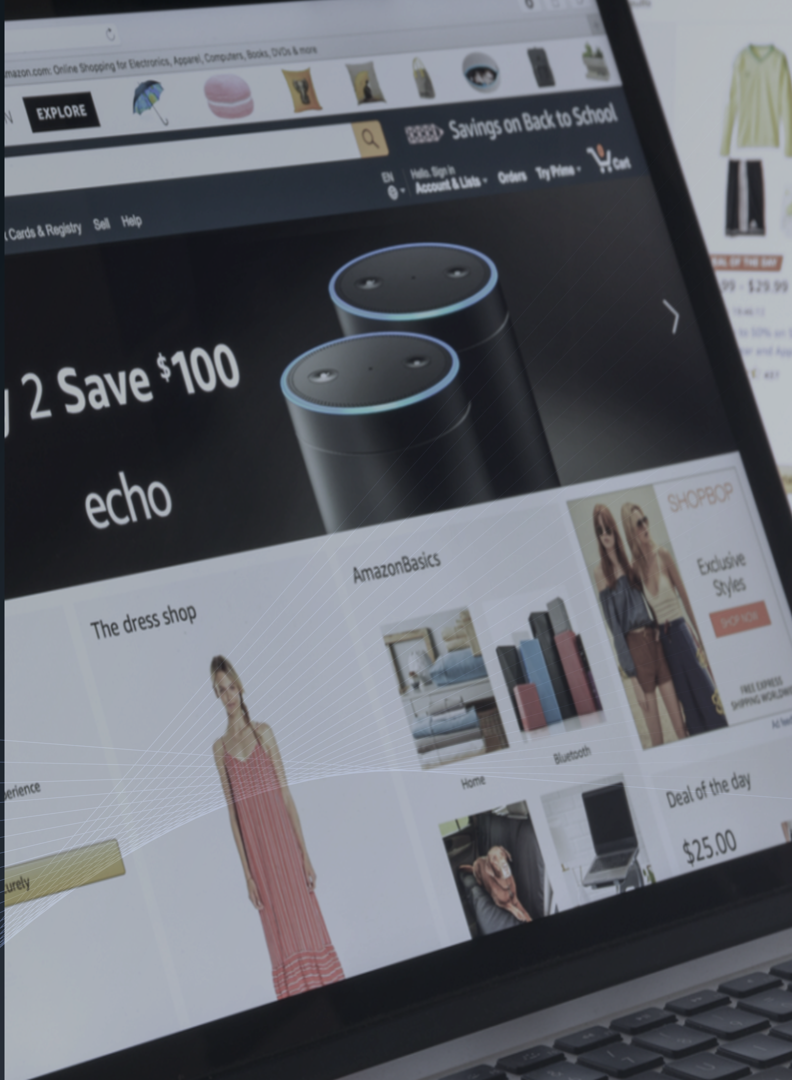
SUSS
SINGAPORE UNIVERSITY
OF SOCIAL SCIENCES

Leveraging Unsupervised Representation Learning with Reviews to Improve Top-N Recommendation in E-commerce

By: Low Wei Teck (E1881107)
Supervisor: Dr Chris Ho

Table of Content

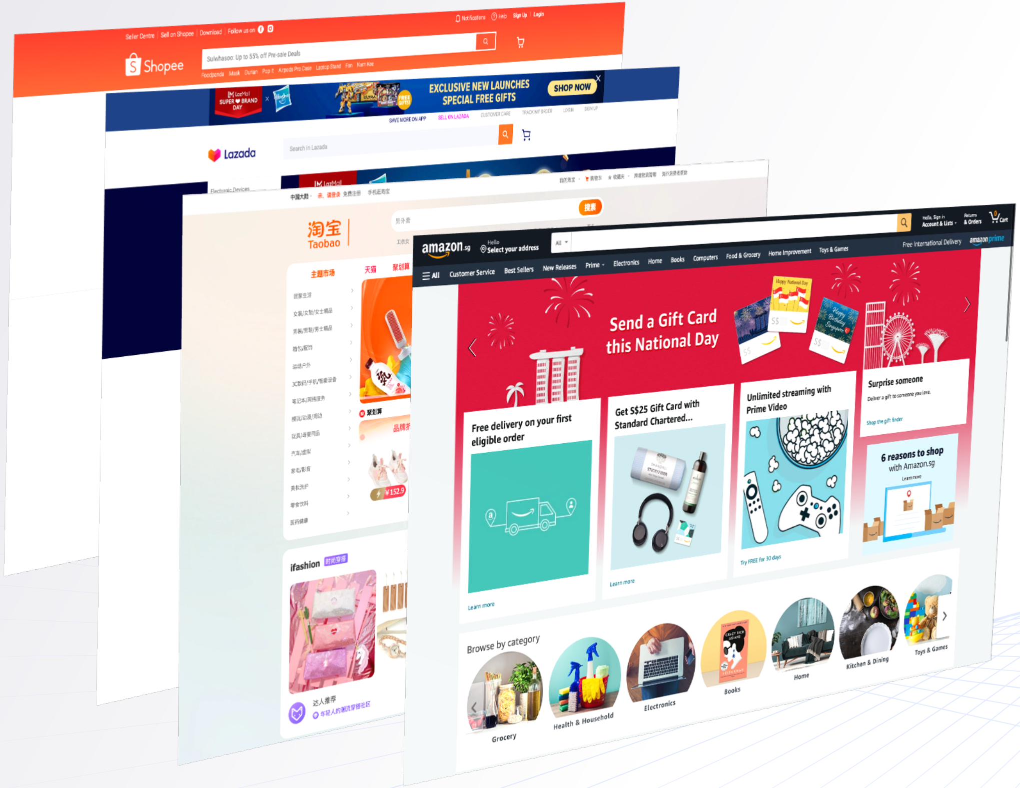


1.

