Session-Based Recommendation System for E-commerce: A Comparative Study of Algorithmic Approaches to Product Matching

Student: Battaglia Andrea

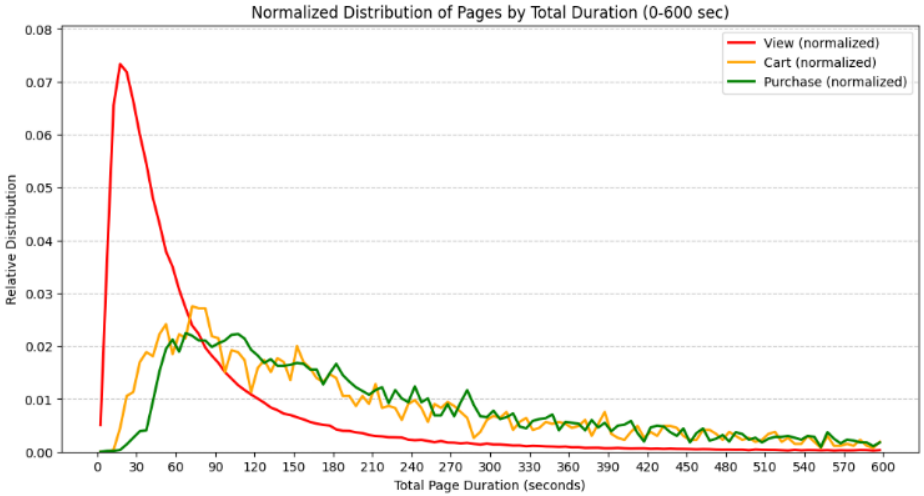
Dataset Structure

- **Source: Data comes from user interactions on an Asian e-commerce platform.**
- **Granularity: Each row represents a single user action on a product page.**
- **Each event is linked to a specific product page, allowing for behavioral analysis.**
- **Rows: 605110 rows**

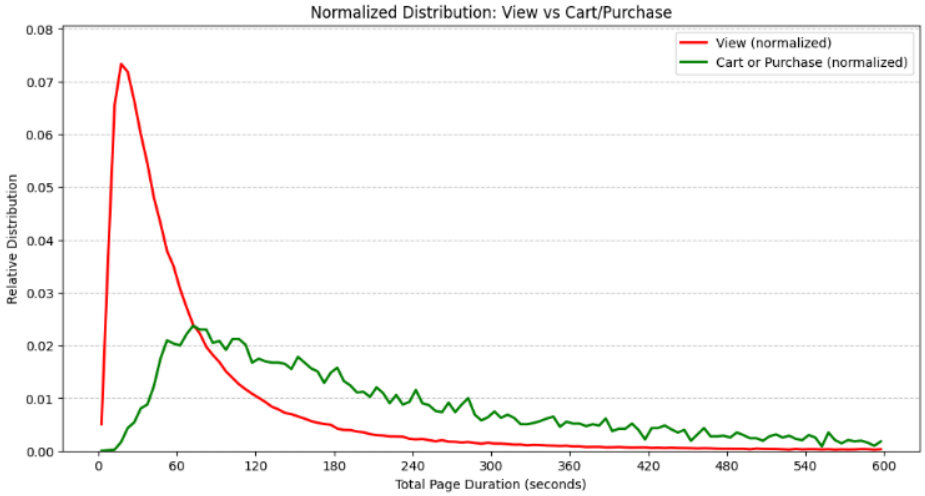
Field	Unique Count	Missing Count
event_time	35822	0
event_type	3	0
product_id	51188	0
category_id	541	0
category_code	123	191292
brand	2093	87492
price	14876	0
user_id	105887	0
user_session	141891	0

