사회적 영향력과 어의 유사도 분석에 기반한 가치정보의 추천 기법

김명훈, 김상욱
경북대학교 컴퓨터학부 Smart Life 실현을 위한 SW 인력양성사업단
경북대학교 소프트웨어기술연구소
e-mail : mhkim@media.knu.ac.kr

Social Influence and Semantic Similarity Concerned Recommendation Technique of Qualitative Information

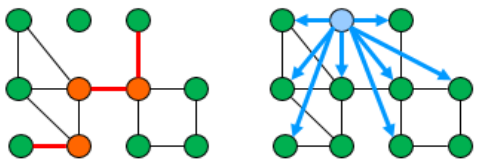
MyeongHun Kim, SangWook Kim
SW Human Resource Development Program for Supporting Smart Life,
School of Computer Science and Engineering, Kyungpook National University
Software Technology Research Center (SWRC), Kyungpook National University

요 약

추천 기법은 개인의 관심사와 상황을 고려한 개인화된 아이템을 제공함으로써 아이템의 소비 과정에서 발생하는 부하를 줄여주고 정보 소비의 효율성을 증대시키는데 중요한 역할을 한다. 본 연구에서는 전통적인 추천 기법인 Content-Based(CB)기법과 최근 온라인 소셜 네트워크의 경향을 반영한 Social Network-based(SN)기법을 접목하여 새로운 복합방식의 정보 추천 기법을 제시한다. CB 기법의 대표적인 한계점인 cold start problem 과 SN 기법의 추천 아이템의 전문성 문제를 상호 보완하며, 특히 최근 소셜 네트워크의 특징인 비신뢰(non-trust) 기반의 영향력 있는 정보 확산자가 존재하는 환경에서 기법을 적용할 수 있도록 하였다. 또한 대부분 사람 추천 중심인 기존의 SN 기법들과는 달리 사람에게 제공할 정보의 추천에 초점을 두며, 정보 선정과정에서 개인의 온라인과 현실(real world)에서의 사회활동 정보를 모두 활용하여 더욱 더 개인화된 가치정보를 제공하고자 한다.

1. 서론

최근 온라인 소셜 네트워크의 특성은 긴밀한 사회적 유대의 형성보다는 정보 확산을 목적으로 대량의 정보를 불특정 다수에게 제공하는 노드가 등장하고 있다. 이러한 노드는 그래프 이론에서 정의하는 특정 노드의 삭제가 전체 네트워크에서 connected components 의 개수를 증가시키는 Bridge Node[1]와는 달리, 수 많은 간선으로 노드들과 연결되어 있지만 노드의 삭제가 전체 네트워크의 구조를 변형시키지 않고 단방향으로만 정보 확산을 일으킨다.



(그림 1) 브릿지 노드(a)와 영향력 확산자(b)의 구조

그림 1 (b)와 같이 많은 인접 노드와 연결되어 있지만 간선을 제거해도 전체 네트워크에서 connected components 의 개수는 변하지 않는데, 이런 노드가 공급하는 정보의 양이 증가하면서 수신자는 오버헤드를 경험한다. 따라서 정보의 적절한 선별적 소비와 정보 과잉공급으로 소비에서 누락된 가치정보를 찾아 재공급하여 정보 소비의 기회를 증대시킬 필요가 있다.

본 연구에서는 이러한 선별 및 재 공급할 가치가 있는 정보를 탐지하는 방법에 있어 전통적인 Content-Based(CB) 추천 기법과 온라인 소셜 네트워크 환경에 적합한 Social Network-Based(SN) 추천 기법을 접목하고, 추가로 실 세계(real world)의 친화도를 반영하여 보다 신뢰성 있는 알고리즘을 제안하고자 한다.

2. 관련 연구

일반적으로 추천 기법은 CB 와 CF(Collaborative Filtering-based), Hybrid 로 나뉘며[2], 최근 소셜 미디어의 발전과 더불어 SN(Social Network-based) 방식이 대두되고 있다[3]. CB 방식은 한 사람이 과거에 선호한 정보, 특히 텍스트로 표현되는 뉴스, 책, 문서, 댓글 등을 수집하여 이와 유사한 컨텐츠의 정보를 추천하며 주로 키워드로 정보의 특성을 표현하는데 TF-IDF 로 정량적인 가중치를 부여한다[4]. CB 기법의 몇 가지 한계점은 다음과 같다: (1) 추천 아이템이 지나치게 전문화(overspecialized)되는 경향이 있어 추천의 범주

가 좁고 환경 변화에 유연하지 못하다; (2) 신규 사용자의 등장 시 즉각적인 추천이 힘들다. Cold-start problem[5] 이라 불리는 이 문제는 CB 기법이 사용자의 개인화된 특성을 분석하기 위해서는 충분한 수의 과거 사용자 정보를 사전에 확보해야 함을 의미한다.