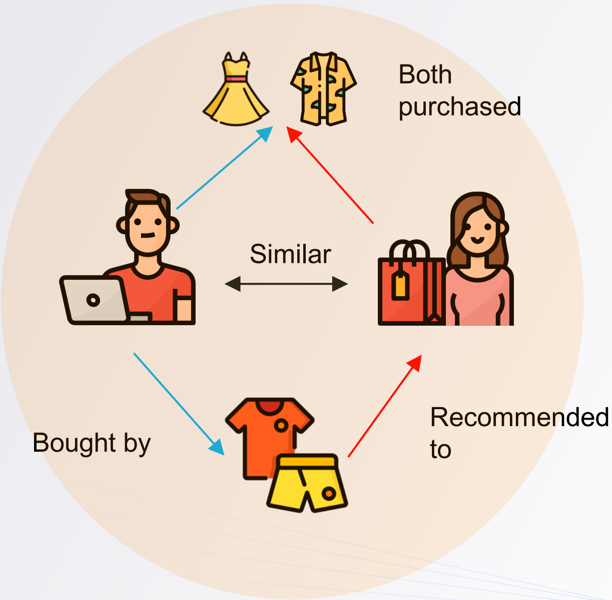
Introduction

Understanding Recommender Systems

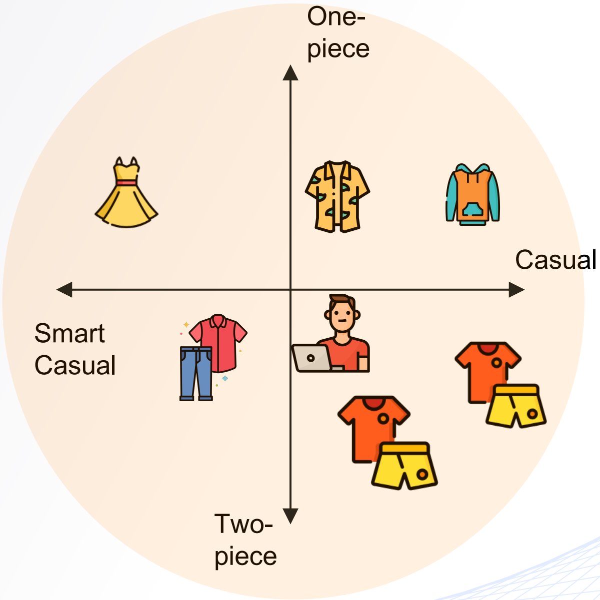
Introduction



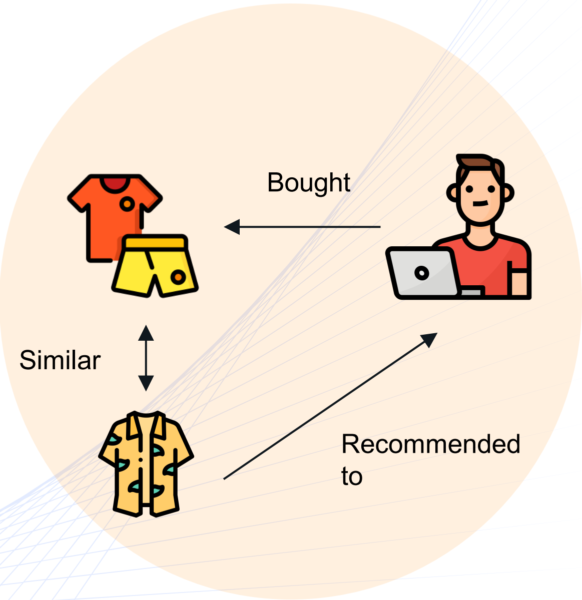
- E-commerce industry have been experiencing tremendous growth, account for **approx. 2.7% of global output.**
- **Recommender System (RS)** is a crucial feature in many E-commerce platform alleviating information overload.
- Several key problems plagued existing RS such as ***Data Sparsity*** and ***Cold Start***.



Memory-based Collaborative Filtering (e.g., *K*-nearest neighbor)



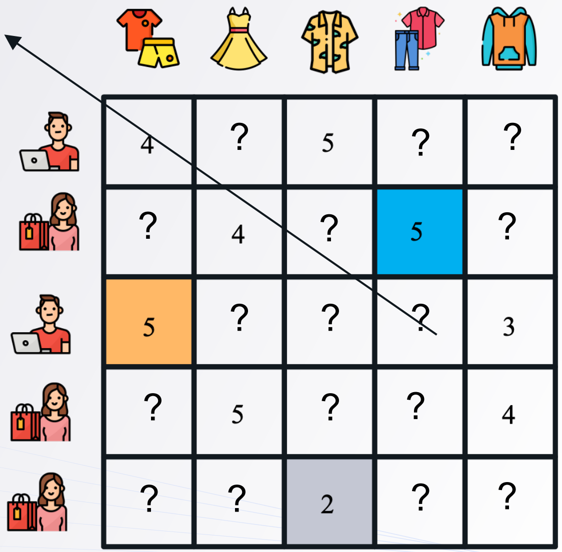
Model-based Collaborative Filtering (e.g., Matrix Factorization)



Content-based Filtering (e.g., Cosine-similarities)

Data Sparsity

Missing Data



User-Item Rating Matrix
($\boldsymbol{M} \in \mathbb{R}^{m \times n}$)

- Real world dataset including explicit ratings given by users on an e-commerce platform
  - ➢ Large number of users/items but few user ratings
- Affects ***Collaborative Filtering*** approaches
- Sparsity is denoted by as ratio of empty and total available user-item entries (ratings):
  - ➢ **Sparsity = 1** $- \frac{R}{U * I}$
  - ➢ *R* = ratings, *U* = # Users, *I* = # Items

SUSS
SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

Purchased

Purchased

Which user is easier to generate recommendation for?

Rated

Rated

5 5 5
5 5 5
5 4 3
5 2 5

5

Which items would you more confidently recommend to an user?

- Increasingly common due to the growth of e-commerce platforms
  - ➢ Rapid increase in new user/items entering the platform
- Mainly affects ***Collaborative Filtering*** approaches
- Difficult to generate recommendations for new user or recommend new items if we barely know any information on what they like or how an item is perceived.

Definition

It is increasingly challenging to generate **quality Top-N recommendations** solely based on *product ratings* given the sheer size of today's E-commerce platforms. We want to explore incorporating **new sources of information** (*product reviews*) to help improve the recommendations.

SUSS
SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

Business Objective

We are expected to recommend more **relevant items to users** and **expose** users to a much **diverse selection of products** in order to drive **user trust** in the platform to forge stronger **customer loyalty**.

Data Mining Objective

The RS is regarded as a **Top-N Recommendation** task, in which we will leverage product reviews as a new source of data to **improve recommendation accuracy** by predicting future purchases based on user purchase history.

2.

