# Developing a hybrid collaborative filtering recommendation system with opinion mining on purchase review

Brain neuro Language Processing Lab.

Youdong Yun

# NOTICE

01 Introduction

02 Related research

03 Proposed recommendation system

04 Experimental evaluation

05 Result and discussion

06 Conclusion and Future work

# 01 Introduction

- 통신 기술의 발달로 인한 스마트 기기 보급 증가
- 인터넷의 생활화 및 결제시스템과 물류시스템의 발달
- 전자상거래의 발전으로 인해 원활한 소비 환경의 조성

소비자들의 온라인 쇼핑 이용 증가!

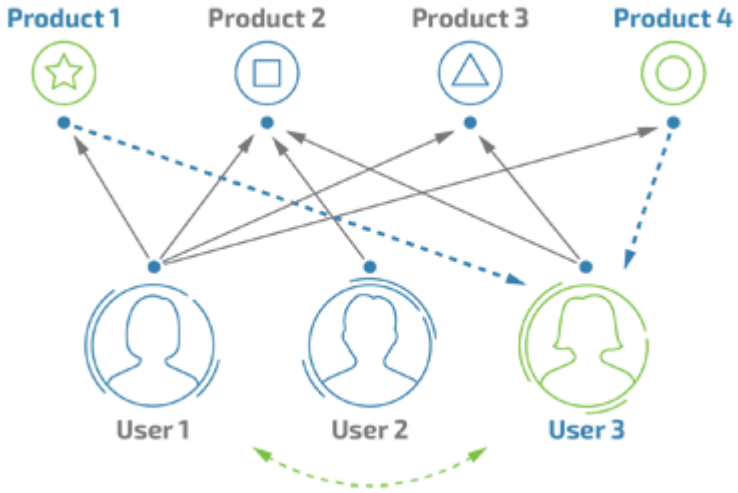
# 01 Introduction

제품들이 너무 많은데....
내가 원하는 것만 찾아줄 수 없나?

소비자들은 온라인 환경의 수많은 제품 사이에서 개인의 취향 및 요구에 가장 적합한 제품을 제공받는 서비스를 원하게 되었고, 소비자의 요구에 따라 추천 시스템에 대한 관심이 높아지고 있다.

# 01 Introduction

Collaborative Filtering은 소비자 선호도 이력을 기반으로 탐색된 유사 성향 이웃 소비자로부터 목표 소비자에게 경험하지 못 했던 제품에 대한 선호도를 예측하는 알고리즘이다.




(Collaborative Filtering : CF)

# 01 Introduction

Collaborative Filtering은 성공적인 개인화 추천 알고리즘으로서 다양한 분야에서 활용되고 있으나, Collaborative Filtering에 대한 몇 가지 문제점은 지속적으로 화두가 된다.

- 알고리즘의 우수성에 비해 높지 않은 알고리즘의 예측력
- 기존의 협업 필터링에서 사용하는 정량적인 데이터의 선호도 대표성 부족

# 01 Introduction

이러한 문제점을 해결하고자 본 연구에서는 Collaborative Filtering 예측력 향상을 위해 정량적인 데이터 뿐만 아니라 정성적인 데이터를 활용한 추천 시스템을 개발하였다.

Quantitative Data (Evaluation rating)

Qualitative Data (Purchase review)

# 02 Related Research

- Opinion Mining

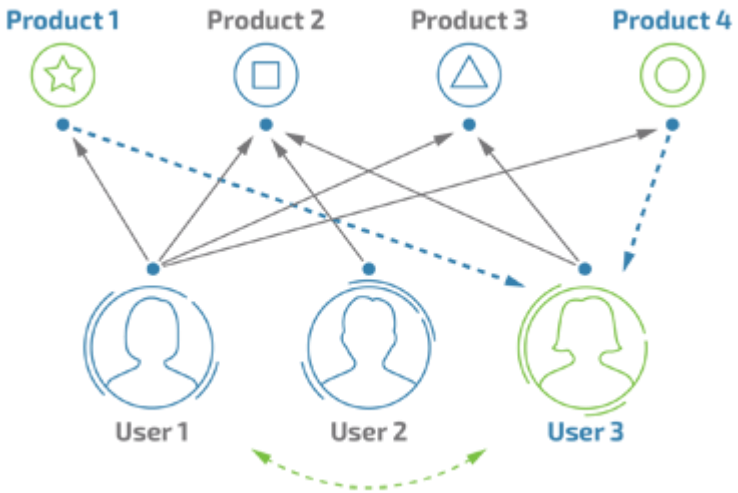
감성 분석 (Sentiment Analysis)이라고도 하며, 정형 및 비정형 텍스트 데이터의 주관적 정보를 식별하기 위해 NLP, Text Analysis, Computational Linguistics를 활용하는 것을 말한다. 일반적으로 Topic 또는 문서의 전반적인 문맥 극성 (Polarity)과 관련하여 작성자의 태도를 판단한다.

- Positive
- Neutral
- Negative

# 02 Related Research

- Collaborative Filtering

협업 필터링 (Collaborative Filtering) 접근 방법은 사용자의 제품에 대한 구매 및 선택 행위와 해당 제품에 대해 부여한 평가 점수 등의 과거 행동과 다른 사람이 내린 유사한 결정을 기반으로 모델을 생성하고, 다음 사용자가 관심을 가질 것 같은 제품을 예측하기 위한 알고리즘이다.




- Product
- Neighbor

# 02 Related Research

- Combination of opinion mining and Collaborative Filtering

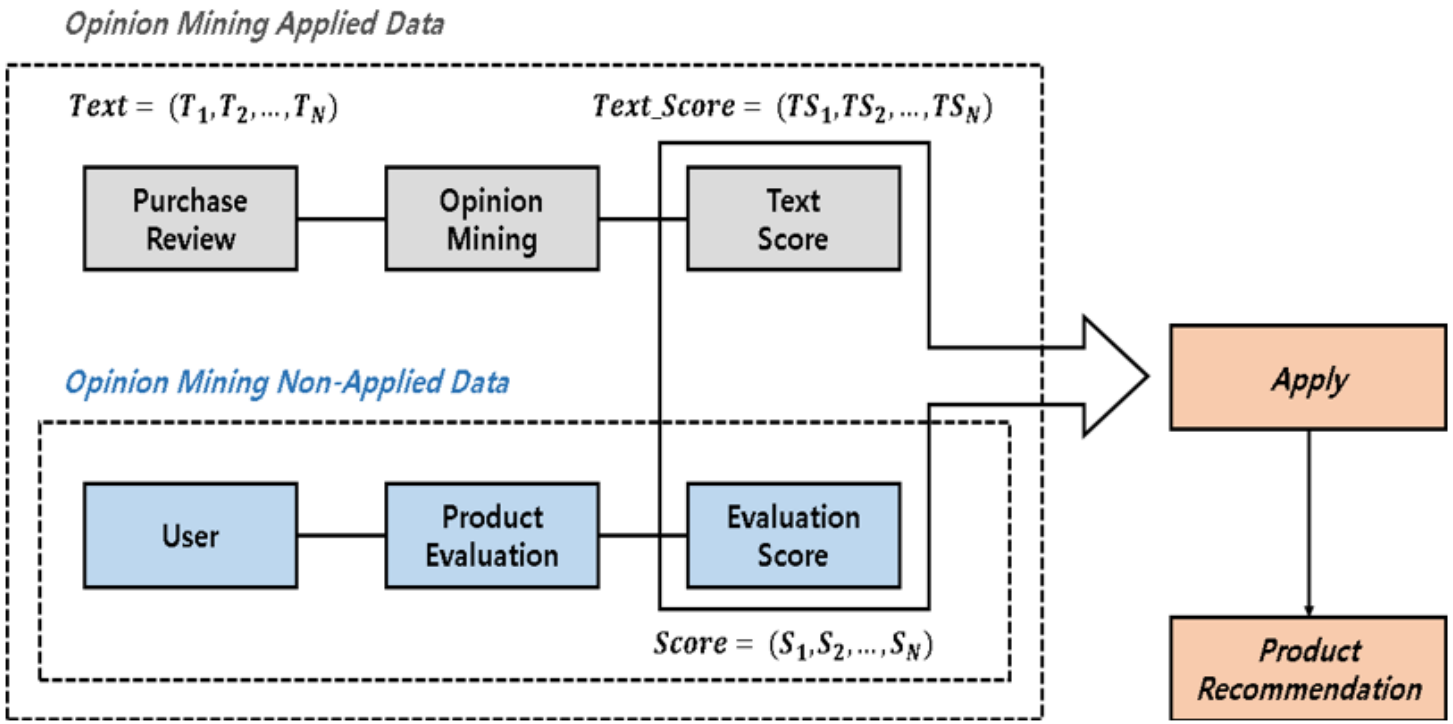
Collaborative Filtering과 Opinion Mining을 결합하려는 시도는 매우 적으나, 기존의 연구에서 소비자의 리뷰를 활용한 Collaborative Filtering의 성능이 더 우수한 결과를 보였으며, Opinion Mining과 결합 된 구매 리뷰는 보다 정확한 제품 추천 목록을 생성할 수 있다.

