Recommender System for Woongjin Book Club

kidaptive

Woongjin Book Club Platform

Overview of Book Club Platform

- Woongjin Learners are provided with 3 different platforms
  - Reading Platform (Book Club)
  - Learning Platform (Book Club Study)
  - Video Call Platform (Together)

- Book Club Platform is a library of resources from which learners can choose titles to read (or listen to), as well as answer comprehension questions and take free-form notes.

- Resource/Media types could be Audio, Video or Books.

Recommender System

Overview of Recommender System

- Recommendation Engines are a subclass of information filtering system that seek to predict the item preference for an user.
- Recommender systems typically produce a list of recommendations in one of two ways – either by Collaborative or Content-based filtering.

- Collaborative filtering approaches build a model from a user's past behaviour as well as similar decisions made by other users. This model is then used to predict items for an active user.
- Content-based filtering approaches utilize a series of discrete characteristics of an item in order to recommend additional items with similar properties.
- These approaches are often combined in Hybrid Recommender Systems.

Types of Recommender Systems (1/6)

- Random-Based Recommender

  Recommend random items, typically used as a benchmark.

- Popularity-Based Recommender

  Items which are viewed by most learners are recommended.

  In addition user features were used to provide popular recommendations matching the active user's age and/or gender

Types of Recommender Systems (2/6)

- Content-Based Recommender

  "If you liked this item, you might also like …"

  Content recommenders don't need ratings or even implicit user preferences. Good at finding items that are similar to other items preferred by user, but not so hot at finding something new.

  Book Name, Author and Keywords (korean-english translation using Google API) were used as item profile. User profile was built by combining profiles for all items used by user.

  Similarity score between a given user profile and item profile was computed using weighted average of the following similarity scores.

  a. Jaccard Similarity for Author Name. (0.25)
  b. Jaccard Similarity for Tokens of Book Name. (0.25)
  c. Log of Term Frequency Similarity for Keywords. (0.50)

  Items which are most similar to user profile were recommended.

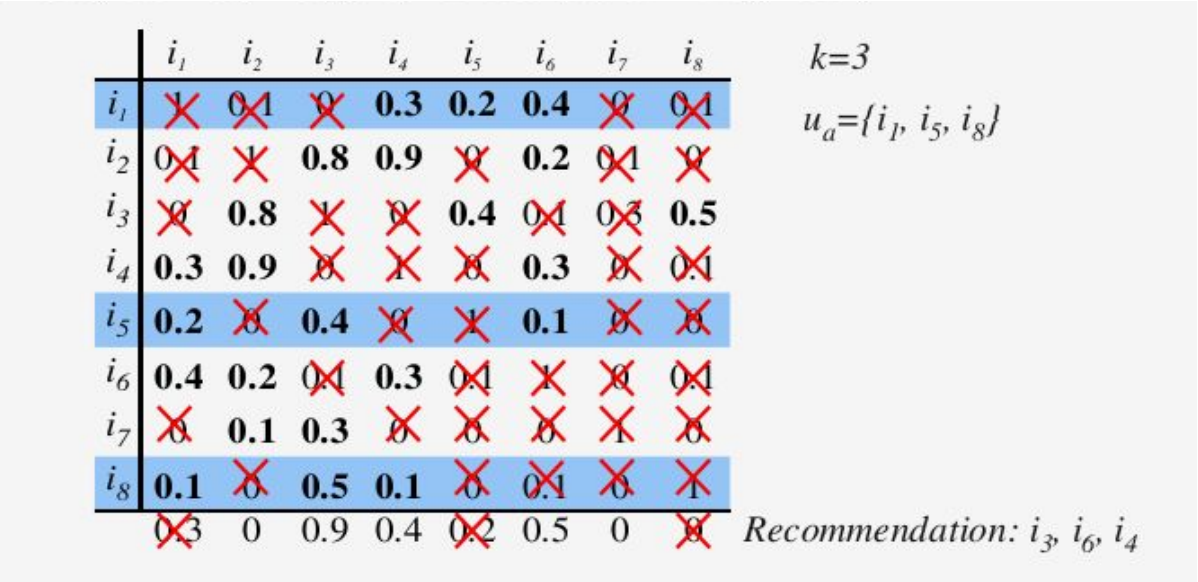
Types of Recommender Systems (3/6)

- Item-Based Collaborative Filtering

  "Customers who liked this item also liked …"

  Item-Item Similarity Matrix is computed using co-occurrence of users.

  Items which are similar to items accessed by a user are filtered and recommended in decreasing order of averaged similarity scores.