Past Works

Summary of Recent Studies

Past Works

SUSS
SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

Recommender System Based on Consumer Product Reviews

Silvana Aciar
University of Girona Campus Montilivi, Building P4, 17071 Girona, Spain
saciar@eia.udg.es

Debbie Zhang
University of Technology, Sydney, PO Box 123 Broadway 2007, Australia
debbiez@it.uts.edu.au

Simeon Simoff
University of Technology, Sydney, PO Box 123 Broadway 2007, Australia
simeon@it.uts.edu.au

John Debenham
University of Technology, Sydney, PO Box 123 Broadway 2007, Australia
debenham@it.uts.edu.au

TopicMF: Simultaneously Exploiting Ratings and Reviews for Recommendation

Yang Bao[†] **Hui Fang** **Jie Zhang**
[†] Nanyang Business School, Nanyang Technological University, Singapore
School of Computer Engineering, Nanyang Technological University, Singapore
{baoyang@ntu.edu.sg, hfang@e.ntu.edu.sg, zhangj@ntu.edu.sg}

Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text

Julian McAuley
Stanford University
jmcauley@cs.stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Combining Rating and Review Data by Initializing Latent Factor Models with Topic Models for Top-N Recommendation

FRANCISCO J. PEÑA, Insight Centre for Data Analytics, University College Dublin, Ireland
DIARMUID O'REILLY-MORGAN, Insight Centre for Data Analytics, University College Dublin, Ireland
ELIAS Z. TRAGOS, Insight Centre for Data Analytics, University College Dublin, Ireland
NEIL HURLEY, Insight Centre for Data Analytics, University College Dublin, Ireland
ERIKA DURIAKOVA, Insight Centre for Data Analytics, University College Dublin, Ireland
ARRY SMYTH, Computer Science, University College Dublin, Ireland
ONGHUS LAWLOR, University College Dublin, Ireland

From Free-text User Reviews to Product Recommendation using Paragraph Vectors and Matrix Factorization

Georgios Alexandridis
School of Electrical and Computer Engineering
National Technical University of Athens
Zografou, Greece
gealexandri@islab.ntua.gr

Thanos Tagaris
School of Electrical and Computer Engineering
National Technical University of Athens
Zografou, Greece
thanos@islab.ntua.gr

Giorgos Siolas
School of Electrical and Computer Engineering
National Technical University of Athens
Zografou, Greece
gsiolas@islab.ntua.gr

Andreas Stafylopatis
School of Electrical and Computer Engineering
National Technical University of Athens
Zografou, Greece
andreas@cs.ntua.gr

- Initial works were **manual** in nature in an ontology-based RS (Aciar et al., 2006).
- **Unsupervised techniques** emerges when incorporating review texts in RS (McAuley & Jeskovec, 2013; Bao et al., 2014; Alexandridis et al., 2019; Peña et al., 2020).
- Latest techniques were complex and review text were generally represented in ways that lack **semantics** and **context**, except in Alexandridis et al., (2019).
- Generally, studies conducted in the form of **rating prediction**.

3.

Data Understanding & Preparation

Exploratory Data Analysis and Text Pre-processing

Table 1

Amazon Dataset Characteristics by Category

Category	#Users	#Items	#Reviews	Sparsity (%)
Pet Supplies	19,738	4,950	94,334	**99.89**
Grocery and Gourmet Food	14,418	4,799	77,604	**99.89**

Note. Characteristics are defined by words per review, with the exception of *Unique*, which covers the entire review corpus.

SUSS
SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

Table 2

Data Dictionary for Finalized Dataset

Feature Name	Description	Data Type	Values
asin	Amazon ASIN	String	A1358PQON9ZAK4
title	Product Title	String	Ateco Food Coloring Kit, 6 colors
categories	Product Category	String	['Grocery & Gourmet Food', 'Cooking & Baking', 'Food Coloring']
reviewerID	Reviewer ID	String	0000013714
overall	**Product Ratings**	**Numeric**	**1-5**
reviewText	**Product Review**	**String**	**“The product is good…”**
reviewTime	**Date of review**	**String**	**“05 28, 2013”**

Table 3

Raw Review Texts Characteristics by Length

Category	Minimum	Maximum	Mean	Median	Unique
Pet Supplies	1	2,652	87.44	58	231,430
Grocery and Gourmet Food	1	3,368	91.84	64	230,912

Note. Characteristics are defined by words per review, with the exception of *Unique*, which covers the entire review corpus.

Text Pre-processing Flowchart

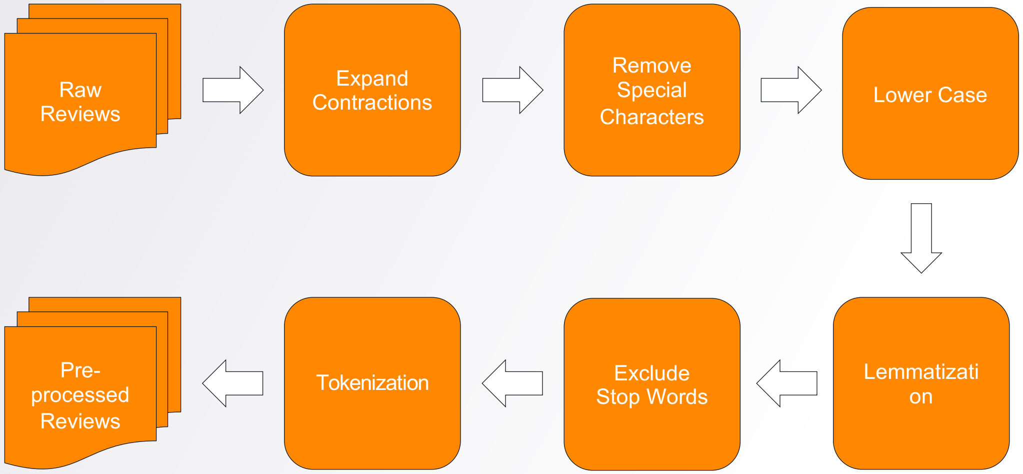


Figure 1. Text Pre-processing Flowchart

NLTK

TextBlob

Raw Review: *e.g., "This curry paste makes a delicious curry. I just love the..."*

Contractions: *e.g., "These aren't delicious -> "These are not delicious"*

Special Characters: *e.g., "$%^&**

Lower Case: *e.g., "The curry paste…" -> "the curry paste…"*

Lemmatization: *e.g., "loving" -> "love"*

Stop Words: *e.g., "your", "a", "the", "and", "me"*

Tokenization: *e.g., ['curry', 'paste', 'delicious', …]*

Pre-processed Review: *"e.g., curry paste delicious curry fry chicken vegetable"*

4.