| | PORE* with MAE | Item-Based CF with MAE | User-Based CF with MAE |
|---|---|---|---|
| Mean Absolute Error (평균 절대 오차) | 1.19 | 3.10 | 2.96 |

*PORE = Preference and Opinion-based REcommendation

# 03 Proposed Recommendation System

- **Proposed recommendation system process**



# 03 Proposed Recommendation System

- Proposed recommendation system model
  - $OMNAD = Opinion\ Mining\ Non - Applied\ Data$
  - $OMAD = Opinion\ Mining\ Applied\ Data$
    - $AS_OMAD = S + TS$
    - $CS_OMAD = \lceil (S + TS)/2 \rceil$
    - $NS_OMAD = \left(\frac{S - min(S)}{max(S) - min(S)}\right) \cdot 100, \left(\frac{TS - min(TS)}{max(TS) - min(TS)}\right) \cdot 100$
    - $NAS_OMAD = \left(\frac{S - min(S)}{max(S) - min(S)}\right) \cdot 100 + \left(\frac{TS - min(TS)}{max(TS) - min(TS)}\right) \cdot 100$

# 04 Experimental Evaluation

- **Data set**

**"Amazon find food reviews data (October, 2012)" 120K개의 데이터를 활용하였다.**

| ID | Product ID | User ID | Score | Purchase review |
|---|---|---|---|---|
| 1 | B006K2ZZ7K | A1UQRSC LF8GW1T | 5 | Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal. |
| 2 | B0064KO0BU | A34D61R QILOKIJ | 1 | Besides being smaller than runts, they look the same and have the same consistency. Unfortunately, they taste nothing like banana runts … nor do they even taste good. Yucky stuff. Trying to return with vendor. |
| 3 | B004K2IHUO | A3D1TXE 98KRKYO | 4 | Arrived slightly thawed. My parents wouldn't accept it. However, the company was very helpful and issued a full refund. |
| 4 | B000UUIYHQ | AWIR8TQZ GSVV9 | 3 | I was a little disappointed with the taste (or lack thereof) of these cookies. They are low in sugar and taste like. If you want a healthy alternative and don't mind sacrificing some taste, then this isn't a bad option. |
| 5 | B002Y2QSLC | A2Y00L1N 2C8F66 | 2 | This coffee is a disappointment. The taste goes nowhere … no full body and aroma. So much for the private reserve alluring trap, really! |

# 04 Experimental Evaluation

- Opinion mining

제품에 대해 소비자가 작성한 텍스트 리뷰가 갖는 긍정, 부정, 중립적인 의견을 판별하는 분석이며, 미리 구축된 감성 사전과 작성된 소비자의 리뷰를 매칭하여 텍스트 내 극성을 분석하게 된다.

| Type of words | Word example |
|---|---|
| Positive words | Accommodative, adaptive, advanced, blossom, confidence, congratulate, encouragingly, endear, endearing, endorse, endorsed, energetic, energise, pleasant, pleasantly, pleased, pleases, pleasing, pleasingly, pleasurable, pleasurably, pleasure, plentiful, slick, smart, smarter, smartest, trivially, trophy, … |
| Negative words | Abominably, abominate, abomination, abort, aborted, aborts, abrade, abrasive, abrupt, abruptly, blind, blunder, bondage, brashness, detestably, detested, detesting, detracted, disaster, disbelieve, grievous, grievously, grim, grimace, grind, gripe, infiltrator, infiltrators, infirm, inordinate, revolting, unwell, … |

# 04 Experimental Evaluation

- **User-Based Collaborative Filtering**

제품을 구매하고자 하는 소비자의 구매 내역 프로파일에서 소비자를 기준으로 다른 소비자들과 유사도를 계산하여 가장 가까운 이웃을 찾아내고 이웃들의 성향에 따라 제품을 추천해준다.

| Actual \ Predicted | Negative | Positive |
|---|---|---|
| Negative | a | b |
| Positive | c | d |

(1) $precision = \left(\frac{a+b}{a+b+c+d}\right),\ recall = \left(\frac{d}{c+d}\right)$