	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
i_1	1	0.1	0	**0.3**	**0.2**	**0.4**	0	0.1
i_2	0.1	1	**0.8**	**0.9**	0	**0.2**	0.1	0
i_3	0	**0.8**	1	0	**0.4**	0.1	0.3	**0.5**
i_4	**0.3**	**0.9**	0	1	0	**0.3**	0	0.1
i_5	**0.2**	0	**0.4**	0	1	**0.1**	0	0
i_6	**0.4**	**0.2**	0.1	**0.3**	0.1	1	0	0.1
i_7	0	**0.1**	**0.3**	0	0	0	1	0
i_8	**0.1**	0	**0.5**	**0.1**	0	0.1	0	1
	0.3	0	0.9	0.4	0.2	0.5	0	0

Types of Recommender Systems (4/6)

- User-Based Collaborative Filtering

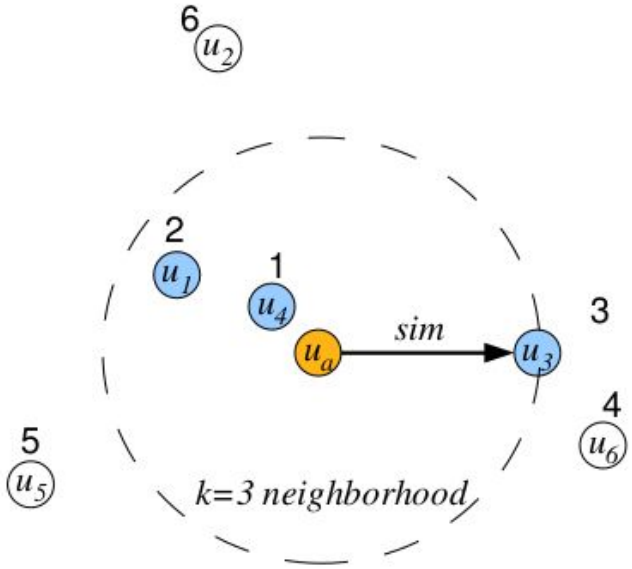
  "Customers who are similar to you also liked …"

  User-User Similarity Matrix is computed using co-occurrence of items.

  Most similar users for a given user are filtered. Items accessed by most similar users are recommended in decreasing order of averaged similarity scores weighted by user similarity.

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_a	0	0	1	1	0	1	0	1
u_1	0	1	1	1	1	0	0	0
u_2	1	0	0	0	1	1	0	0
u_3	0	0	0	1	0	1	0	1
u_4	0	0	0	1	1	1	0	1
u_5	1	1	0	0	0	0	0	1
u_6	0	1	0	0	1	1	0	1
	0	1	~~1~~	~~3~~	2	~~2~~	0	~~2~~

Recommendation: i_5, i_2



Types of Recommender Systems (5/6)

- Customized User-Based Collaborative Filtering

  User-User Similarity Matrix is computed using cosine similarity of user age and/or item_type preference.

  Item types includes video, audio and books. For each user, ratio of no of item_types used is computed as item type preference.

  3 models based on age, item type preference, and combination of age and item type preference are generated

- Content Boosted Collaborative Filtering

  Uses content-based predictions to convert a sparse user-item ratings matrix into a full ratings matrix; and then uses CF to provide personalized suggestions through collaborative filtering.

  2 models based on item-based and user-based colloborative filtering are generated.

kidaptive

Types of Recommender Systems (6/6)

- Custom Hybrid Recommender

  Similarity of users is computed by weighted average of :

  1) jaccard similarity from user based collaborative filtering (0.75)
  2) cosine similarity from age and/or item_type preference (0.25)

  3 models based on age, item type preference, and combination of age and item type preference are generated

- Generic Hybrid Recommender

  Framework to configure and run multiple recommenders, and combine item recommendations using weighted average of item scores.

  {books_rec_item_based_cf.ItemBasedCFRecommender : 0.5,
  books_rec_user_based_cf.UserBasedCFRecommender : 0.5}

Descriptive Analytics

Insights about Data

Metadata

- Metadata of books contained Book_Name, Author, and Keywords in Korean Text, these were translated using Google API.
- Metadata of learners contained Gender and Data of Birth(DOB). Age was derived from DOB. Item type preference was derived based on percentage of item access.

Data Preprocessing

- Unique key for item : “Book_Code ”, Unique key for user:“Learner_Id”
- Eliminated item and users with null id.
- Number of open and close events were unequal, so time spent on a resource couldn’t be computed. Hence we decided to filter on close events (book_close, audio_close, video_close) events.