Modelling

Building a review-based Recommender System

Understanding Paragraph Vector Model (Doc2Vec)

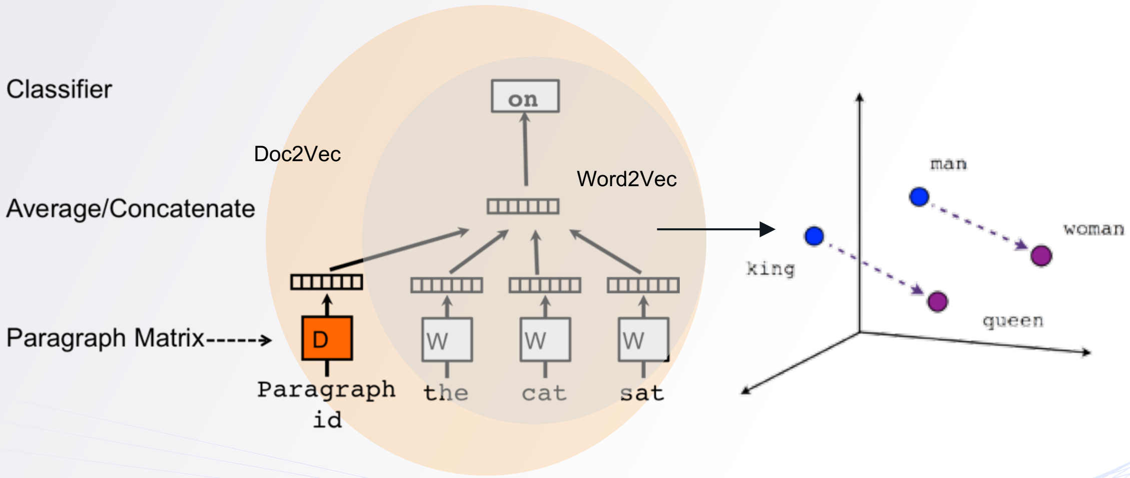


Figure 2. Example of PV-DM model. (Le & Mikolov, 2014).

- **Word2Vec:** representation of words in real-valued vectors, encoding meaning of the words such that similar words are closer to each other in a vector dimension space.
- **Doc2Vec**: sum/mean aggregation of *word representation vectors*, to represent an entire document (*reviews*) as a vector representation, where similar documents are closer in vector dimension space.

Table 5

Comparison of NLP Methods for Feature Extractions on Text

Methods	TF-IDF	LDA	**Doc2Vec**
Type	Bags-of-Word	Probabilistic	**Neural Network**
Usage	Vectorization	Topic Modelling	**Vectorization**
Semantics	No	No	**Yes**
Context	No	No	**Yes**

“ Paragraph vectors unlike TF-IDF, Topic Modelling vectors, considers word semantics and order.

We can assume that by better capturing the rich information within reviews, we can generate better user and item profile representations.

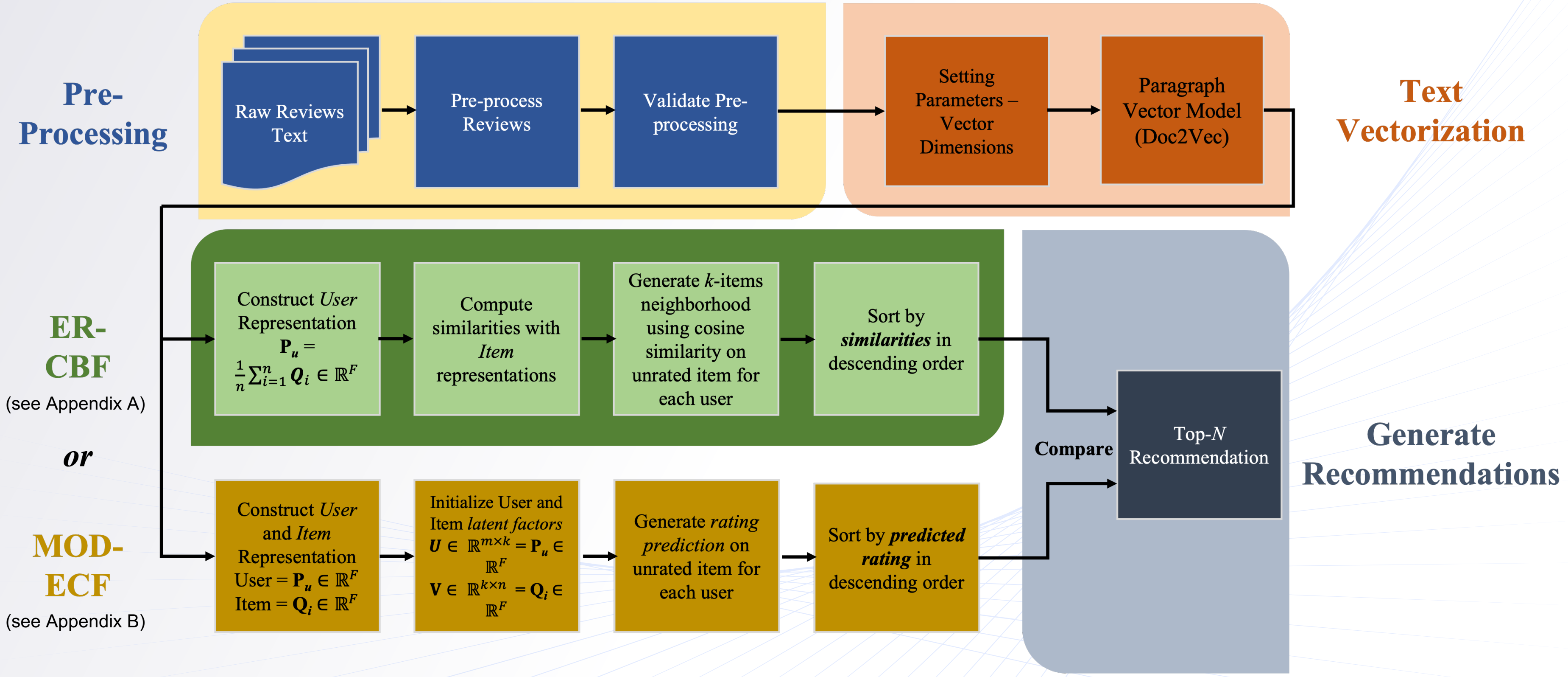


Figure 3. Proposed Modelling Process

SUSS
SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

"Baseline Comparisons

Traditional

- User-based Collaborative Filtering (UB-CF)
- FunkSVD (Matrix Factorization)

Previous Work

- Topic-initialized Matrix Factorization (TI-MF)
  - Initializing FunkSVD using Topic Modelling vectors
  - Similar to Peña et al. (2020)

Random

- Random Normal
  - Predict random ratings given normal distribution.

5.

Evaluation

Evaluating the Recommender System’s Performance

SUSS
SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

“Predict future purchases based on past purchases”

Why

Generate Top-*N* recommendations and compare to latest holdout test items based on user’s purchase history, mimicking a **typical consumer lifecycle.**

How

Order user’s rating history chronologically based on *review time*, and split history by **70/30** ratio, leaving 30% of the items for evaluation.

