Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity

M. Firoz Kabir
Technical University of Braunschweig
Outline

• Introduction
• Research Objectives
• Data & Metrics
  • Dataset
  • Recommendation takers
  • Content Diversity
  • Effect of RS
• Results
• Discussion
• Bibliography
Introduction

- **Recommender System**: Recommender systems apply knowledge discovery techniques to the problem of making personalized recommendations for information, products or services during a live interaction.
- **Filter Bubble**: Online personalization effect on people diversity of viewpoint or content.
- **Content Diversity**: How one content is different from another content based on tag genome data.
Research Objectives

RQ1: Do recommender systems expose users to narrow content over time?

RQ2: How does the experience of users who take recommendations differ from that of users who do not regularly take recommendations?
Dataset

- MovieLens is a movie recommender system
- 217267 unique users, 20M ratings and 20K movies (2013)
- MovieLens Longitudinal data of user rating data
- Logs data of rating and recommendation page
- It uses well known RS with item-item collaborative filtering (CF)
- Tag-genome data to compute content diversity
Collaborative filtering

- Collaborative filtering works by building a database of preferences for items by users.
- MovieLens uses an item based collaborative filtering algorithm for recommendation that is robust and highly scalable.
- The goal of a collaborative filtering algorithm is to suggest new items or to predict the utility of a certain item for a particular user based on personalized data.
- Scalability and high quality recommendation are main challenges with CF algorithms.
Collaborative filtering (2/2)

• **Two categories**: Memory-based (user-based) and Model-based (item-based) CF algorithms

• **Memory based**: Memory-based algorithms utilize the entire user-item database to generate a prediction

• **Model based**: Model-based collaborative filtering algorithms provide item recommendation by first developing a model of user ratings
Identify Recommendation takers

• Removed first 15 ratings for each user
• Each rating block consist of 10 consecutive ratings
• Removed last rating block incase of insufficient rating
• User is only counted if s/he has 3 or more ratings blocks
• 1405 users, 3-203 ratings blocks, 173K ratings, 10K movies
• Top Picks For You accessed 150K times

Figure 2: Rating Block Illustration
Identify Consumed Recommendation

• We have to identify which movie in each rating block were recommended explicitly to user
• Based on recommended movie in rating block and rating we can measure user experience
• Movie in top picks between 3 hours and 3 months before rating
Ignoring Vs Following Group

• We first look at whether the user took at least 1 recommendation in one of his rating blocks.
• Then we compute the percentage of that user rating blocks in which the user took at least one recommendation.
• Taker in 50% rating blocks classified as following group
Measuring Content Diversity

• We used tag-genome information space to describe movie
• Tag-genome has relevance score based on description
• Relavance score 1 to 5 for tag-genome
• We compute Euclidean distance between two movie vectors to examine content similarity or diversity.

\[ d(m_i, m_j) = \sqrt{\sum_{k=1}^{m} [rel(t_k, m_i) - rel(t_k, m_j)]^2} \]
Measuring the effect

- They compute movie distance distribution for each group to find out content diversity distribution.
- Compute content diversity for top 15 movies per user.
- They measure given ratings to understand user experience.
- Investigate the effect by measuring the shift of both group.
## Results (1/5)

<table>
<thead>
<tr>
<th></th>
<th>At the beginning</th>
<th>At the end</th>
<th>Within-group p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All users</strong></td>
<td>25.02</td>
<td>24.67</td>
<td>2.43e-06</td>
</tr>
<tr>
<td><strong>Following Group</strong></td>
<td>25.22</td>
<td>24.80</td>
<td>0.014</td>
</tr>
<tr>
<td><strong>Ignoring Group</strong></td>
<td>24.74</td>
<td>24.51</td>
<td>0.087</td>
</tr>
<tr>
<td><strong>Between-group p-value</strong></td>
<td>0.0037</td>
<td>0.0406</td>
<td></td>
</tr>
</tbody>
</table>

- For all users content distance becomes similar over time.
- Following group movies are also becoming similar over time.
- Ignoring group doesn’t have significant drop as they don’t take recommendation and movielens learn their preferences.
- Both group content diversity becomes smaller over the time.
Results (2/5)

Average and maximum content diversity of both group

<table>
<thead>
<tr>
<th>Rating Block</th>
<th>The First</th>
<th>The Last</th>
<th>Within-group p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All users</strong></td>
<td>26.60</td>
<td>26.01</td>
<td>1.542e-12</td>
</tr>
<tr>
<td><strong>Following Group</strong></td>
<td>26.67</td>
<td>26.30</td>
<td>0.01007</td>
</tr>
<tr>
<td><strong>Ignoring Group</strong></td>
<td>26.59</td>
<td>25.86</td>
<td>8.236e-07</td>
</tr>
<tr>
<td>Between-group p-value</td>
<td>0.6162</td>
<td>0.006468</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rating Block</th>
<th>The First</th>
<th>The Last</th>
<th>Within-group p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All users</strong></td>
<td>34.56</td>
<td>34.00</td>
<td>8.903e-07</td>
</tr>
<tr>
<td><strong>Following Group</strong></td>
<td>34.73</td>
<td>34.36</td>
<td>0.127</td>
</tr>
<tr>
<td><strong>Ignoring Group</strong></td>
<td>34.45</td>
<td>33.73</td>
<td>0.000</td>
</tr>
<tr>
<td>Between-group p-value</td>
<td>0.237</td>
<td>0.008</td>
<td></td>
</tr>
</tbody>
</table>
Results (3/5)

