# UCC 추천을 위한 협업 필터링 기법과 사회 네트워크 분석의 통합적 활용 방안

『Integrated Usage of Collaborative Filtering Techniques and Social Network Analysis for UCC』

**김종우, 정중희, 오재훈**
**kjw@hanyang.ac.kr**

**Hanyang University, Dept. of Management Information System**

# CONTENTS

- 연구 배경
- 연구 동기
- 선행 연구
- 연구 목적
- 연구 방법
- 실험 결과
- 결론
- 향후 추진계획 및 연구과제

# 연구 배경

- 정보시스템 기술의 발달로 인하여 다양하고 방대한 정보가 인터넷 사용자들에게 제공
  - 방대한 정보 중에서 사용자가 필요로 하는 정보만을 걸러내기 위해서는 추가적인 시간과 노력이 필요

▪**추천 시스템(Recommendation System)**은 사용자의 정보 과부하를 줄여줄 수 있는 좋은 수단 [Kautz H., B. Selman, M. Shah, 1997, Moody, James and Douglas R. White, 2003]

# 연구 배경

- **추천시스템** : 사용자가 부여한 평점, 관심 콘텐츠 저장 등의 명시적 프로파일과 구매 이력, 행동 패턴 등의 잠재적 프로파일을 고려하여 필터링에 반영하며, 얼마나 정확하게 사용자의 선호도가 반영된 결과를 제시하느냐에 목적[Frederic P. Miller et al., 2009]

  - **협업 필터링** : 고객 자신의 선호 경향과 선호도가 유사한 이웃 고객을 선정하고 이들의 관계를 이용하여 선호도를 예측[Yehuda Koren, 2010]
  - **콘텐츠 기반 추천 :** 사용자의 선호 정보를 표현하는 사용자 모델과 아이템의 속성을 비교하여 유사도가 높은 아이템을 추천해 주는 방식[Michael J. Pazzani, Daniel Billsus, 2007]

# 연구 배경

- **사회네트워크 분석** : 의사소통 집단 내 개체의 상호작용에 관심을 두고, 개체 간 연결 상태 및 연결 구조의 특성을 계량적으로 파악하여 시각적으로 표현하는 분석기법[김용학, 2003, J Warmbrodt, 2008, Yusoon Kim, Thomas Y. Choi, Tingting Yan and Kevin Dooley, 2011]

| 척도 | | 의미 |
|---|---|---|
| 중심성 (Centrality) | 연결중심성 (Degree centrality) | 네트워크의 한 행위자가 몇 사람과 직접적으로 연결되어 있는지를 측정하는 지표 |
| | 근접 중심성 (Closeness centrality) | 직접적인 연결 뿐 아니라 간접적인 연결까지 포함해서 중심성을 측정하는 지표 |
| | 매개 중심성 (Betweenness centrality) | 순수하게 한 행위자가 중계자(브로커) 역할 만들 측정하는 지표 |
| | 권력 중심성 (Power Centrality) | 행위자의 in/out 중심성과 각 행위자가 연결한 행위자의 in/out 중심성 지수를 함께 고려하여 중심성을 측정하는 지표 |
| | 위세 중심성 (Eigenvector Centrality) | 한 행위자가 네트워크 내에서 중요한 위치에 있는 다른 행위자와 연결된 정도를 측정하는 지표 |
| 응집성 (Cohesion) | 밀도 (Density) | 네트워크 내 행위자들이 얼마나 많은 관계를 갖고 있는가를 파악할 수 있는 대표적 양적 지표 |
| | 중심화 (Centralization) | 행위자들이 중심성을 중심으로 그 주위를 조직화하고 있는 정도를 나타내는 지표 |