Summary of fields with counts of unique and Missing values

Distributions of Dwell times

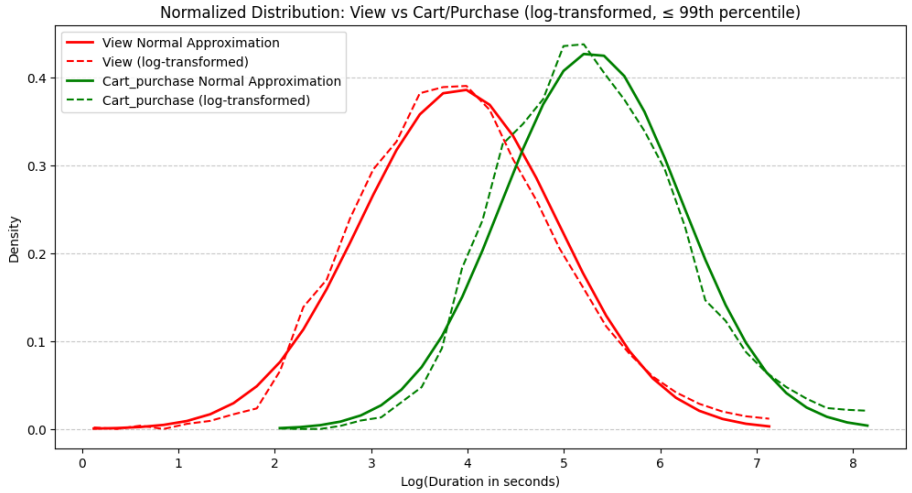


Normalized distribution of View duration by Interaction type

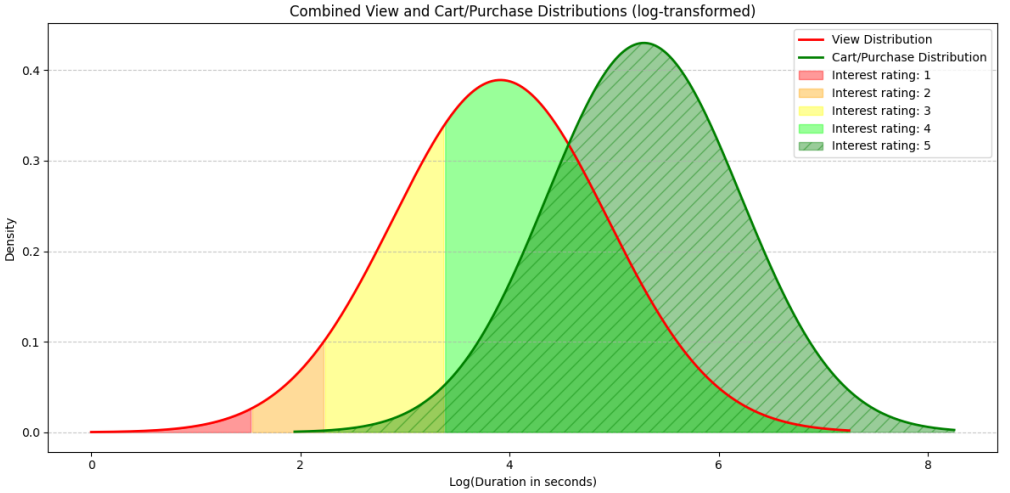


Comparison of View and Cart/Purchase Distributions

Distribution model and Interest score assignment



Normal approximation of Log-transformed distributions for View and Cart/Purchase



Visualization of Interest Scores based on Duration and Event

Application of the algorithms to the dataset

1. **Centralized Ratings:** aggregation and normalization of user-product interaction scores.
2. **Category Similarity:** Jaccard Similarity with MinHashing to compute similar product categories (two categories are considered similar if both are frequently viewed within the same session)**.**
3. **Product comparisons with two different approaches:**
    a. **Cosine Similarity** for precise vector-based similarity**.**
    b. **SimHash + LSH** with random hyperplanes for fast approximate matching.

Measuring Category Similarity with Jaccard Similarity and MinHashing

Category_1_Code	Category_2_Code	Similarity
furniture.kitchen.table	sport.tennis	0.25
apparel.jacket	apparel.trousers	0.23
accessories.bag	accessories.wallet	0.22
computers.cpu	computers.motherboard	0.2
apparel.shoes.slipons	apparel.shoes.moccasins	0.18
auto.accessories.player	electronics.audio.subwoofer	0.17
country_yard.hammok	kids.swing	0.16

Some of the most similar Product categories based on User behavior

Categories: furniture.kitchen.table and sport.tennis

Categories: apparel.shoes.slipons and apparel.shoes.moccasins

Categories: country_yard.hammock and kids.swing

Analysis of Product Recommendations by 2 approaches

Example: Laptop recommendations

Selected Product:

- **Product ID: 1307112**
- **Category: computers.notebook**
- **Brand: HP**
- **Price: €1146.14**

Approach 1: Cosine Similarity on candidates

product_id	category_code	brand	price	Similarity
1305843	computers.notebook	dell	1183.81	0.50
1306751	computers.notebook	lenovo	1415.71	0.27
1306361	computers.notebook	acer	1152.93	0.26
1307329	computers.notebook	hp	1145.20	0.23
1307211	computers.notebook	hp	1201.83	0.23
1306424	computers.notebook	hp	1003.63	0.23
1304106	computers.notebook	lenovo	978.07	0.22
1306249	computers.notebook	hp	1029.37	0.22
1306422	computers.notebook	hp	1019.07	0.22
1307080	computers.notebook	asus	957.81	0.19