Caveat

Minimal user history should have **two items** to allow possible splitting. User and item in test set should both appear in the train set.

SUSS
SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

Recall@N: *how many items recommended did users already purchased?*

$$\text{Recall@}N = \frac{\boldsymbol{\#\, relevant\ items\ recommended}}{\boldsymbol{\#\, relevant\ items}} = \frac{\#\, test\ set\ items\ recommended}{\#\, test\ set\ items}$$

*Novelty@N: *are we able to recommend items that are not as popular?*

$$Novelty@N = \frac{\boldsymbol{\sum_{i=1}^{N}(1 - popularity(i))}}{\boldsymbol{N}} = \frac{\sum_{i=1}^{N}\left(1 - \frac{\#reviews_i}{\#\, max\ reviews}\right)}{N}$$

Results: Recall@N

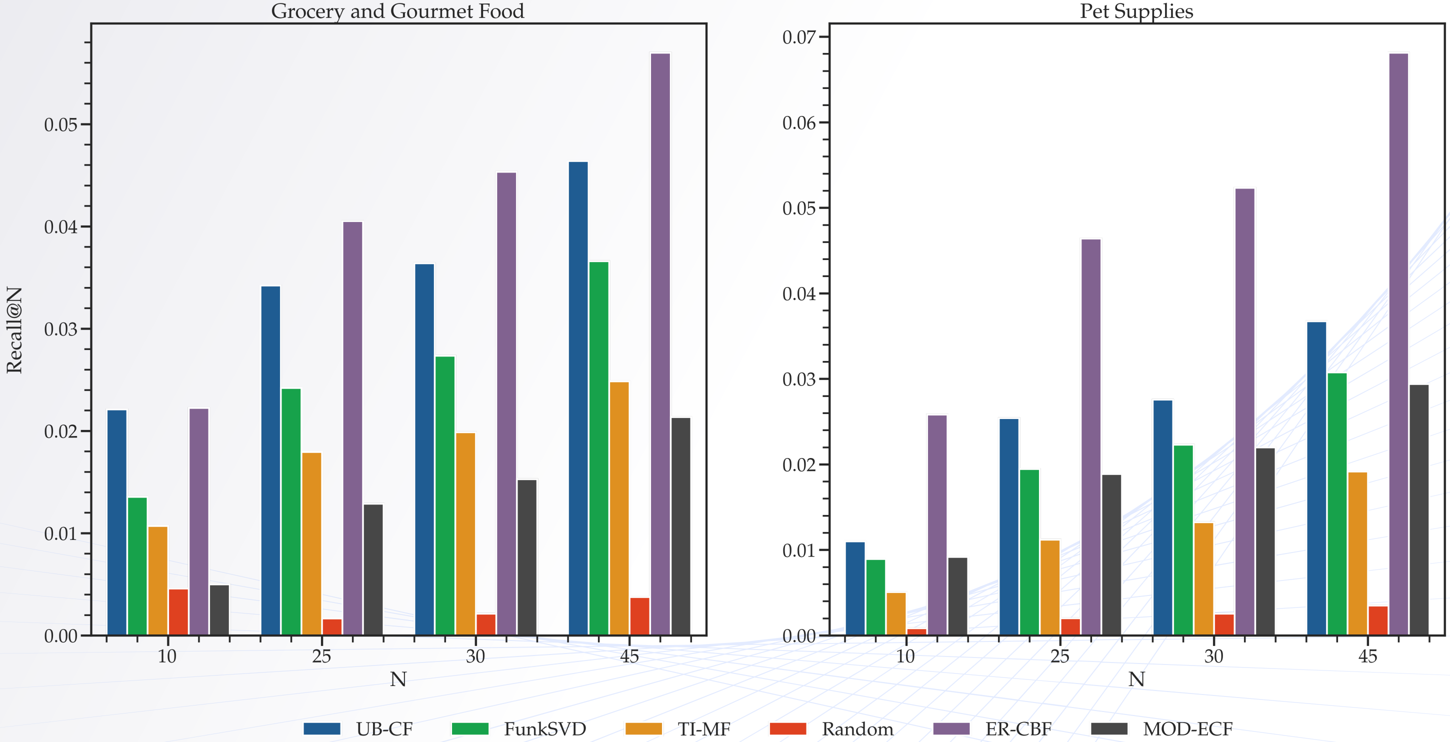


Figure 4. Recall@N, where N = {10, 25, 30, 45}

SUSS
SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

Table 6

Top-N performance evaluated by Recall@N

Algo.	Amazon Grocery and Gourmet Food				Amazon Pet Supplies			
	Recall@10	Recall@25	Recall@30	Recall@45	Recall@10	Recall@25	Recall@30	Recall@45
UB-CF	0.02210	0.03423	0.03640	0.04639	0.01101	0.02542	0.02758	0.03673
FunkSVD**	0.01355	0.02420	0.02735	0.03659	0.00893	0.01945	0.02230	0.03075
TI-MF**	0.01070	0.01795	0.01987	0.02485	0.00507	0.01119	0.01325	0.01916
Random	0.00046	0.00166	0.00213	0.00376	0.00085	0.00202	0.00256	0.00349
*ER-CBF**	**0.02225**	**0.04052**	**0.04535**	**0.05699**	**0.02583**	**0.04642**	**0.05234**	**0.06814**
*MOD-ECF****	0.00499	0.01288	0.01528	0.02136	0.00919	0.01888	0.02200	0.02940

Note. All matrix factorization-based algorithms are defined with a factor size of k=50. Similarly, the review representations in our proposed approaches are defined with a dimension size of n=50.

- *Doc2Vec settings: {epochs=10, window=5, min_count=1}
- **Matrix Factorization models settings: {epochs=10, learning_rate=0.005, regularization=0.1}