[표 1] 사회적 네트워크 분석의 대표적인 척도

# 연구 동기

**“CF 추천 시스템 + 사회 네트워크 분석”**

- **고객의 구매 데이터를 이용하여 고객과 제품 네트워크를 구축**

  **->일종의 사회 네트워크**

# 선행 연구 및 본 연구의 차별점

- 사회 네트워크 분석을 이용한 추천 시스템에 관한 연구가 다양한 방면으로 이루어지고 있음.
  - 협업필터링의 근간이 되는 유사도를 제외하고 사회 네트워크의 척도만으로 추천에 이용

"사회연결망 : 신규고객 추천문제의 새로운 접근법", 박종학, 조윤호, 김재경, 2009 : 협업필터링의 기반이 되는 유사도를 제외하고 구매여부와 연결정도 중심성만으로 신규고객 추천에 이용

"신상품 추천을 위한 사회연결망 분석의 활용", 조윤호, 방정혜, 2009 : 구매여부와 연결정도 중심성, 근접중심성, 매개중심성, 위세중심성을 이용하여 신상품 추천에 이용

"A model of a trust-based recommendation system on a social network", Frank Edward Walter, Stefano Battiston , Frank Schweitzer, 2007 : 소셜 네트워킹과 Trust-relation 을 결합하여 추천에 이용

"A Distributed Trust-based Recommendation System on Social Networks", Karan Sarda, Priya Gupta, Debdoot Mukherjee, Smruti Padhy, Huzur Saran, 2008 : Trust의 두 가지 양상(friendship-trust, domain-expertise)을 고려하여 추천에 이용