SN 기법은 통화, 메시지, 이메일 등의 사회적 행위를 기반으로 특정인과의 관계나 아이템을 추천하는 것을 말하며, 대표적으로 TidalTrust[6], SocialMF[7], LOCALBAL[8]가 있다. SN 방식은 전통적 기법과는 달리 현재 상황의 정보(context)가 결정적인 역할을 하므로 사용자의 과거 정보에 대한 의존도가 적고 추천 아이템의 범주 또한 넓은 특성이 있다.

본 연구에서는 온라인 소셜 네트워크 환경에서 정보 전달의 특성, 즉 사용자가 누구의 정보를 적극적으로 전달하는가를 조사하여 온라인에서의 친화도를 측정하며, 더불어 이 노드와 실세계(real world)에서 어떤 친화도를 가지는지를 추가적으로 조사하여 보다 더 정확한 사회적 친화도를 산출한다. 더불어 개인이 받는 정보의 컨텐츠 특성을 정량적으로 판별하기 위해 TF-IDF 를 사용한 CB 기법으로 사용자가 선호하는 정보의 특성을 도출하며 최종적으로 앞의 두 방법을 접목하여 각 정보에 대한 점수를 할당(Rating)하고 이를 기준으로 추천 정보를 선정한다. CB 방식의 cold-start problem 을 SN 방식이 보완하고 또한 CB 기법으로 SN 기법에 추천 전문성(specialization)을 더하여 최종 추천 아이템의 정확도를 높이게 될 것이다.

3. 가치 정보의 식별 알고리즘

3.1 Qualitative Information 의 정의

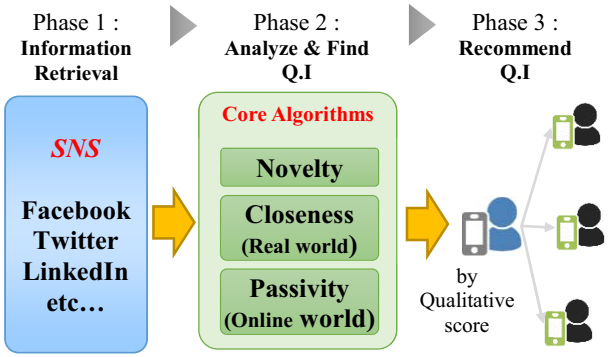
본 연구에서 최종적으로 사용자에게 제공하는 정보를 가치정보(*Qualitative Information, QI*)라 명명하며, 정보의 가치를 표현하는 식(1)의 값이 4 장에서 언급할 Threshold 값을 초과하는 정보를 *QI* 라 정의한다.

$$Qualitative\ Score = Novelty + Affinity \quad (1)$$

$$Affinity = Affinity_{real} + Affinity_{online} = Closeness - Passivity \quad (2)$$

3.2 Qualitative Information 의 판정 방법

그림 2 와 같이 온라인 소셜 네트워크의 모든 수신 정보를 수집한 후(Phase 1) 각 정보의 Novelty, Closeness, Passivity 의 Score 를 도출하고(Phase 2), 이 세 가지 Score 를 종합한 Qualitative Score(QS)를 기준으로



(그림 2) 제안하는 정보 추천 시스템

최종 추천 정보를 결정한다(Phase 3). 일정 값 이상의 QS 를 가지는 정보를 다시 제공함으로써 소비되지 않은 정보는 다시 소비될 기회를 가지고 더불어 다른 노드로 확산될 확률이 높아진다. Novelty 는, Passivity 는 각각 CB, SN 기법을 근간으로 한다. 다음 서브 챕터에서 구체적인 알고리즘을 정의 한다.