Results: Novelty@N

SUSS SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

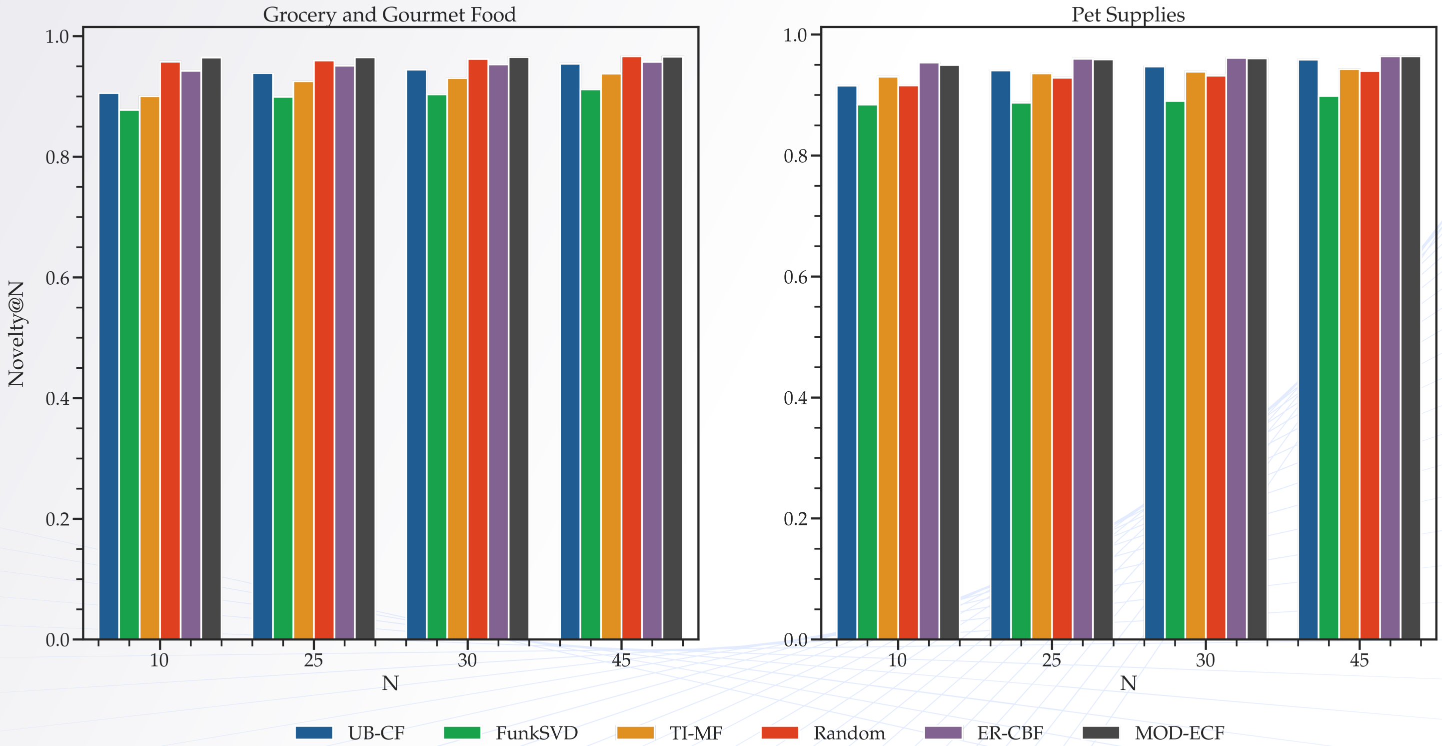


Figure 5. Novelty@N, where N = {10, 25, 30, 45}

SUSS SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

Table 7

Top-N performance evaluated by Novelty@N

Algo.	Amazon Grocery and Gourmet Food				Amazon Pet Supplies			
	Novelty@10	Novelty@25	Novelty@30	Novelty@45	Novelty@10	Novelty@25	Novelty@30	Novelty@45
UB-CF	0.90557	0.93876	0.94462	0.95401	0.91555	0.94071	0.94731	0.95870
FunkSVD**	0.87762	0.89942	0.90351	0.91186	0.88427	0.88741	0.89002	0.89838
TI-MF**	0.90050	0.92528	0.93025	0.93792	0.93035	0.93601	0.93856	0.94257
Random	0.95766	0.95966	0.96197	**0.96668**	0.91587	0.92872	0.93213	0.93958
*ER-CBF**	0.94243	0.95109	0.95319	0.95722	**0.95381**	**0.96009**	**0.96143**	**0.96418**
*MOD-ECF****	**0.96466**	**0.96496**	**0.96514**	0.96580	0.94959	0.95902	0.96069	0.96408

Note. All matrix factorization-based algorithms are defined with a factor size of k=50. Similarly, the review representations in our proposed approaches are defined with a dimension size of n=50.

- *Doc2Vec settings: {epochs=10, window=5, min_count=1}
- **Matrix Factorization models settings: {epochs=10, learning_rate=0.005, regularization=0.1}

Findings & Discussion

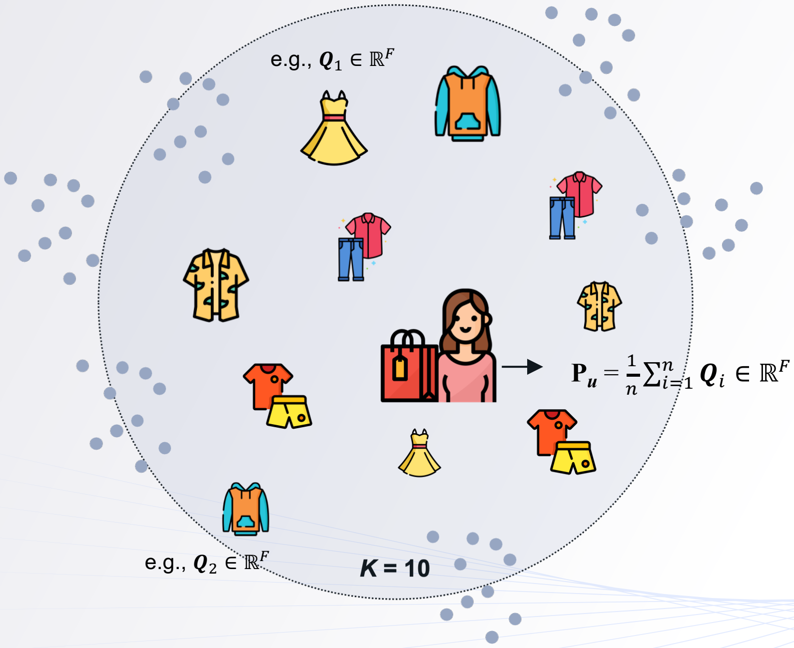


Figure 6. Visual Representation of the *ER-CBF* model.

- **ER-CBF** has the **BEST** accuracy (*Recall@N*) and adequate *Novelty@N* in relative to other models.
- **Traditional algorithms**, although reliable but were still suffering from *data sparsity* and *cold start.*
- **MOD-ECF** were unreliable in performance, despite outperforming **TI-MF** in *Pet Supplies.*
- Requires deeper cross-analysis to better understand how using review-based recommenders perform across different groups of users.