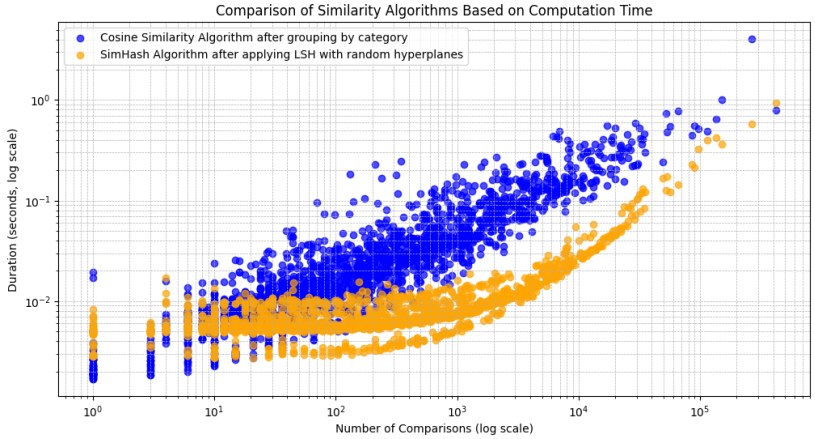
Recommended laptops based on Cosine Similarity

Approach 2: SimHash with LSH using random hyperplanes

product_id	category_code	brand	price	Similarity
1306424	computers.notebook	hp	1003.63	0.66
1307244	computers.notebook	hp	1317.67	0.61
1306296	computers.notebook	asus	746.22	0.61
1307420	computers.notebook	lenovo	2470.72	0.61
1305843	computers.notebook	dell	1183.81	0.60
1307211	computers.notebook	hp	1201.83	0.60
1306484	computers.notebook	hp	1106.59	0.59
1307107	computers.notebook	asus	1282.14	0.59
1307080	computers.notebook	asus	957.81	0.59
1305840	computers.notebook	dell	1209.55	0.59

Recommended laptops based on SimHash Similarity

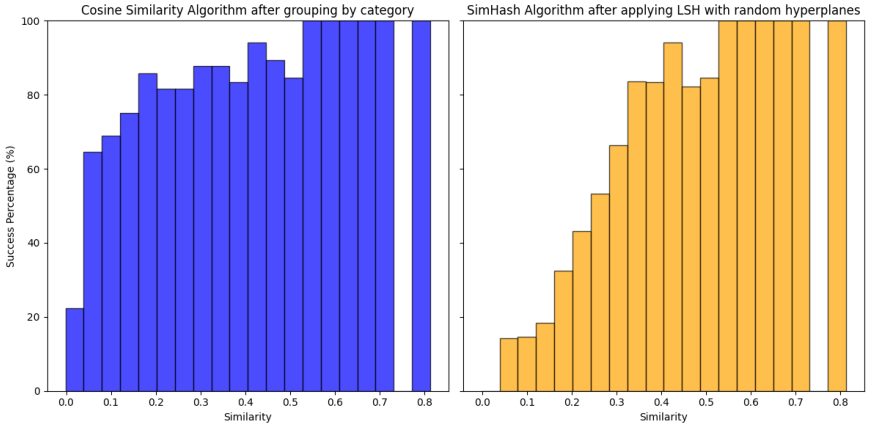
Algorithms comparison: Efficiency vs Accuracy



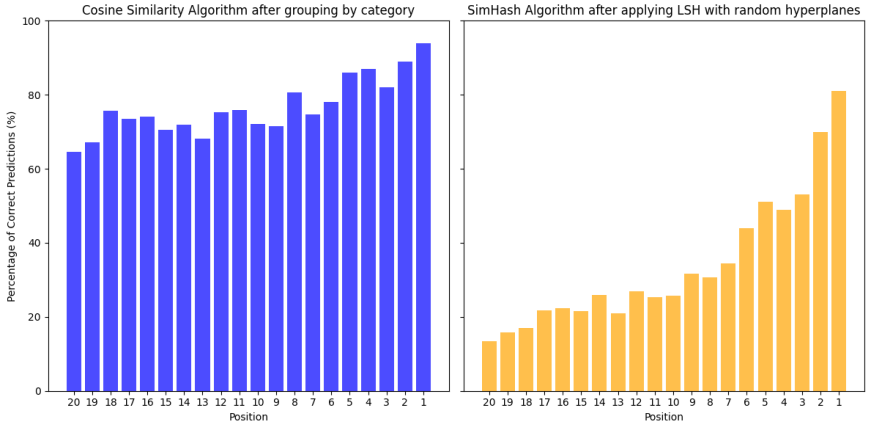
Algorithm 1 with Cosine Similarity using similar categories

Algorithm 2 with SimHash after using random Hyperplanes

Computation time vs. Number of comparisons for both algorithms



Precision of recommended products by true Cosine Similarity



Precision of recommended products by true position on ranking