3.3 Novelty

0 에서 1 사이의 값을 가지는 $\boldsymbol{Novelty_i}$는 사용자가 받은 정보 $\boldsymbol{i}$의 컨텐츠 신규성을 나타내며, 컨텐츠 익숙도를 나타내는 $\boldsymbol{Familiarity}$의 역수가 된다(식 (3)).

$$Novelty_i = \frac{1.0}{Similiarity_i \times \sum_{k=1,k\in E}^{n} W_{extracted_i}} \quad (3)$$

식(3)의 $Similarity$는 수신자가 과거에 받은 누적된 정보와 개별 정보를 토큰으로 나누어 TF-IDF 에서 변형된 TF-IIDF(TF-Inverse IDF)로 가중치 벡터를 표현하고 이 두 벡터를 cosine similarity 로 유사도를 계산한 값이다. 다수의 누적 정보에 대한 벡터를 $\vec{W}_{extract} = \{ts_1, ts_2, ts_3, ts_4, ts_5, \dots\}$, 개별 정보의 벡터를 $\vec{W}_{document} = \{ds_1, ds_2, ds_3, ds_4, ds_5, \dots\}$ 라 할 때, TF-IIDF 는 식(4, 5)과 같이 정의되고 $Similarity = \vec{W}_{extract} \cdot \vec{W}_{document}$가 된다. TF-IDF 와는 달리 각 벡터의 성분 값은 토큰의 발생빈도수(tf_t)와 토큰 t가 발생한 정보의 수(df_t)에 비례한다.

$$ts_n = tf_t \times iidf_t = \frac{tf_t}{\log \frac{N}{df_t}} \quad (4)$$

$$ds_n = tf_{t(d)} \times iidf_t = \frac{tf_{t(d)}}{\log \frac{N}{df_t}} \quad (5)$$

3.4 Closeness (Affinity in real world)

0 에서 1 사이의 값을 가지는 $Closeness$는 현실에서의 사회적 활동기록을 나타내는 폰의 통화, SMS, 이메일의 사용내역을 바탕으로 각 노드 n과의 친화도를 측정한 값이다. 식(6)과 같이 노드 n과의 $Closeness$는 커뮤니케이션의 유형($type$)별 커뮤니케이션의 강도($COINT$)의 합으로 산출하며, $COINT$는 값은 식(7)과 같이 5 개의 변수로 구성된다. 커뮤니케이션의 유형은 $type = email, SMS, phone$으로 3 가지로 정의한다.

$$Closeness_n = Affinity_{real_n} = \sum_{type} COINT_{n(type)} \quad (6)$$

$$COINT_n = ATI_n \times F_n \times D_n^* \times \frac{\sigma_n}{COEFF^{MRTI_n}} \quad (7)$$

ATI는 활동적으로 커뮤니케이션을 한 시간적 간격(Active Time Interval), F는 전체 커뮤니케이션 횟수(Frequency), D는 통화시간(Duration), σ_n는 커뮤니케이션의 산포, $MRTI_n$는 커뮤니케이션의 최근 시점(Most Recent Time Interval)으로 현 시점과의 시간적 차이($|current\ time - event\ time|$)를 나타낸다.

3.5 Passivity (Affinity in online world)

0 에서 1 사이의 값을 가지는 Pas_{ij}는 온라인 소셜 네트워크 환경에서 사용자의 정보 전달 성향을 나타낸다. 즉 수신자의 정보 전달 유무는 정보 전달자와의 관계에 영향을 받으며, 수신자의 이웃 노드의 영향력을 각각 정량적으로 판별할 수 있게 한다.

$$Pas_{ij} = \varphi_{ij}\,\omega_{ij} \Big/ \sum_{k:(i,\ k)\in E} 1 - \tau_{ik} \qquad (8)$$

$$1 - \tau_{ik} = \varphi_{ik} = 1 - \frac{\sum Info_{spread_{ik}}}{\sum_{(i,\,k)\in E} Info_{ik}} \qquad (9)$$

Pas_{ij}는 정보 수신자 i의 정보 제공자 j에 대한 정보 전달 수동성을 나타내며 모든 이웃 노드의 거절률 $(\sum \varphi_{ik})$에 대한 노드 j의 거절률의 비로 표현한다(식 8). 거절률은 식(9)과 같이 i가 받은 k의 정보의 총 개수 $(\sum_{(i,\,k)\in E} Info_{ik})$와 i가 전달한 k의 정보의 총 개수$(\sum Info_{spread_{ik}})$로 산출한다.