Data Preprocessing

Total:

No of learners : 243,344, No of books : 11,490

After merging with user metadata

No of learners : 89,519 (- 153,825) No of books : 10,564 (- 926)

After merging with books metadata

No of learners : 85,047 (- 4,472) No of books : 9,093 (- 1,471)

No of Book Access (events_count)

min : 1, mean : 1.5, 75% : 2, max : 299

Filtered users who have access books at-least 3 times

No of learners : 22,435 (- 62,612) No of books : 5,203 (- 3,890)

kidaptive

Data Preprocessing

Presence of Age Outliers

min : 0, mean : 8.35, max : 118

Filtering Learners in 5<=Age Range<=20

No of learners : 21802 (- 633) No of books : 5156 (- 47)

Presence of No of Books Outliers

min : 1, mean : 2.82, max : 135

Filtering Learners who have at-least accessed 20 books

No of learners : 326 (- 21,476) No of books : 2426 (- 2730)

Random Split on Users (70%-30%)

Train Data: No of learners : 229 No of books : 2048

Test Data: No of learners : 97 No of books : 1280

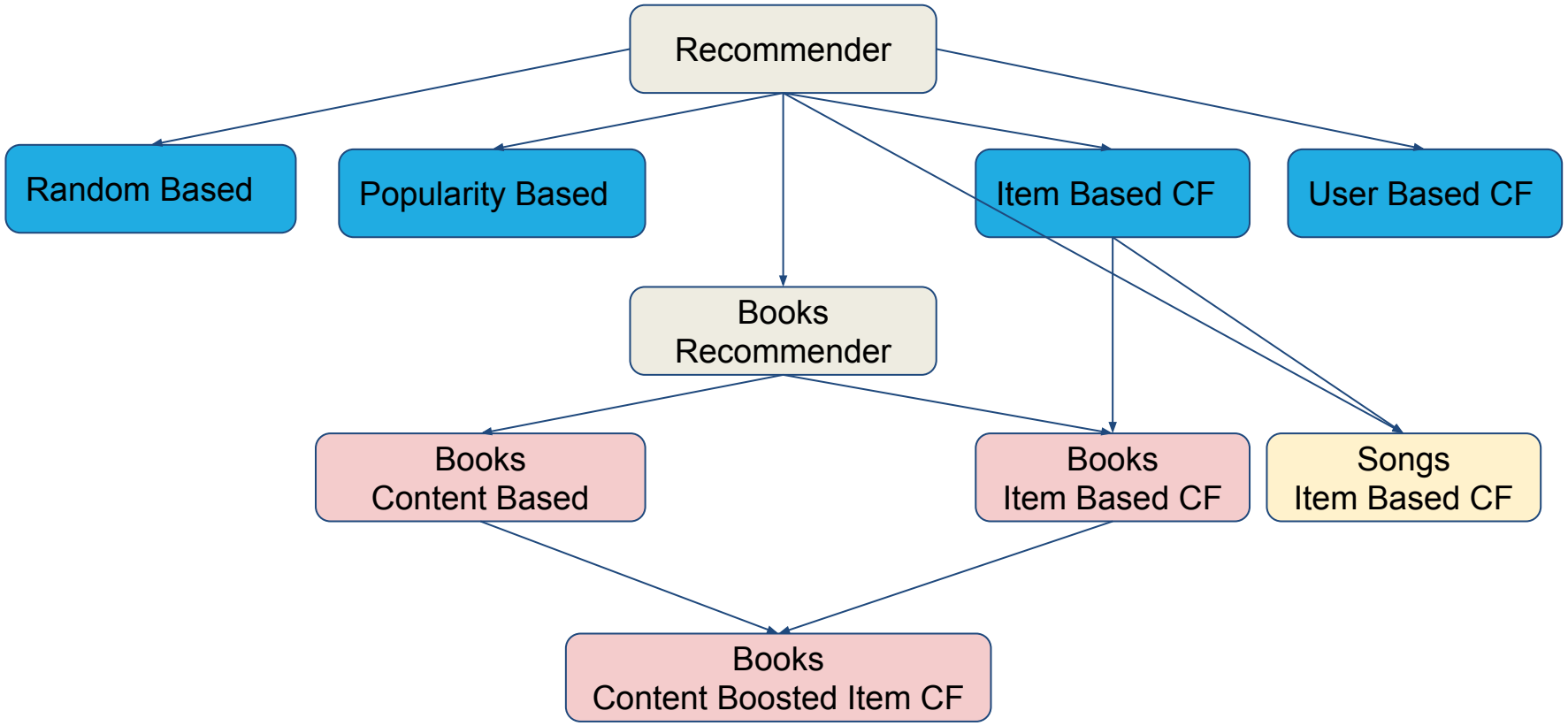
Kfolds (10) Split on Users