- Following group user experience

<table>
<thead>
<tr>
<th>Rating Block</th>
<th>All Users</th>
<th>The First</th>
<th>The last</th>
<th>Within-group p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3.69</td>
<td>3.57</td>
<td>2.2e-16</td>
</tr>
<tr>
<td>Following Group</td>
<td>3.69</td>
<td>3.68</td>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td>Ignoring Group</td>
<td>3.74</td>
<td>3.55</td>
<td></td>
<td>3.128e-11</td>
</tr>
<tr>
<td>Between-group p-value</td>
<td>0.2129</td>
<td>0.001719</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The Percentage of rated movies

<table>
<thead>
<tr>
<th>Rating Block</th>
<th>Rating Block</th>
<th>0.5 - 1 stars</th>
<th>1.5 - 2 stars</th>
<th>2.5 - 3 stars</th>
<th>3.5 - 4 stars</th>
<th>4.5 - 5 stars</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Users</td>
<td>The First</td>
<td>2.7%</td>
<td>5.3%</td>
<td>17.8%</td>
<td>46.5%</td>
<td>27.7%</td>
</tr>
<tr>
<td></td>
<td>The Last</td>
<td>2.8%</td>
<td>6.3%</td>
<td>22%</td>
<td>46.4%</td>
<td>22.5%</td>
</tr>
<tr>
<td>Following Group</td>
<td>The First</td>
<td>2.2%</td>
<td>6.0%</td>
<td>17.8%</td>
<td>46.2%</td>
<td>27.8%</td>
</tr>
<tr>
<td></td>
<td>The Last</td>
<td>1.8%</td>
<td>5.1%</td>
<td>19.0%</td>
<td>49.2%</td>
<td>24.9%</td>
</tr>
<tr>
<td>Ignoring Group</td>
<td>The First</td>
<td>2.4%</td>
<td>4.6%</td>
<td>18.0%</td>
<td>45.3%</td>
<td>29.7%</td>
</tr>
<tr>
<td></td>
<td>The Last</td>
<td>3.6%</td>
<td>6.9%</td>
<td>21.5%</td>
<td>45.1%</td>
<td>22.9%</td>
</tr>
</tbody>
</table>
Discussion

• Recommender system narrower the content over time
• Taking recommendation lessened the filter bubble
• Following users narrow content diversity slowly than ignoring group
• We might have natural narrowing tendency specially for movie
• Collaborative filtering algorithm is recommended for designing RS
• Content based algorithm might push more narrow consumption
Bibliography


Thank You for your attention!
Semantics derived automatically from language corpora necessarily contain human biases

M. Firoz Kabir
Technical University of Braunschweig
Outline

• Introduction
• Research Objectives
• Methods
  • WEAT
  • WEFAT
  • IAT
• Dataset
• Results
  • Gender, racial, occupational etc.
• Discussion
• Bibliography
Introduction

• Machine Learning is able to capture the knowledge and computation discovered and transmitted by human and human culture
• Historic biases and prejudices are being refined in machines
• Neural Language Processing tools share same biases human states
• Bias result is expected even we used unbiased algorithm
• Machine prejudice are now coming to the fore as AI should always be applied transparently
Research Objectives

RQ1: Finding human biases from language corpora

RQ2: Introducing two new methods the Word Embedding Association Test (WEAT) and the word embedding factual association test (WEFAT)
Bias in Humans and Machines

• Harmful bias refer as prejudice that is identified by negative impact
• Prejudice needs deliberate action based on konwledge of society
• Human bias was documented by Implicit Association Test (IAT)
• IAT applied to semantic representation of words in AI, termed as word embeddings
• IAT applies to human subjects, embedding of interest derived from aggregate writing of humans
Dataset

• We use state of the art GloVe word embedding method
• We used pre-trained GloVe embeddings distributed by its author
• We pick four corpora which GloVe provides trained embeddings
• So dataset has 840 billion tokens (roughly words)
• Tokens are case sensitive in our corpus, 2.2 million different ones
Implicit Association Test (IAT)

- Implicit Association Test first introduced by Greenwald (1998)
- IAT usually used for pair of categories and two different concepts
- The IAT follows a reaction time paradigm, which means subjects are encouraged to work as quickly as possible
- If pairing is faster than it means task is more easy or pleasant
- The IAT has been used to describe wide range of prejudice and other phenomena
- We compared our result with original IAT findings in result section
Word Embedding Association Test (WEAT)