4. 실험 결과 및 고찰

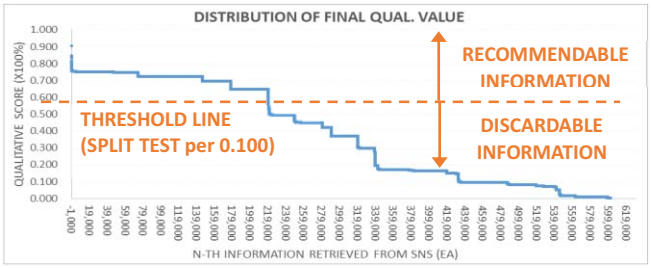
4.1 정보 재 공급 기준(Threshold)의 설정

본 실험의 목적은 제안하는 알고리즘이 실질적으로 사용자에게 정보를 추천하기 위한 기준 값(Threshold)을 설정하는 것이며 사용자가 확산한 정보의 개수가 가장 많은 경우로 판정하였다. 챕터 3 에서 정의한 3 가지 알고리즘에 기반한 Qualitative Score(QS)값이 모든 정보에 할당이 되고(그림 3), $QS \geq Threshold$인 정보를 사용자에게 재 공급한다.

실험은 아래 두 가지 방법으로 수행되었으며 다수 사용자의 다양한 Threshold 값들은 Least Square Method 로 하나의 수렴 값으로 표현하였다.

① 실제 SNS(Facebook)의 정보 700,970 개를 취득하여 한 사용자에 대한 Threshold 도출.
② 가상의 정보와 사용자를 생성하고 각 사용자당 700,970 개의 정보 제공 및 Threshold 도출.

각 실험은 한 사용자 당 0.0 ~ 1.0 사이의 Threshold 값을 0.1 단위로 구분하여 총 11 번씩 실시하였으며,



(그림 3) 수집한 정보의 최종 QS 값 분포

매회 신규 정보와 추천정보(QS≥Threshold)의 개수를 총 700,970 개로 구성하여 사용자에게 제공하고 이때 사용자가 다시 인접 노드로 전달한 정보의 개수를 관찰하였다. 그림 4 (a)와 같이 한 사용자의 소셜 네트워크 서비스(Facebook)에 존재하는 실제 정보를 기반으로 한 실험은 Threshold=0.7 인 경우에 가장 많은 정보를 전달했으며, 가상의 다수 사용자 (3644 명)에 대한 가상 정보 기반의 실험 결과 Threshold 값이 0.645 로 수렴하는 결과를 얻을 수 있었다(그림 4 (b)).

즉, $QS \geq 0.645$인 정보를 신규 정보와 함께 재 공급하면 최대 다수의 사용자가 수신 정보에 가장 활발한 반응을 보일 것이며, 이는 가장 많은 사용자에게 적절한(accurate) 정보를 추천 할 수 있음을 의미한다.

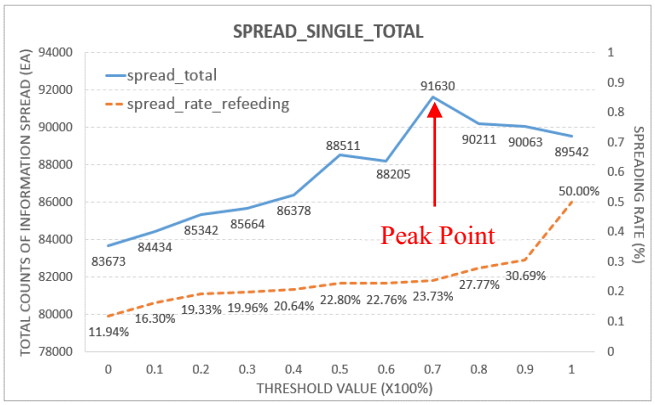
4.2 유사 기법과의 성능(SR 지수) 비교

제안 기법의 성능 평가는 챕터 3 에서 언급한 세가지 기법의 개별 평가와 종합 평가를 동시에 수행하며 아래 식 10 의 SR 지수로 정량적인 성능 비교가 될 수 있도록 설계하였다.