Train Data: No of learners : 293 No of books : 2327

Test Data: No of learners : 33 No of books : 612

Framework

Recommender Framework



kidaptive

Evaluation

Evaluation Setup

Train and Test Data

- A single instance can be generated by considering specified percentage of users as test users and remaining as train users
- Multiple sets of train and test data can be generated using k-folds of users for cross-validation.

K-folds	No of Users	No of Items	Avg No of Items/User
Train Data	293	~2300	32
Test Data	33	~625	33

Offline Evaluation

For each user in test data, 50% of items interacted where held out, recommender was used to predict a set of items that the user will use.

Evaluation Metrics

Each recommendation is classified as:

a) true positive (TP, an interesting item is recommended)

b) true negative (TN, an uninteresting item is not recommended)

c) false negative (FN, an interesting item is not recommended)

d) false positive (FP, an uninteresting item is recommended)

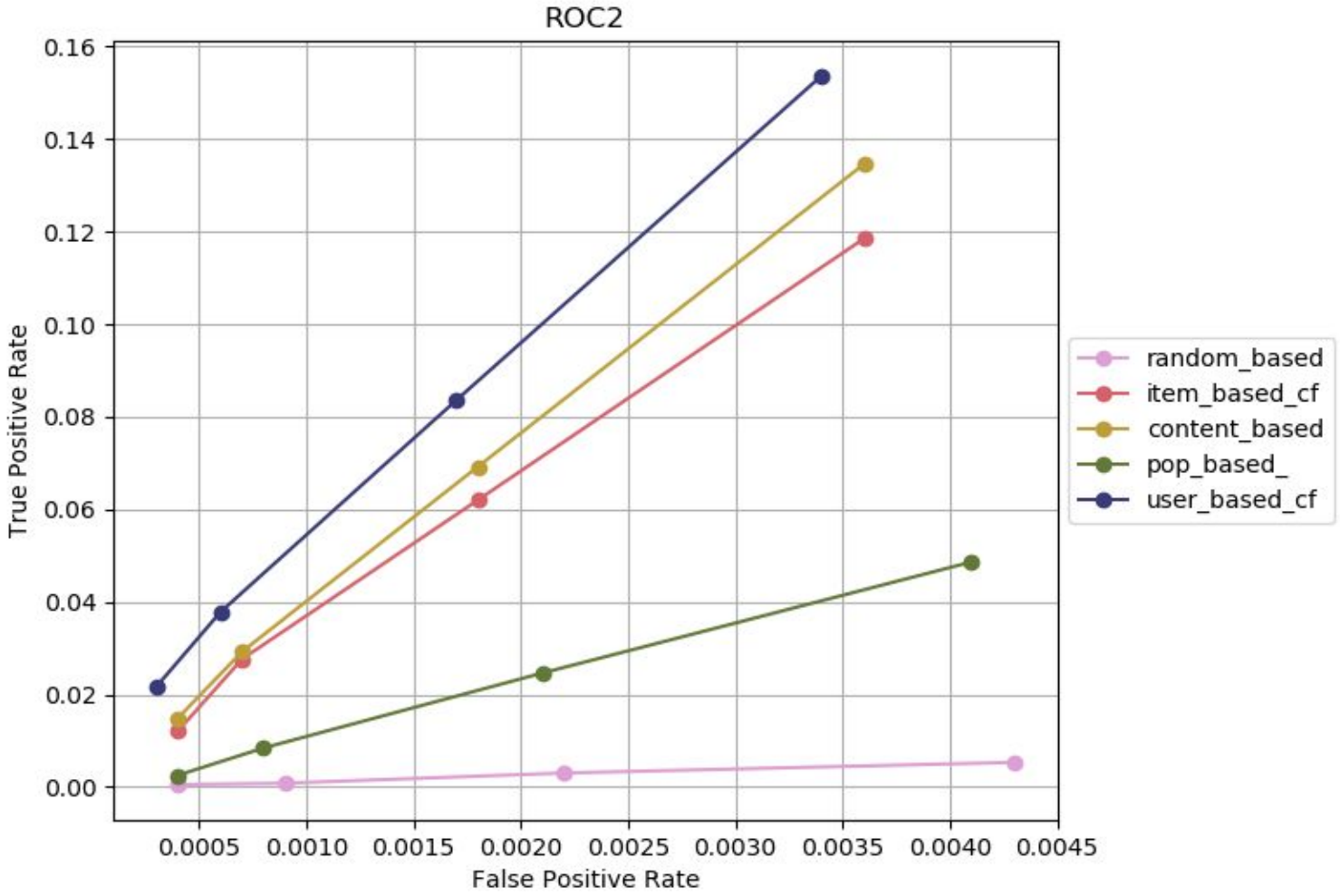
ROC2 is suitable for evaluating binary and non binary dataset.