(2) $TPR(True\ Positive\ Recommendation) = \frac{d}{b+d}$

(3) $FPR(False\ Positive\ Recommendation) = \frac{b}{b+d}$

(4) $F-measure = \frac{2 \times precision \times recall}{precision + recall}$

# 04 Experimental Evaluation

- **User-Based Collaborative Filtering**

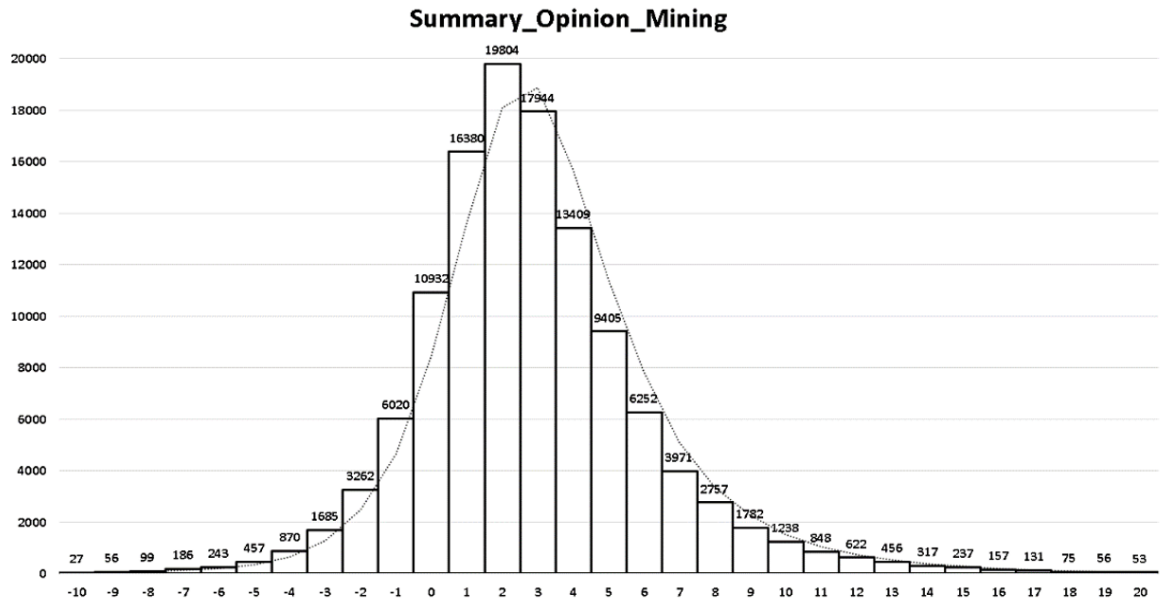
  - ***Precision*** *= 추천 시스템이 추천한 제품 중 실제 소비자가 좋아하는 제품의 비율*
  - ***Recall*** *= 소비자가 실제로 좋아하는 제품 중 몇 개의 제품을 정확하게 추천 했는지의 비율*
  - ***TPR*** *= 추천 시스템이 추천한 제품 중 사용자가 선택한 제품의 비율*
  - ***FPR*** *= 추천 시스템이 추천한 제품 중 소비자가 선택하지 않은 제품의 비율*

**본 연구에서는 80%의 Training 데이터, 20%의 Test 데이터로 분석을 수행하였으며, 최소한 3번 이상 구매된 제품으로 분석을 진행하였고, 사용자 유사도 측정에 Cosine Similarity를 사용했다.**

# 05 Result and Discussion

- **Opinion Mining**

일부 의견이 과도하게 낮거나 높게 나타나 이러한 269개의 데이터를 노이즈(Noise) 데이터로 분류하여 제거하여 총 -10에서 20점 사이의 119,731개의 데이터로 분석을 진행하였다.



# 05 Result and Discussion

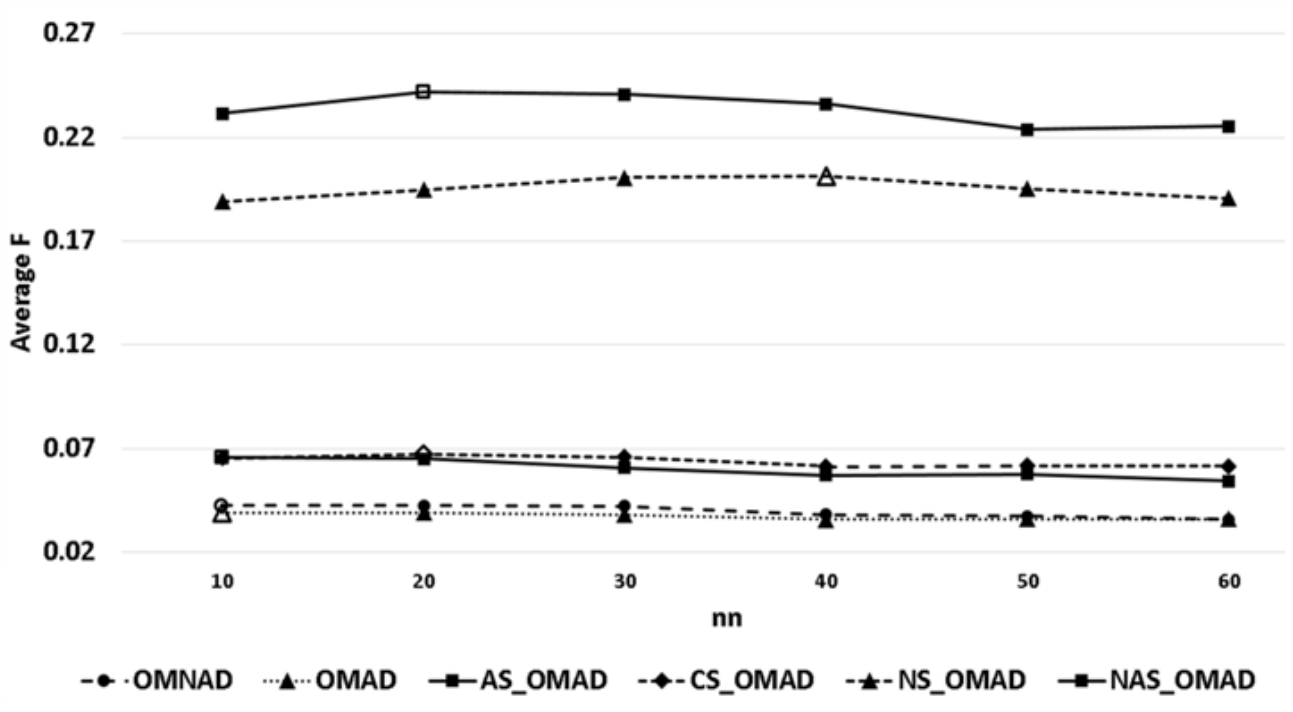
- Opinion Mining

| ID | Product ID | User ID | Score | Text_Score | Purchase Review |
|---|---|---|---|---|---|
| 1 | B001SATU8E | A1CR1DS 15Z7XO1 | 4 | 5 | Makes a tasty, super easy meal, fast. BUT high in calories. The instructions say to saute the veggies first but I recommend cooking the chicken first. The chicken takes longer to cook and the raw chicken on top of veggies just. |
| 2 | B001EPPI84 | A27TZ4W BU7N0YF | 1 | −1 | No tea flavour at all. Just whole brunch of artificial flavours. It is not returnable. I wasted 20+ bucks. |
| 3 | B004K2IHUO | A1ZKFQL HFZAEH9 | 3 | 2 | Not what I was expecting in terms of the company's reputation for excellent home delivery products. |
| 4 | B0093NIWVO | A1JT114 SOITFFO | 5 | 2 | Fresh, a great way to get a little chocolate in my life without a million calories. They taste just like chocolate pudding. |
| 5 | B001EPPI84 | A3Q0IDQ 03S0158 | 2 | 0 | It is okay. I would not go out of my way to buy it again. |

# 05 Result and Discussion

- User-Based Collaborative Filtering

F-measure를 통해 총 모델에서 가장 적합한 이웃의 수를 찾기 위해 분석을 실시한 결과이다.