Interpreting Recommendations – Grocery and Gourmet Food

SUSS
SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

User: *A5PC6KU9TQFRY*

Purchased: *Oolong Tea x2, Green Tea, Table Salt, Soup*

ER-CBF@10:

```
For user: A5PC6KU9TQFRY:
Purchase History:
             asin                                              title
346    B00014HS2S  Prince of Peace Oolong Tea - 100 Tea Bags net ...
351    B00014HS2S  Prince of Peace Oolong Tea - 100 Tea Bags net ...
8980   B000H1195C  Kikkoman Egg Flower, Hot and Sour Soup, 0.88-O...
34315  B004772OWE  Celestial Seasonings Authentic Green Tea, K-Cu...
43219  B006GCMI5Q  Benson's - Table Tasty Salt Substitute - No Po...

Recommending:

         asin                                              title
0  B002750C06  Harney &amp; Sons Caffeinated Green Tea with C...
1  B007FRAYKC  Teavana Youthberry Wild Orange Blossom Loose-L...
2  B001EQ5Q3U  Tetley Premium Black Tea, British Blend, 80 Te...
3  B001E5E268  Stash Tea Organic Jasmine Green Tea, Loose Lea...
4  B000E63LAQ  Celestial Seasonings Green Tea, Authentic, 40 ...
5  B001EQ5JLY  PG Tips Black Tea, Pyramid Tea Bags, 80Count B...
6  B000FFS09I  Numi Organic Tea Iron Goddess of Mercy, Full L...
7  B000KOUKMY  Luzianne Specially Blended for Iced Tea, 100-C...
8  B00286KM2A  Lipton Gallon-Sized Black Iced Tea Bags, Unswe...
9  B0009TMZIM  China Tea Loose Leaf Sampler Gift Pack - 6 Tin...
```

Items Recommended:

1. Green Tea
2. Black Tea
3. Youthberry Tea

UB-CF@10:

```
For user: A5PC6KU9TQFRY:
Purchase History:
             asin                                              title
346    B00014HS2S  Prince of Peace Oolong Tea - 100 Tea Bags net ...
351    B00014HS2S  Prince of Peace Oolong Tea - 100 Tea Bags net ...
8980   B000H1195C  Kikkoman Egg Flower, Hot and Sour Soup, 0.88-O...
34315  B004772OWE  Celestial Seasonings Authentic Green Tea, K-Cu...
43219  B006GCMI5Q  Benson's - Table Tasty Salt Substitute - No Po...

Recommending:

         asin                                              title
0  B000BXSRT2    World Confections Candy Cigarettes, Pack of 24
1  B0042LH7TK  Green Mountain Coffee Fair Trade Pumpkin Spice...
2  B006846XQ2  Starbucks Tazo Tea * Awake * Black Tea, 16 K-C...
3  B000H11C6I  Blue Diamond Almond Nut-Thins Cracker Crisps, ...
4  B000J4IDO2  Coffee People K-Cups, Black Tiger Extra Bold, ...
5  B0029XDZIK  Keurig, The Original Donut Shop, Medium Roast ...
6  B004OINZZQ  Twinings of London Winter Spice Herbal Tea Bag...
7  B007PA33MA  Timothy's Decaf Colombian K-Cup Counts for Keu...
8  B002HQCWYM  Twinings of London English Breakfast Tea K-Cup...
9  B0027DUJFE  Navitas Organics Antioxidant Superfruit Blend ...
```

Items Recommended:

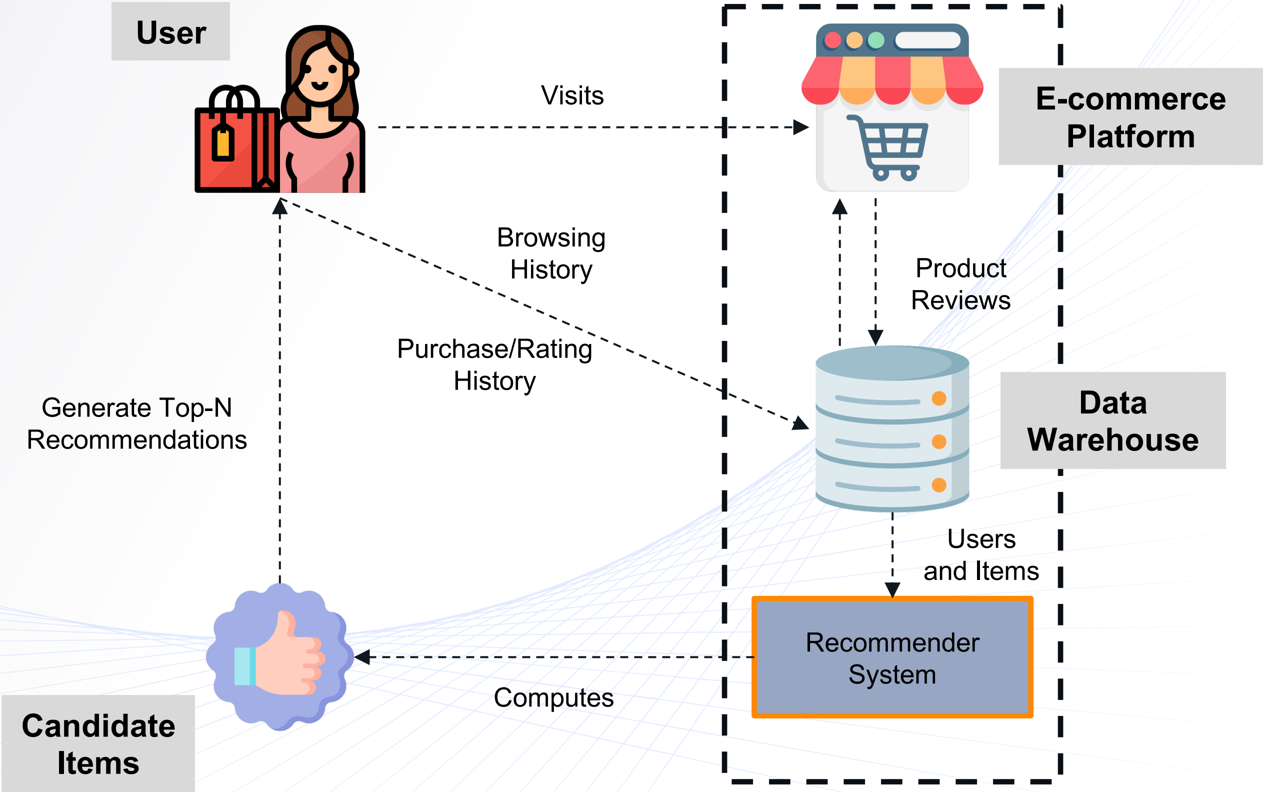
1. Candies
2. Coffee
3. Crackers
4. Breakfast/Herbal Tea

5.