**차별점**

- CF 알고리즘에 사회적 네트워크의 척도를 결합적으로 활용
- 사용자 뿐만 아니라 상품의 네트워크의 척도를 활용

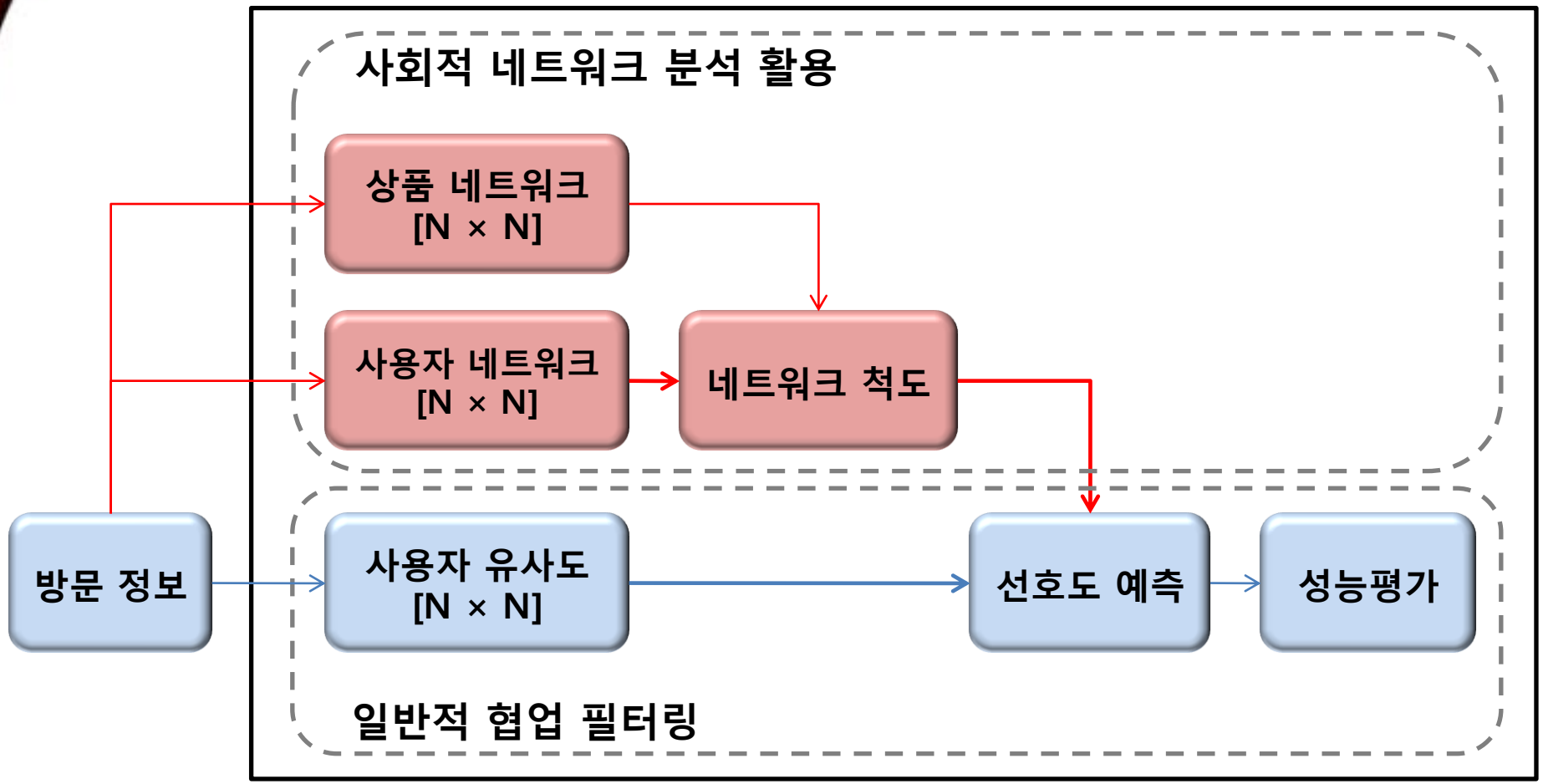
# 연구 목적

**『사회 네트워크 분석의 구조적 척도들을 활용하여
협업 필터링 기반 추천시스템의 성능을 높일 수 있는 방법들을 모색』**

- **연구질문 1** UCC를 이용한 사용자들의 사회 네트워크를 구성하고 분석하여, 이를 사용자 기반 협업 필터링에 어떻게 적용하여 추천의 성능을 높일 수 있는가?
  - **연구 질문1.1** UCC를 이용한 사용자들의 사회 네트워크 척도 중 어떠한 것이 추천 성능을 높일 수 있는가?
  - **연구 질문1.2** UCC를 이용한 사용자들의 사회 네트워크 척도를 어떻게 가중하여야 추천의 성능을 높일 수 있는가?

- **연구질문 2** 사용자들이 이용한 UCC들의 사회 네트워크를 구성하고 분석하여, 이를 사용자 기반 협업 필터링에 어떻게 적용하여 추천의 성능을 높일 수 있는가?
  - **연구 질문2.1** 사용자들이 이용한 UCC들의 사회 네트워크 척도 중 어떠한 것이 추천 성능을 높일 수 있는가?
  - **연구 질문2.2** 사용자들이 이용한 UCC들의 사회 네트워크 척도를 어떻게 가중하여야 추천의 성능을 높일 수 있는가?

# 접근 방법




[그림 1] 사용자 네트워크를 활용한 협업 필터링의 확장

# 연구 방법 (모형)

## 메타데이터 추출

사용자의 Web log 데이터

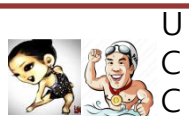
| 방문날짜 | 사용자 | URL | 체류시간 |
|---|---|---|---|
| 20100212 | 14324 | http://youtube.com/... | 38초 |
| 20100213 | 235 | http://youtube.com/... | 20초 |
| | | | |

| 방문날짜 | 사용자 | UCC 제목 | 태그 | 체류시간 |
|---|---|---|---|---|
| 20100212 | 14324 | 김연아 금메달 | 김연아, 피겨.. | 38초 |
| 20100213 | 235 | 소녀시대 GEE | 소시, 소녀시대,GEE | 20초 |
| | | | | |

메타 데이터 추출

## 협업필터링에 적용

사용자-UCC행렬 생성



사용자

UCC간 유사도 계산

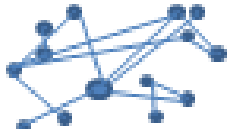
사용자간 유사도 계산

실험1.