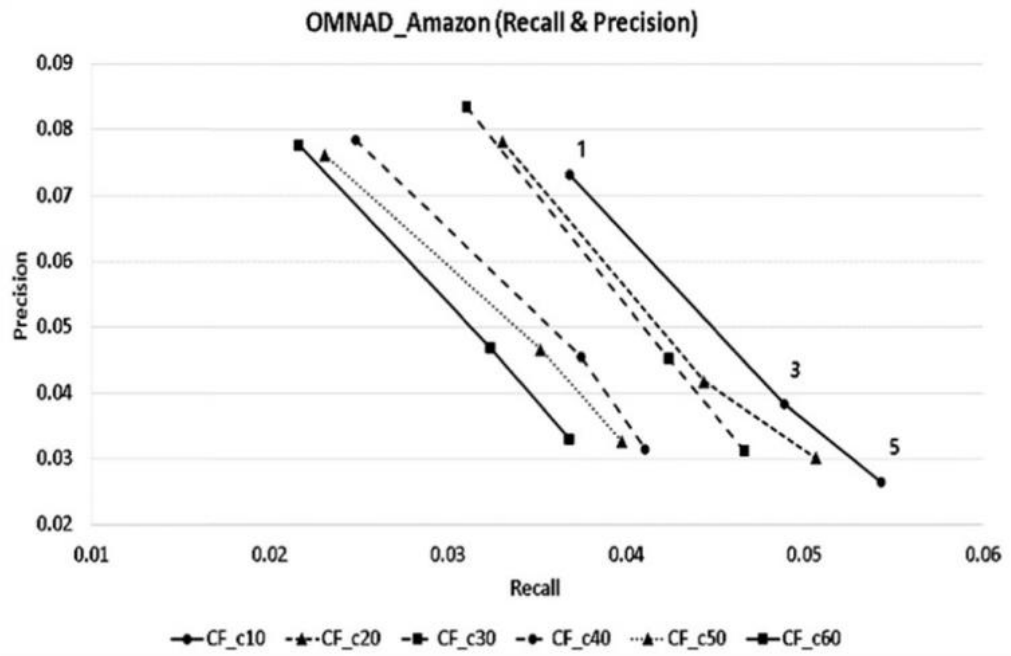


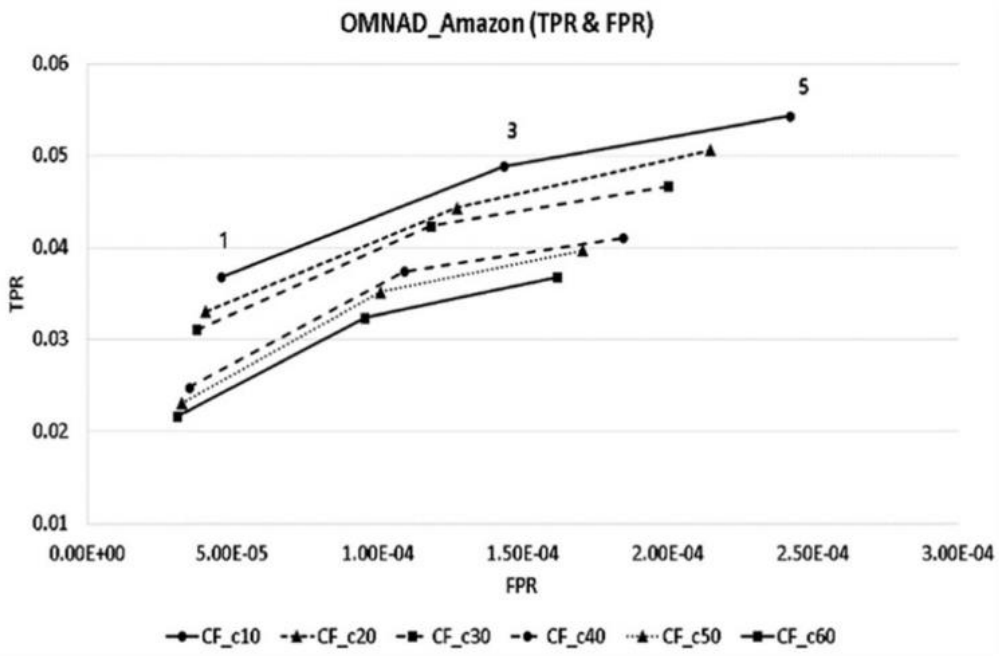

- *OMNAD = 10명 (F = 0.0425)*
- *OMAD = 10명 (F = 0.039174)*
- *AS_OMAD = 10명 (F = 0.0655)*
- *CS_OMAD = 20명 (F = 0.0672)*
- *NS_OMAD = 40명 (F = 0.2007)*
- *NAS_OMAD = 20명 (F = 0.2417)*

# 05 Result and Discussion

- User-Based Collaborative Filtering (OMNAD)

10명의 이웃 수로 설정했을 때 가장 우수한 결과를 보였다.



# 05 Result and Discussion

- User-Based Collaborative Filtering (OMNAD)

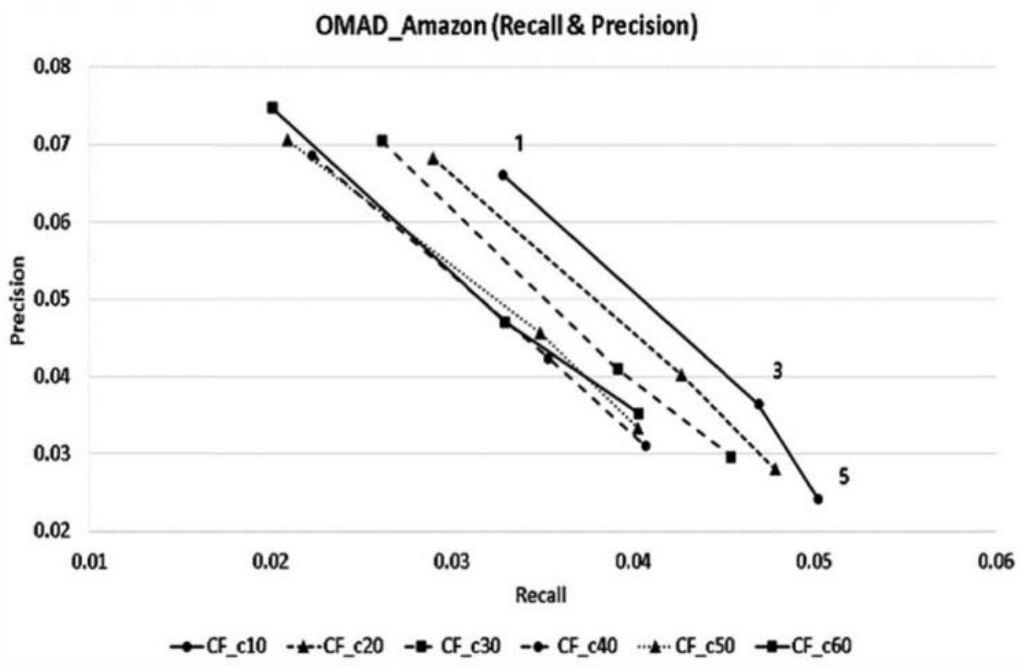
10명의 이웃 수로 설정했을 때 가장 우수한 결과를 보였다.