• Consider two sets of target and attribute words, have null hypothesis
• The permutation test measures the unlikelihood of null hypothesis
• IAT measures differential association for single pair but WEAT
  • The test statistic is
    \[ s(X,Y,A,B) = \sum_{x \in X} s(x,A,B) - \sum_{y \in Y} s(y,A,B) \]
    where
    \[ s(w,A,B) = \text{mean}_{a \in A} \cos(\bar{w},\bar{a}) - \text{mean}_{b \in B} \cos(\bar{w},\bar{b}) \]
    In other words, \( s(w,A,B) \) measures the association of the word \( w \) with the attribute, and \( s(X,Y,A,B) \) measures the differential association of the two sets of target words with the attribute.
Word Embedding Factual Association Test (WEFAT)

- We also wish to examine imparical information in word embedding
- Consider a set of target concept s, and real valued, factual property associated with each concept
- Consider a single set target words W, two set of attribute words A,B. There is a property Pw associated with each word w€W
- The null hypothesis is there is no association between s(w,A,B) and Pw: we test it with linear regression analysis

\[ s(w,A,B) = \frac{\text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)}{\text{std-dev}_{x \in A \cup B} \cos(w, x)} \]
Results (1/8)

Flowers and Insects

• Original Finding: Greenwald et al. Report that flower is pleasant and insects are unpleasant. Based on reaction of 32 participants IAT effect size 1.35 and p value of $10^{-8}$

• Our Finding: We applied WEAT method to same stimuli for Glo Ve. We observe the effect size of 1.50 and with p-value<10-7 for statistical significance
Results (2/8)

Musical Instruments and Weapons

• Original Finding: Greenwald et al. Report that musical instrument is pleasant and weapons are unpleasant. Based on reaction of 32 participants IAT effect size 1.66 and p-value of $10^{-10}$

• Our Finding: We applied WEAT method to same stimuli for GloVe. We observe the effect size of 1.53 and with p-value $<10^{-7}$ for statistical significance
Results (3/8)

Racial Biases

• Original Finding: Greenwald et al. report that race is indicated by name. European American names are more pleasant than African American. Based on 26 subjects, IAT effect size of 1.17 and p-value of $10^{-06}$

• Our Finding: We again replicate attitude towards two races using WEAT method to Glo Ve. We observe the effect size of 1.41 and with p-value $<10^{-8}$ for statistical significance
Results (4/8)

Resume Study

• Original Finding: Greenwald et al. report that race is indicated by name. European American names are more pleasant than African American. Based on 26 subjects, IAT effect size of 1.17 and p-value of $10^{-06}$

• Our Finding: We again replicate attitude towards two races using WEAT method to GloVe. We observe the effect size of 1.41 and with p-value $<10^{-8}$ for statistical significance
Results (5/8)

Gender Biases

- Original Finding: Nosek et al. Investigate with 38797 subjects, female names were found to be associated with family than career with an effect size of 0.72 and \( p \)-value < \( 10^{-2} \)
- Our Finding: We use the same stimuli found in Nosek. We found the same result man are more connected with career than women with an effect size of 1.81 and \( p \)-value < \( 10^{-3} \) for statistical significance
Results (6/8)

Arts and Mathematics

• Original Finding: 28108 subjects, effect size of 0.82 and $p$-value $< 10^{-2}$

• Our Finding: We found same result with an effect size of 1.06 and with $p$-value $< 10^{-2}$ for statistical significance

Arts and Sciences

• Original Finding: 83 subjects, with effect size 1.47 & $p$-value $< 10^{-24}$

• Female with arts and male with science, with an effect size of 1.24 and $p$-value of $10^{-2}$
Results (7/8)

Occupational Statistics

• Our Finding: By applying WEFAT we predict the percentage of women in the 50 most relevant occupations.

• Person correlation coefficient 0.9

With $p$-value < $10^{-18}$
Results (8/8)

Androgynous Names
- The 1990 U.S. census data about Gender and names in population
- By WEFAT, we found the percentage Of people who were women with Coefficient of 0.84 and \( p \)-value <10\(^{-13} \)

**Figure 2.** People with androgynous names
Pearson’s correlation coefficient \( \rho = 0.84 \) with \( p \)-value < 10\(^{-13} \).
Discussion (1/2)

• The origin of predjudice is the implicit transmission of ingroup / outgroup identity information though language
• Our null hypothesis helps to eliminate or atleast to quantify predjudice
• AI can and does inherit human biases that human exhibit
• Machines are artifacts, artifacts could persist and perpetuate biases in society for long time
• AI application results are explicit, so we can monitor and correct
Discussion (1/2)

• Sentiment analysis for marketing or finance such as review or market trends might have impact from bias
• NLP application result might be influenced by bias in language
• Statistical Machine Translation can reflect gender stereotypes
• Bias is not application of AI but basic representation of knowledge
• Defining prejudice algorithmically is very difficult
• Use corpora with prejudice as little as possible, Complex AI such as cognitive system or heterogenous approach can be great
Bibliography


Thank You for your attention!