사용자

사용자 네트워크

실험2.



척도계산

UCC 네트워크

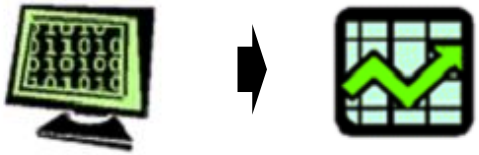
실험3.

척도계산

상품에 대한 고객 선호도 점수예측

고객

## 추천 성과 비교



통계분석을 통한 추천 성과 비교

[그림 2] 연구 과정

# 알고리즘(Algorithm)

- 사용자 선호도가 이진 데이터인 경우, 사용자간의 유사도 구하는 식
  : Jaccard Similarity Coefficient

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad \text{(식 1)}$$

- 사용자 선호도 예측 식

$$P_{ik} = \frac{\sum_{l \in Rater(k)} J(i,l)}{n(Rater(k))} \quad \text{(식 2)}$$

- 사회적 네트워크 척도를 활용한 경우
  - 사용자 네트워크 척도를 활용한 경우

$$P_{ik} = \frac{\sum_{l \in Rater(k)} (w_1 J(i,l) + w_2 c_l)}{n(Rater(k))}, w_1 + w_2 = 1 \quad \text{(식 3)}$$

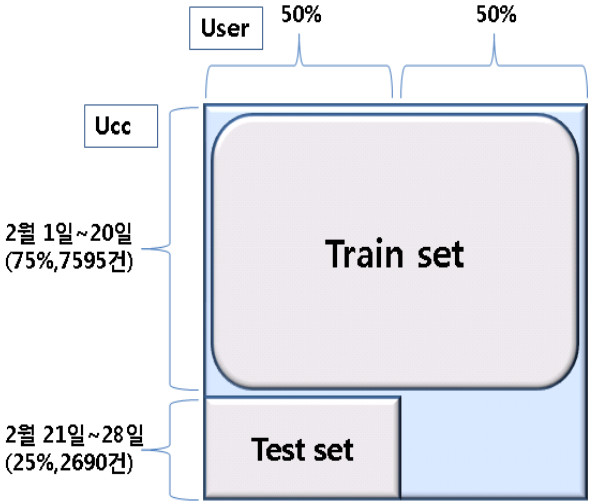
  - 상품 네트워크 척도를 활용한 경우

$$P_{ik} = \frac{w_1 \sum_{l \in Rater(k)} J(i,l)}{n(Rater(k))} + w_2 c_k, \; w_1 + w_2 = 1 \quad \text{(식 4)}$$

# 실험 데이터

- ㈜ 코리안클릭으로부터 2010년 2월 1달간의 www.youtube.com 에 방문한 패널의 웹 로그(Web log) 데이터를 제공받아 활용
- 방문날짜, 개인 정보를 알 수 없도록 임의로 수정된 패널 ID, 방문자 URL, 체류시간 등으로 구성 : 15, 496건의 페이지 정보
- 패널 데이터 : UCC를 한달 동안 10회 이상 시청한 239명

  상품 데이터 : 방문자 URL정보에서 추출한 9998개의 UCC 중에 239명의 패널들이 1회라도 시청한 7967개의 UCC를 실험에 사용
  - ❖ JAVA 프로그램 언어를 이용하였으며, DBMS로는 MS SQL Sever2005를 활용




[그림 3] 훈련 집합과 테스트 집합

# 실험 방법

## ▪실험방법 1(기존 CF)

(1) 사용자×UCC콘텐츠의 행렬을 이용하여 [그림 2]와 같이 사용자 네트워크를 생성한다.
(2) 사용자 네트워크를 통해 얻은 구조 정보로 고객의 유사도를 구한다.
(3) 이를 활용하여 상품에 대한 고객의 선호 예측 점수를 계산하고 성능을 평가한다.