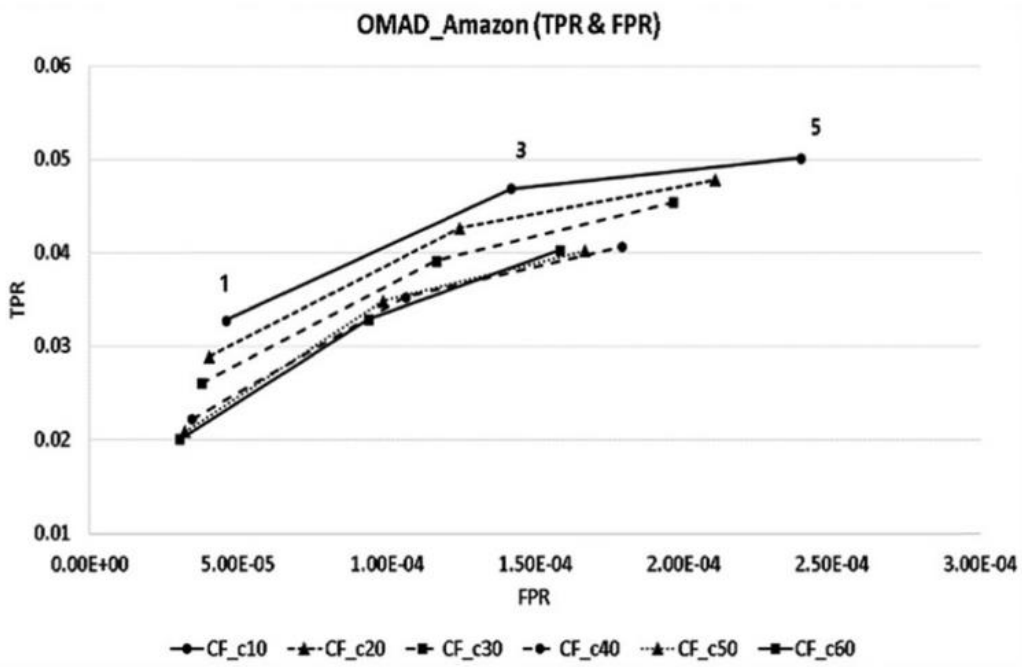
| Number of item | 1 | 2 | 3 |
|---|---|---|---|
| **Precision** | 0.0732 | 0.0385 | 0.0265 |
| **Recall** | 0.0368 | 0.0488 | 0.0543 |
| **TPR** | 0.0368 | 0.0488 | 0.0543 |
| **FPR** | 0.0000460 | 0.0001434 | 0.0002420 |

# 05 Result and Discussion

- User-Based Collaborative Filtering (OMAD)

10명의 이웃 수로 설정했을 때 가장 우수한 결과를 보였다.



# 05 Result and Discussion

- User-Based Collaborative Filtering (OMAD)

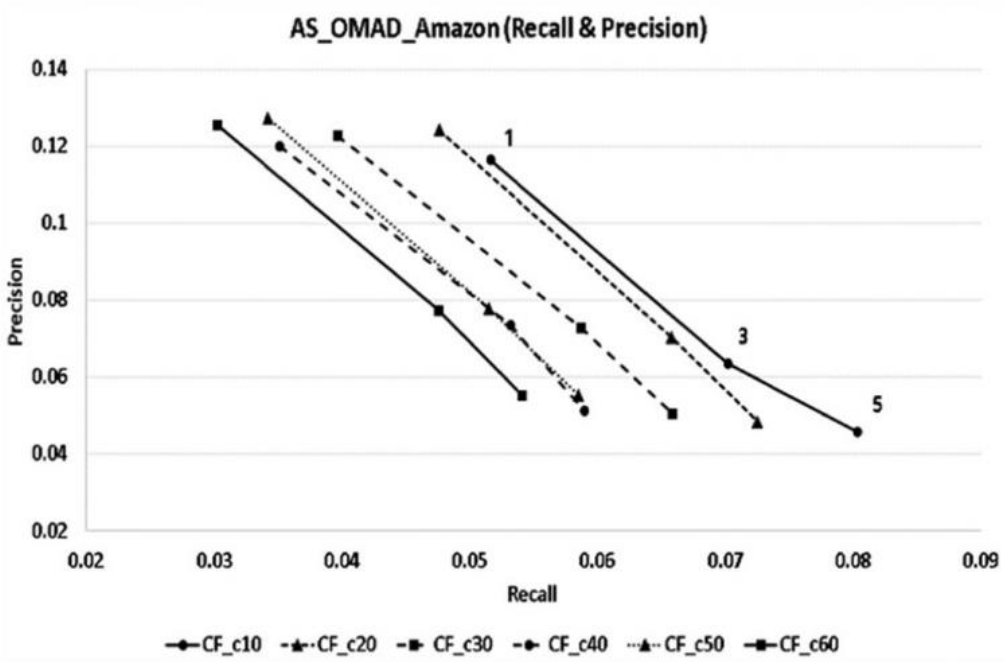
10명의 이웃 수로 설정했을 때 가장 우수한 결과를 보였다.

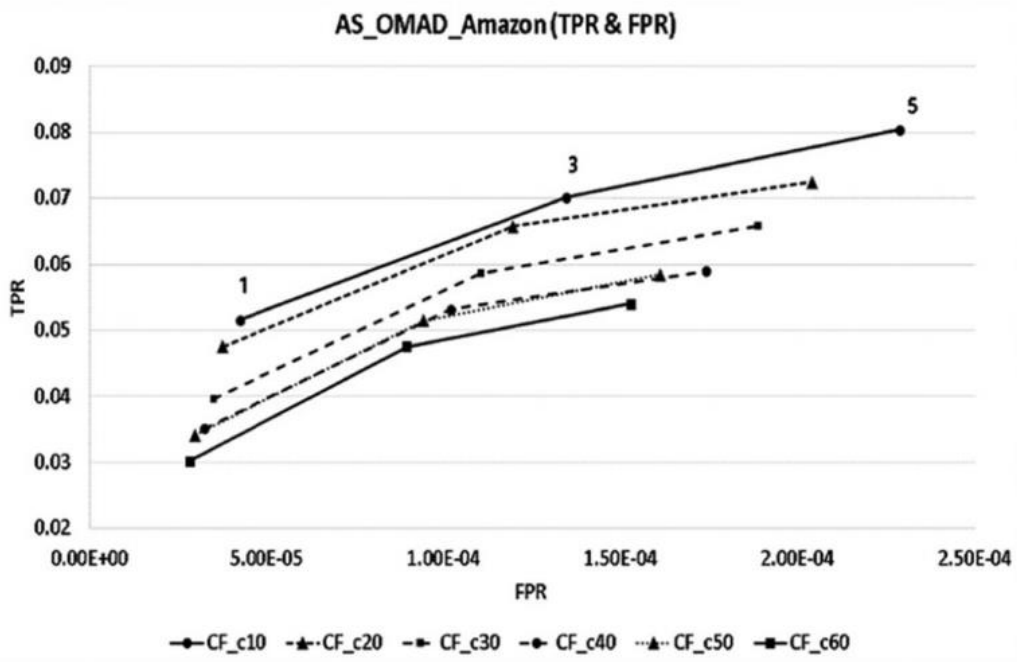
| Number of item | 1 | 2 | 3 |
|---|---|---|---|
| **Precision** | 0.0662 | 0.0365 | 0.0242 |
| **Recall** | 0.0328 | 0.0469 | 0.0502 |
| **TPR** | 0.0328 | 0.0469 | 0.0502 |
| **FPR** | 0.0000458 | 0.0001419 | 0.0002395 |

# 05 Result and Discussion

- User-Based Collaborative Filtering (AS_OMAD)

10명의 이웃 수로 설정했을 때 가장 우수한 결과를 보였다.