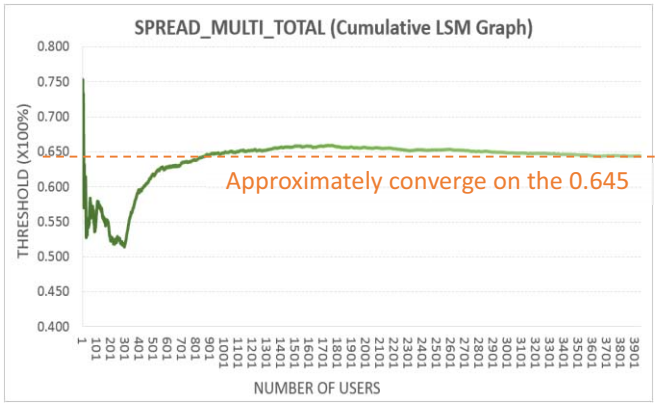
$$Spread\_Rate = \frac{\sum Post_{spread}}{\sum Post_{received}} \times 100\ (\%) \qquad (10)$$

Spread_Rate(SR) 지수는 한 사용자가 받은 전체 정보에서 인접 노드로 전달한 정보의 비율을 나타내며, 적절히 개인화된 정보를 공급 할수록 사용자는 정보 전달행위에 더 적극적이라는 가정을 바탕으로 한다.

Novelty 와 유사한 CB 기반의 Peaky Topic(PT)) 기법과 Affinity 와 같이 SN 기반의 신뢰도 측정 기법인 TidalTrust 를 제안기법과의 비교대상으로 선정하였으