## ▪실험방법 2(CF + 사용자 Network 척도 사용)

(1) 사용자×UCC콘텐츠의 행렬을 이용하여 [그림 2]와 같이 사용자 네트워크를 생성한다.
(2) 사용자 네트워크를 통해 얻은 구조 정보와 고객의 유사도 계산 결과를 가중 평균하여 수정된 사용자 유사도를 만든다.
(3) 이를 활용하여 상품에 대한 고객의 선호 예측 점수를 계산하고 성능을 평가한다.
(4) 가중치와 네트워크 구조 정보 변수를 수정해 가며 추천의 성능을 높일 수 있는 변수를 도출한다.

# 실험 방법

**▪실험방법 3(CF + 상품 Network 척도 사용)**

(1) 사용자×UCC콘텐츠의 행렬을 이용하여 [그림 2]와 같이 UCC 네트워크를 생성한다.

(2) UCC 네트워크를 통해 얻은 구조 정보와 고객의 유사도 계산 결과를 가중 평균하여 수정된 사용자 유사도를 만든다.

(3) 이를 활용하여 상품에 대한 고객의 선호 예측 점수를 계산하고 성능을 평가한다.

(4) 가중치와 네트워크 구조 정보 변수를 수정해 가며 추천의 성능을 높일 수 있는 변수를 도출한다.

# 실험 결과(5-Sample)

- **실험방법2의 결과(CF + 사용자 Network 척도 사용)**

| 실 험 | | 정 확 도 | | |
|---|---|---|---|---|
| | | Top 5 | Top 10 | Top 20 |
| 1 | Random | 0.000196 | 0.000196 | 0.000196 |
| 2 | 기존 CF | 0.006101 | 0.005469 | 0.003507 |
| 3 | CF + Degree Centrailty | 0.013474 | 0.009766 | 0.005962 |
| 4 | CF + Closeness Centraility | 0.008576 | 0.005813 | 0.004419 |
| 5 | CF + Betweeness Centrality | 0.007929 | 0.006101 | 0.003969 |
| 6 | CF + Eigenvector Centrality | 0.010415 | 0.007959 | 0.005644 |
| 7 | CF + Power Centrality | 0.006101 | 0.004298 | 0.002606 |

$$P_{ik} = \frac{\sum_{l \in Rater(k)} J(i,l)}{n(Rater(k))}$$

$$P_{ik} = \frac{\sum_{l \in Rater(k)} (w_1 J(i,l) + w_2 c_l)}{n(Rater(k))}, w_1 + w_2 = 1$$

[표 2] 기존 협업필터링에 사용자 네트워크 정보를 결합한 결과값

# 결론

- 협업필터링과 사회적 네트워크 분석을 활용한 선호도 예측 방안 제시
  - 사용자-네트워크 척도 사용
  - 상품-네트워크 척도 사용
- 협업필터링에 사용자-네트워크 척도 활용한 결과 기존 협업필터링보다 더 나은 성능을 보임.
  - Degree centrality가 가장 좋은 성능을 보임.
  - Power centrality 제외

# 향후 추진계획 & 연구과제

- 사용자 네트워크에 대한 사회네트워크분석 척도의 활용 뿐만 아니라 상품 네트워크에 대한 사회네트워크분석 척도의 활용 방안에 대한 추가적인 연구
- 네트워크 척도간의 최적 가중치에 대한 분석
- 실험의 결과의 일반화를 위해서는 반복적인 실험이 추가적으로 필요 (20회 이상)
  - 통계적 검증

- 다른 Data set에 실험 적용
- 본 연구에서 사용한 사회적 네트워크 척도 외에 다른 척도 활용

Thank you.