# 05 Result and Discussion

- User-Based Collaborative Filtering (AS_OMAD)

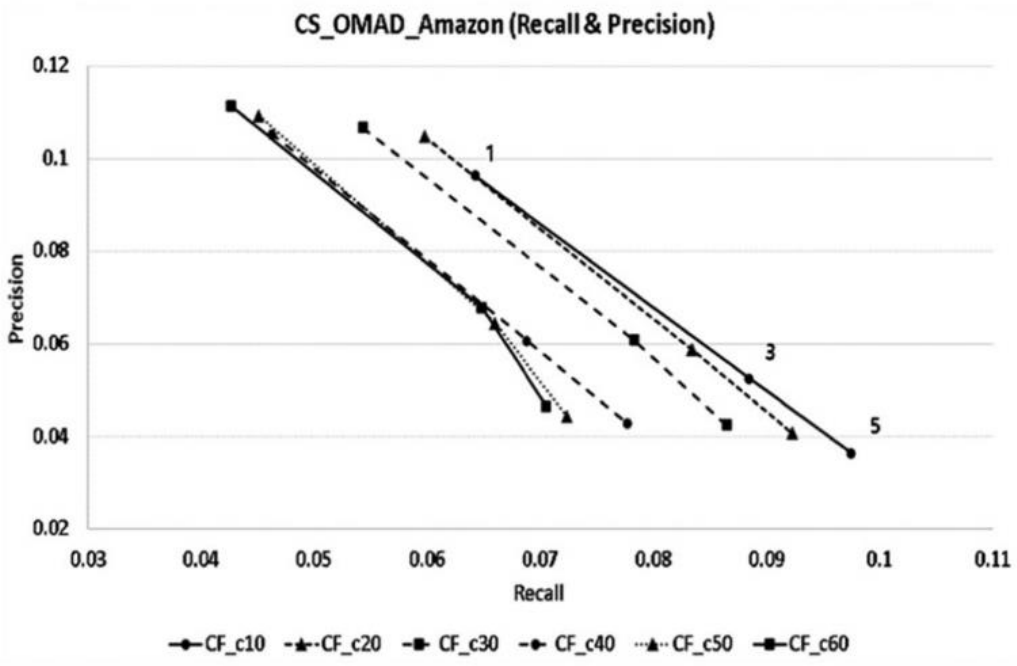
10명의 이웃 수로 설정했을 때 가장 우수한 결과를 보였다.

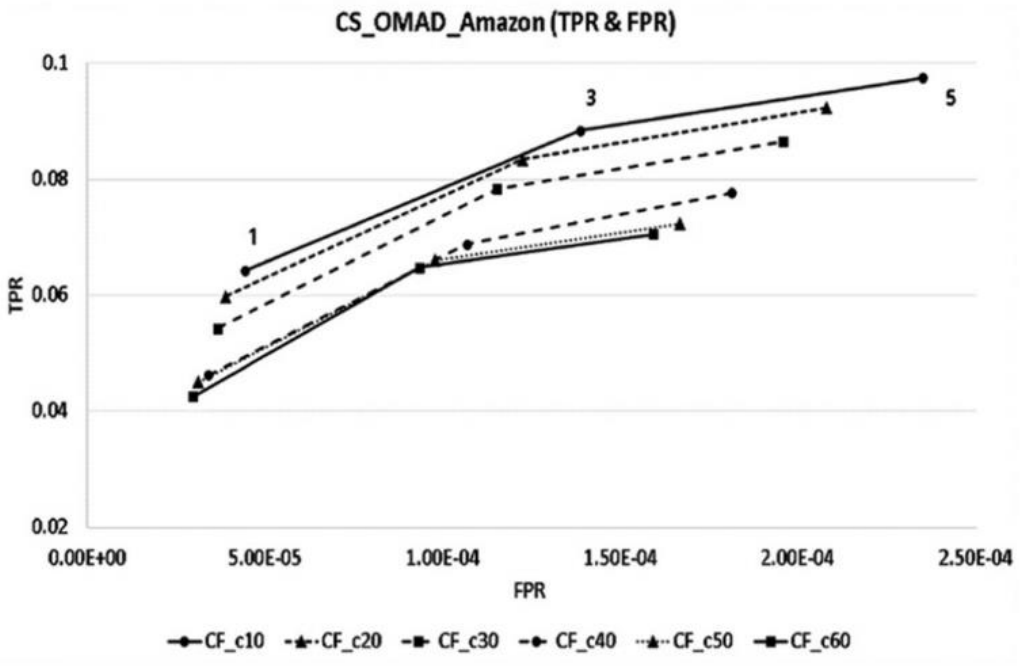
| Number of item | 1 | 2 | 3 |
|---|---|---|---|
| **Precision** | 0.1165 | 0.0635 | 0.0459 |
| **Recall** | 0.0516 | 0.0702 | 0.0803 |
| **TPR** | 0.0516 | 0.0702 | 0.0803 |
| **FPR** | 0.0000423 | 0.0001347 | 0.0002288 |

# 05 Result and Discussion

- User-Based Collaborative Filtering (CS_OMAD)

20명의 이웃 수로 설정했을 때 가장 우수한 결과를 보였다.



# 05 Result and Discussion

- User-Based Collaborative Filtering (CS_OMAD)

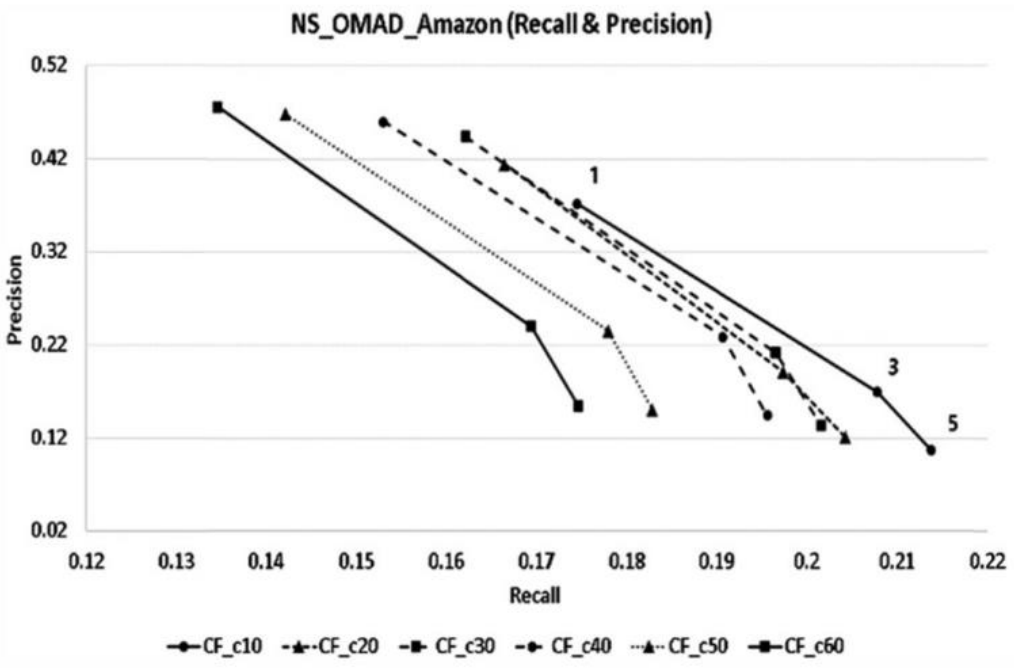
20명의 이웃 수로 설정했을 때 가장 우수한 결과를 보였다.