Further Works

Current developments and further improvements

- Deeper Result Analysis
  - Recall@*N* & Novelty@*N* for specific groups of users (e.g., **Frequent, Average, New** users)
- **Offline** vs **Online** Evaluation
- Methods of **Deployment**
  - A/B Strategy



Thanks!

Any questions?

Appendices

Appendix A: Embedded Review Content-based Filtering (ERCBF)

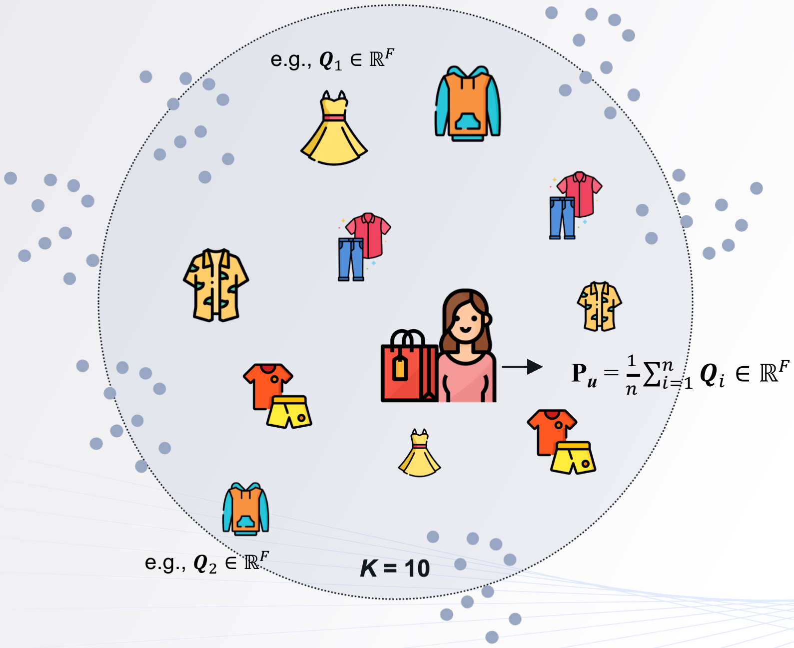


Figure 1. *K*-neighborhood of Similar Items within Paragraph Vector Dimension Space

Algorithm:

1. Find User $\mathbf{P}_U$ Rating History:
   - Sundress, $\mathbf{n}_1 = Q_1 \in \mathbb{R}^F$ = [0.252, -0.234, …]
   - Hoodie, $\mathbf{n}_2 = Q_2 \in \mathbb{R}^F$ = [0.012, 0.224, …]
   - Summer Shirt , $\mathbf{n}_3 = Q_3 \in \mathbb{R}^F$ = [-0.242, 0.111, …]
   - Sports Attire, $\mathbf{n}_4 = Q_4 \in \mathbb{R}^F$ = [-0.124, -0.104, …]
2. Represent User $\mathbf{P}_U$ as the **mean** aggregation of all rated items to:

$$\mathbf{P}_u = \frac{1}{n}\sum_{i=1}^{n} Q_i \in \mathbb{R}^F$$

3. Compute **cosine** similarities with all unrated items:

$$sim(\boldsymbol{P}_u, \boldsymbol{Q}_i) = cos\theta = \frac{P_u \cdot Q_i}{\|P_u\| * \|Q_i\|}$$

4. Formulate ***k*-neighborhood** of similar items

SUSS SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

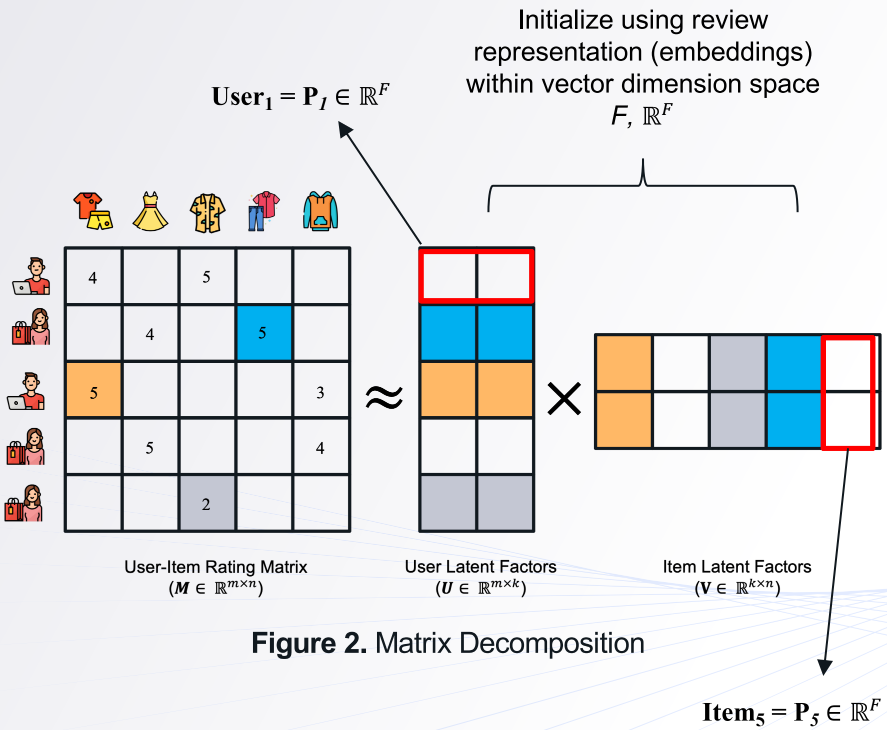


Figure 2. Matrix Decomposition

Algorithm:

1. **Initialize** latent factors of *Matrix Factorization* (MF) using review embeddings from Doc2Vec:

$$\boldsymbol{U} \in \mathbb{R}^{m \times k} = \mathbf{P}_{\boldsymbol{u}} \in \mathbb{R}^F$$

$$\mathbf{V} \in \mathbb{R}^{k \times n} = \mathbf{Q}_{\boldsymbol{i}} \in \mathbb{R}^F$$

2. Add in **biases** to factor systematic variations in user behavior in predicted rating, $\widehat{y_{iu}}$:

$$\widehat{\boldsymbol{y_{iu}}} = \mu + b_u + b_i + Q_i \cdot P_u$$

3. **Regularize** MF to prevent overfitting due to model complexity and sparse matrix:

$$\min_{p*, q*, b*} \sum_{(u,i) \in \kappa} (y_{iu} - \mu + b_u + b_i - Q_i \cdot P_u)^2 + \lambda \left(\|Q_i\|^2 + \|P_u\|^2 + {b_u}^2 + {b_i}^2 \right)$$

4. Optimize (3) using *Stochastic Gradient Descent* (**SGD**)