 설정하는 알고리즘을 새로이 제시하였다.

향후 연구로 전문가 추천을 위해 제시한 알고리즘을 적용한 시스템의 구현과 실제 점수 자료에 기반한 성능평가가 필요하다.

참고 문헌

[1] Gözde Özbal, H˙ilal Karaman and Ferda N. Alpaslan, “A Content-Boosted Collaborative Filtering Approach for Movie Recommendation Based on Local and Global Similarity and Missing Data Prediction”, The Computer Journal, Volume54, Issue9 Pp. 1535-1546, 2011.

[2] Sarwar, B., Karypis, G., Konstan, J., and Reidl, J., “Item-based Collaborative Filtering Recommendation Algorithms,” Proceedings of the 10th International World Wide Web Conference, ACM Press, pp. 285-295, 2001.

[3] Massa, P. and Avesani, P., “Trust-aware Collaborative Filtering for Recommender Systems,” Proceedings of International Conference on Cooperative Information Systems, LNCS 3290, Springer, pp. 492-508, 2004.

[4] Papagelis, M., Plexousakis, D., and Kutsuras, T., “Alleviation the Sparsity Problem of Collaborative Filtering Using Trust Inferences,” Proceedings of the 3rd International Conference on Trust Management, LNCS 3477, Springer, pp. 224 - 239, 2005.

[5] Floyd, R.W., "Algorithm 97: shortest path," Commun. ACM, 5, 345 - 348, 1962.

[6] Ma, H., King, I. and Lyu, M.R., "Effective Missing Data Prediction for Collaborative Filtering," ACM SIGIR'07, 39-46, 2007

[7] 박상원, 최은정, 박민수, 김정규, 서은석, 박영택, “온톨로지 기반 소셜 네트워크 분석을 이용한 전문가 추천 시스템,” 정보과학회, 컴퓨팅의 실제 및 레터, 제 15권 제 5호, 2009.

[8] 한승민, 허의남, 이필우, 이승연, “그리드 상의 소셜-네트워크를 이용한 전문가검색 시스템,” 한국인터넷정보학회, 정기총회 및 추계학술발표, 제 9권 제 2호, pp.317-321, 2008.