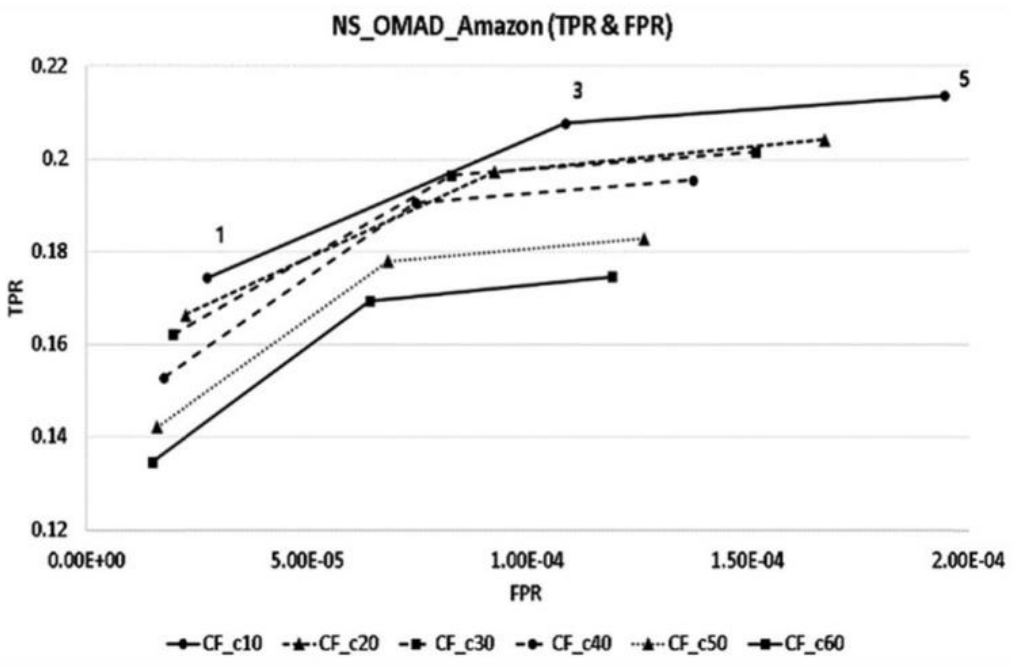
| Number of item | 1 | 2 | 3 |
|---|---|---|---|
| **Precision** | 0.1048 | 0.0588 | 0.0408 |
| **Recall** | 0.0598 | 0.0834 | 0.0923 |
| **TPR** | 0.0598 | 0.0834 | 0.0923 |
| **FPR** | 0.000388 | 0.0001225 | 0.0002081 |

# 05 Result and Discussion

- User-Based Collaborative Filtering (NS_OMAD)

40명의 이웃 수로 설정했을 때 가장 우수한 결과를 보였다.



# 05 Result and Discussion

- User-Based Collaborative Filtering (NS_OMAD)

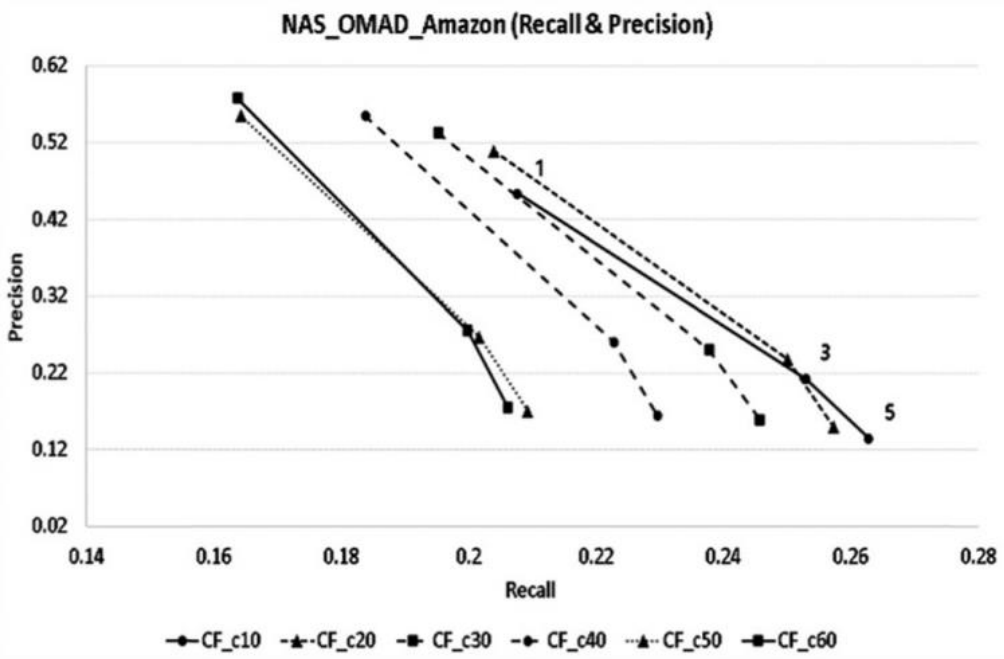
40명의 이웃 수로 설정했을 때 가장 우수한 결과를 보였다.

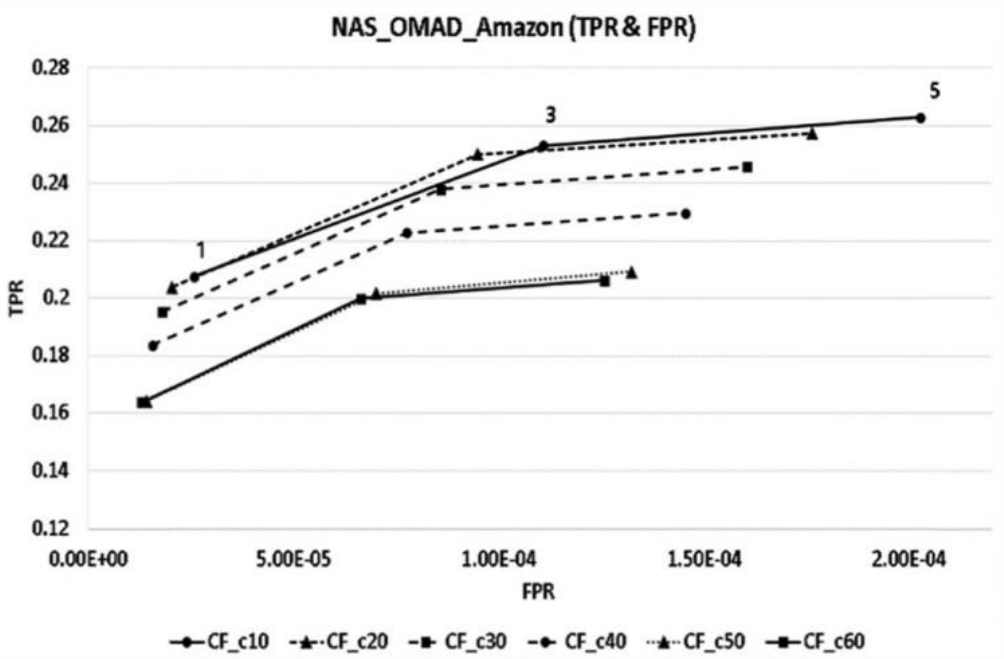
| Number of item | 1 | 2 | 3 |
|---|---|---|---|
| **Precision** | 0.4601 | 0.2289 | 0.1451 |
| **Recall** | 0.1529 | 0.1906 | 0.1955 |
| **TPR** | 0.1529 | 0.1906 | 0.1955 |
| **FPR** | 0.0000174 | 0.0000744 | 0.0001375 |

# 05 Result and Discussion

- User-Based Collaborative Filtering (NAS_OMAD)

20명의 이웃 수로 설정했을 때 가장 우수한 결과를 보였다.