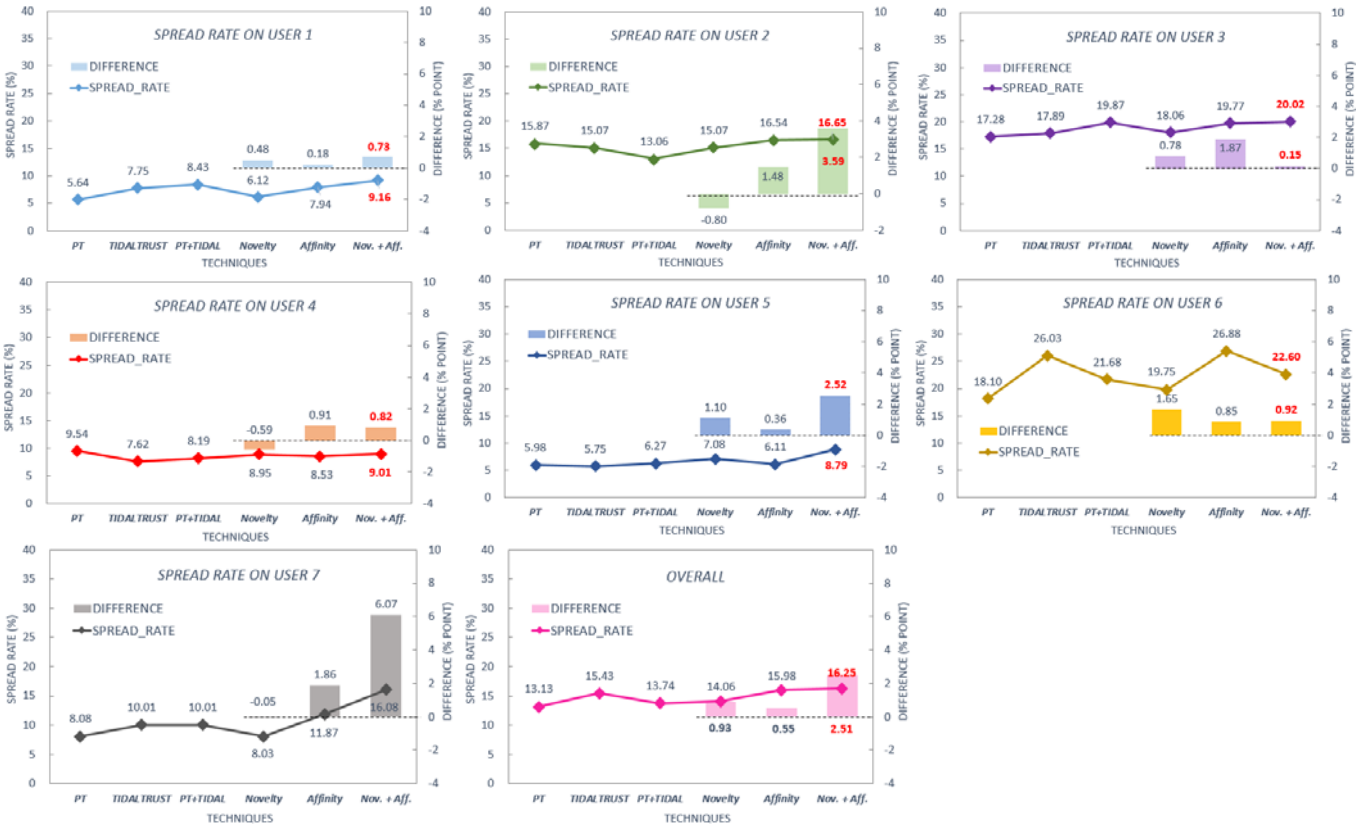


(a)



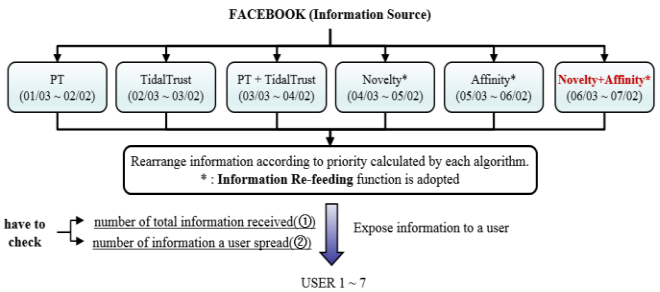
(b)

(그림 4) Threshold 값의 도출: 한 사용자의 실제 정보 기반(a), 다수 사용자의 임의 생성 정보 기반(b) 결과



(그림 6) SR 지수의 최종 평가 결과

며, 아래 그림 5와 같이 6개월 동안 각 기법들을 1달씩 적용하여 7명의 SR 지수를 측정하였다. 7명의 6개



(그림 5) 개별/종합 평가를 위한 방법

월간의 Facebook 사용기간 동안 총 505 만개의 정보, 12 만개의 간선을 수집하였으며 각 사용자 별로 인접 노드로 전달한 정보의 갯수를 조사하였다.

SR 지수의 측정 결과 Novelty, Affinity 의 개별, 종합 성능 평가에서 각각 0.93%, 0.55%, 2.51%P 더 높은 값을 얻을 수 있었으며, 표 1 과 그림 6 에 이를 요약하였다.

개별, 종합 성능 모두 유사 기법 대비 정보에 대한 사용자 반응이 더 적극적이었으며, 더 개인화된 정보를 제공하여 높은 정보 전달성을 유도할 수 있었다.

<표 1> SR 지수 측정 결과를 통한 성능 비교

CB			SN			Hybrid		
PT	No	Diff.	Tidal.	Aff.	Diff.	PT+Tidal	Nov.+Aff.	Diff.
13.13%	14.06%	**0.93%**	15.43%	15.98%	**0.55%**	13.74%	16.25%	**2.51%**

참고문헌

[1]B. Bollobás, Modern graph theory, volume 184 of Graduate Texts in Mathematics, Springer-Verlag, 1998.
[2]F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook, pp. 1-35, Springer US, 2011.
[3]I. Guy, D. Carmel, “Social recommender systems,” Proc. of the 20th international conference companion on World wide web, ACM, pp. 283–284, 2011.
[4]M. Balabanović, Y. Shoham, “Fab: content-based, collaborative recommendation,” Communications of the ACM, 40.3: 66-72, 1997.
[5]G. Adomavicius, A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” Knowledge and Data Engineering, IEEE Transactions on, 17.6: 734-749, 2005.
[6]J. Golbeck, Generating predictive movie recommendations from trust in social networks, Springer, 2006.
[7]M. Jamali, and M. Ester. "Trustwalker: a random walk model for combining trust-based and item-based recommendation," Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 397-406, 2009.
[8]J. Tang, X. HU, H Gao, H Liu, "Exploiting Local and Global Social Context for Recommendation." International Joint Conference on Artificial Intelligence, 2013.

사사

“본 연구는 미래창조과학부 및 정보통신기술진흥센터의 SW 중심대학지원사업의 연구결과로 수행되었음”