# 05 Result and Discussion

- User-Based Collaborative Filtering (NAS_OMAD)

20명의 이웃 수로 설정했을 때 가장 우수한 결과를 보였다.

| Number of item | 1 | 2 | 3 |
|---|---|---|---|
| **Precision** | 0.5093 | 0.2383 | 0.1503 |
| **Recall** | 0.2039 | 0.2500 | 0.2572 |
| **TPR** | 0.2039 | 0.2500 | 0.2572 |
| **FPR** | 0.0000203 | 0.0000948 | 0.0001763 |

# 05 Result and Discussion

- User-Based Collaborative Filtering

| Recommended number | | One item | Three items | Five items |
|---|---|---|---|---|
| OMNAD | Recall | 0.0368 | 0.0488 | 0.0543 |
| | Precision | 0.0732 | 0.0385 | 0.0265 |
| | TPR | 0.0368 | 0.0488 | 0.0543 |
| | FPR | 0.0000460 | 0.0001434 | 0.0002420 |
| OMAD | Recall | 0.0328 | 0.0469 | 0.0502 |
| | Precision | 0.0662 | 0.0365 | 0.0242 |
| | TPR | 0.0328 | 0.0469 | 0.0502 |
| | FPR | 0.0000458 | 0.0001419 | 0.0002395 |
| AS_OMAD | Recall | 0.0516 | 0.0702 | 0.0803 |
| | Precision | 0.1165 | 0.0635 | 0.0459 |
| | TPR | 0.0516 | 0.0702 | 0.0803 |
| | FPR | 0.0000423 | 0.0001347 | 0.0002288 |
| CS_OMAD | Recall | 0.0598 | 0.0834 | 0.0923 |
| | Precision | 0.1048 | 0.0588 | 0.0408 |
| | TPR | 0.0598 | 0.0834 | 0.0923 |
| | FPR | 0.0000388 | 0.0001225 | 0.0002081 |
| NS_OMAD | Recall | 0.1529 | 0.1906 | 0.1955 |
| | Precision | 0.4601 | 0.2289 | 0.1451 |
| | TPR | 0.1529 | 0.1906 | 0.1955 |
| | FPR | 0.0000174 | 0.0000744 | 0.0001375 |
| NAS_OMAD | Recall | 0.2039 | 0.2500 | 0.2572 |
| | Precision | 0.5093 | 0.2383 | 0.1503 |
| | TPR | 0.2039 | 0.2500 | 0.2572 |
| | FPR | 0.0000203 | 0.0000948 | 0.0001763 |

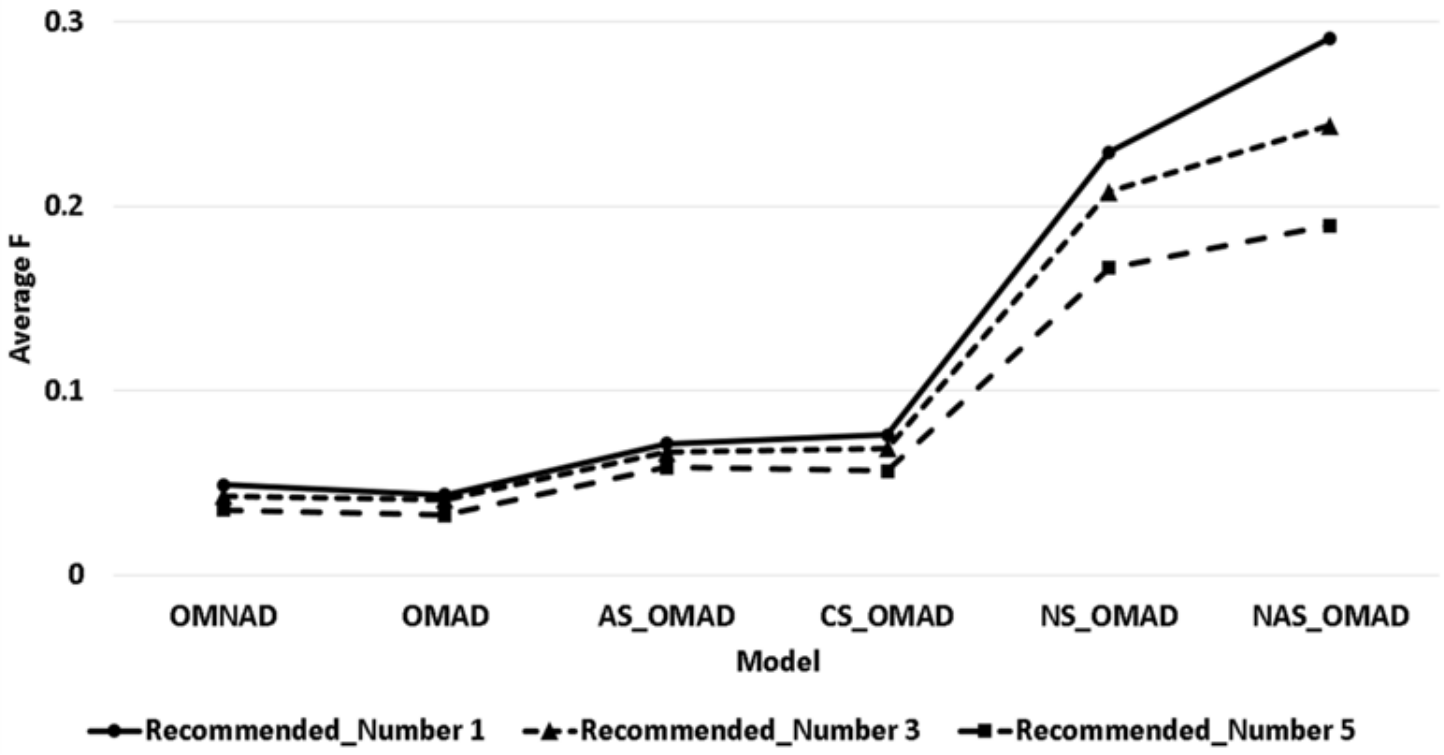
# 05 Result and Discussion

- User-Based Collaborative Filtering

| Recommended number | | One item | Three items | Five items |
|---|---|---|---|---|
| OMNAD | Average *F* | 0.0492 | 0.0430 | 0.0356 |
| OMAD | | 0.0438 | 0.0410 | 0.0326 |
| AS_OMAD | | 0.0715 | 0.0667 | 0.0584 |
| CS_OMAD | Average *F* | 0.0761 | 0.0689 | 0.0566 |
| NS_OMAD | | 0.2296 | 0.2080 | 0.1666 |
| NAS_OMAD | | 0.2912 | 0.2440 | 0.1897 |

# 05 Result and Discussion

- User-Based Collaborative Filtering



# 06 Conclusion and Future Work

- **Conclusion**

  - *소비자에게 가장 적합한 제품을 추천하기 위해 사용자가 직접 제품을 경험해보고 작성한 구매후기를 사용한다면, 정량적 데이터만 사용하는 기존의 연구에 비해 더욱 객관적으로 사용자의 선호도를 고려할 수 있다.*

  - *CF 추천 시스템을 통해 소비자들이 기존에 경험하지 못 했던 제품에 대한 선호도를 예측할 때, Opinion Mining을 통해 분석한 소비자들의 구매후기를 활용함으로써 CF 추천 시스템의 성능이 향상됨을 증명했다.*

# 06 Conclusion and Future Work

- **Future Work**

  - *“Not Good”과 같이 두 단어 이상으로 구성된 부정을 의미하는 구매후기 문장에서 “Not”과 “Good”을 따로 인식하여 올바른 구매후기에 대한 해석이 어려울 수 있다.*

  - *실제로 제품을 구매하는 모든 소비자가 구매후기를 작성하지 않기 때문에 본 연구에서 제안하는 시스템의 적용에 대한 어려움이 문제로 제기되었다.*

# Thank you!

Brain neuro Language Processing Lab.

Youdong Yun