P. Cremonesi E. Lentini M. Matteucci and R. Turrin "An evaluation methodology for recommender systems " 4th Int. Conf. on Automated Solutions for Cross Media Content and Multi-channel Distribution pp. 224-231 Nov 2008.

Receiver Operating Characteristic (ROC2) is a graphical technique that uses two metrics, true positive rate (TPR) and false positive rate (FPR) to visualize the trade-off between TPR and FPR by varying the no of recommendations returned to the user.

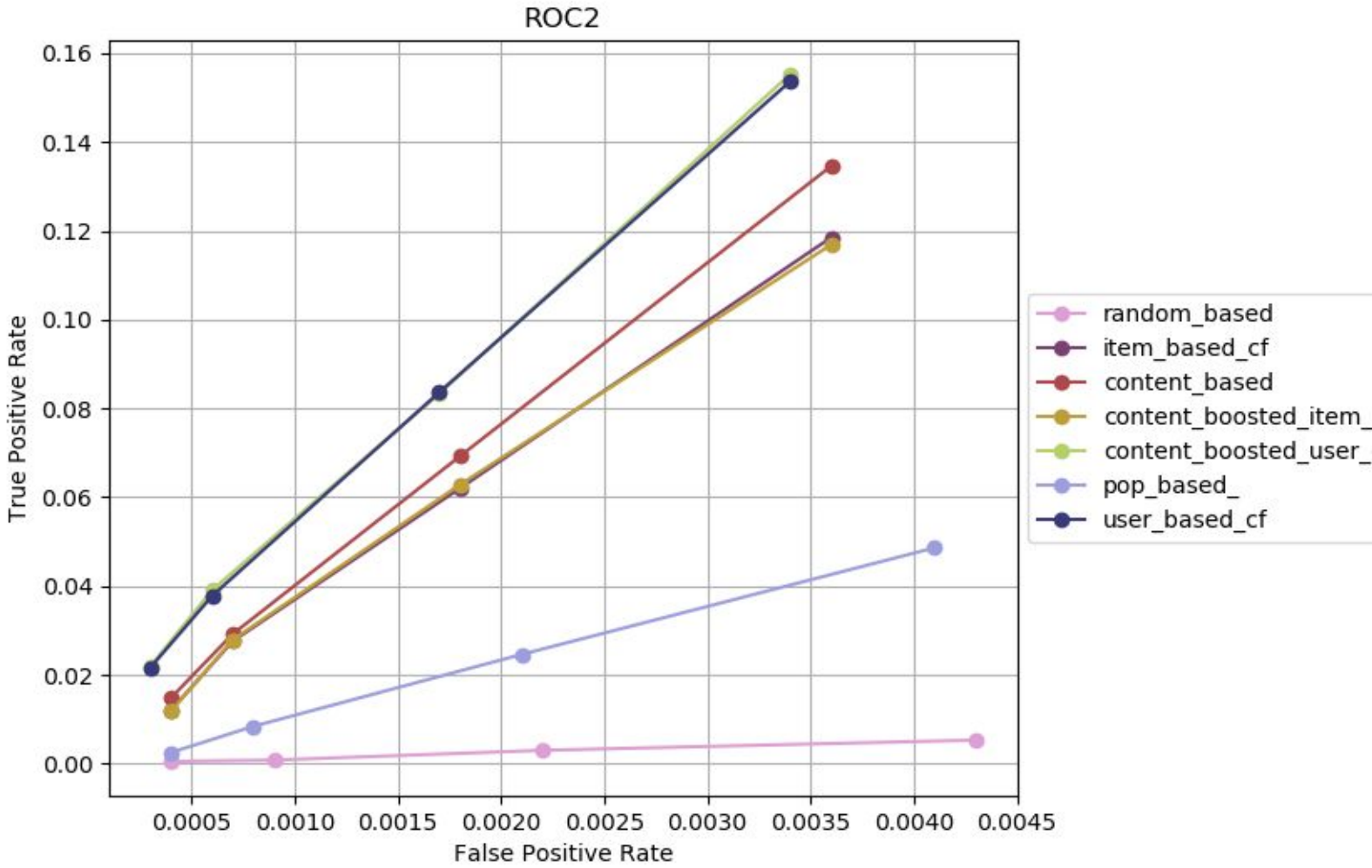
Results

Generic Recommenders

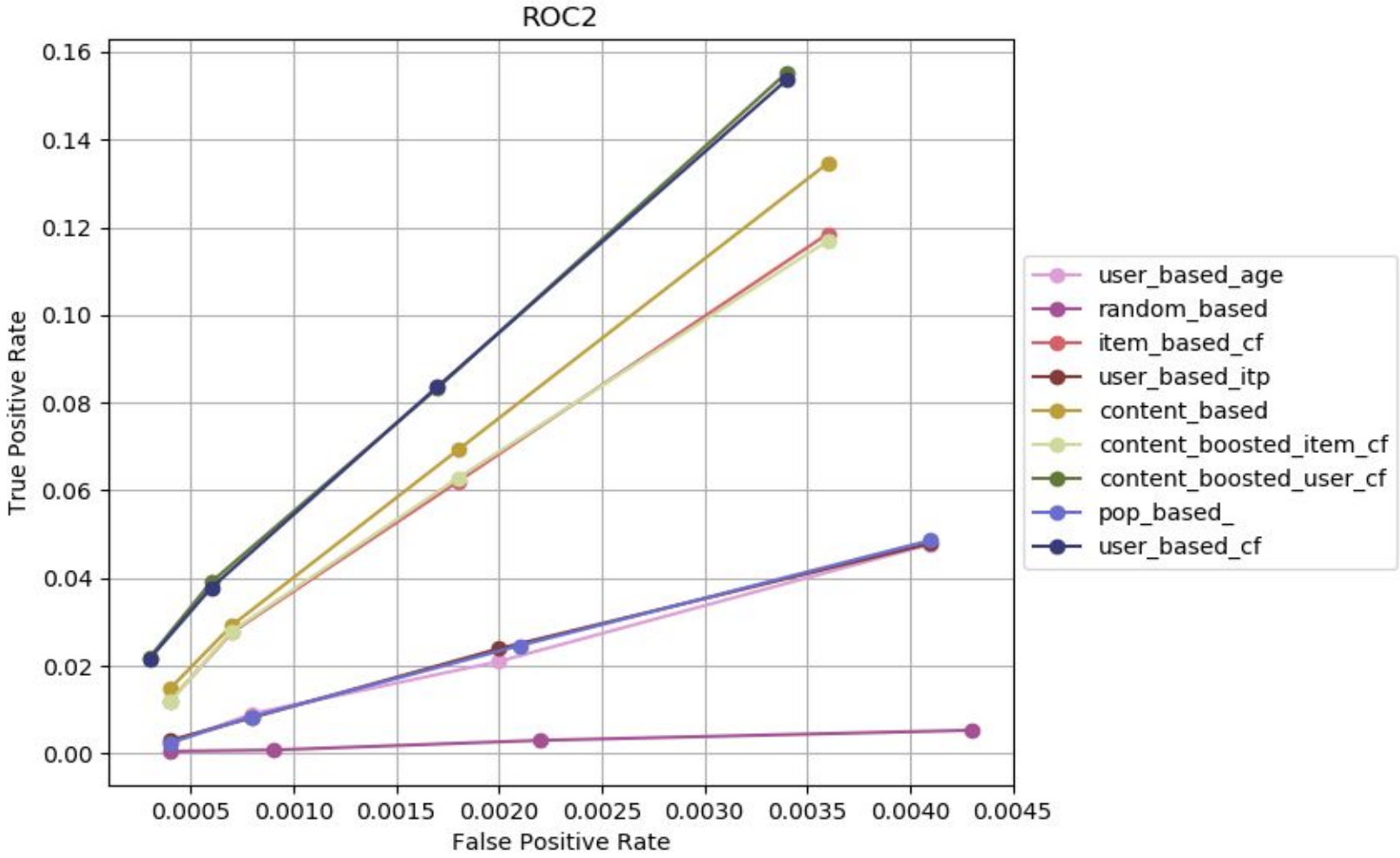


kidaptive

Content Boosted CF Recommenders

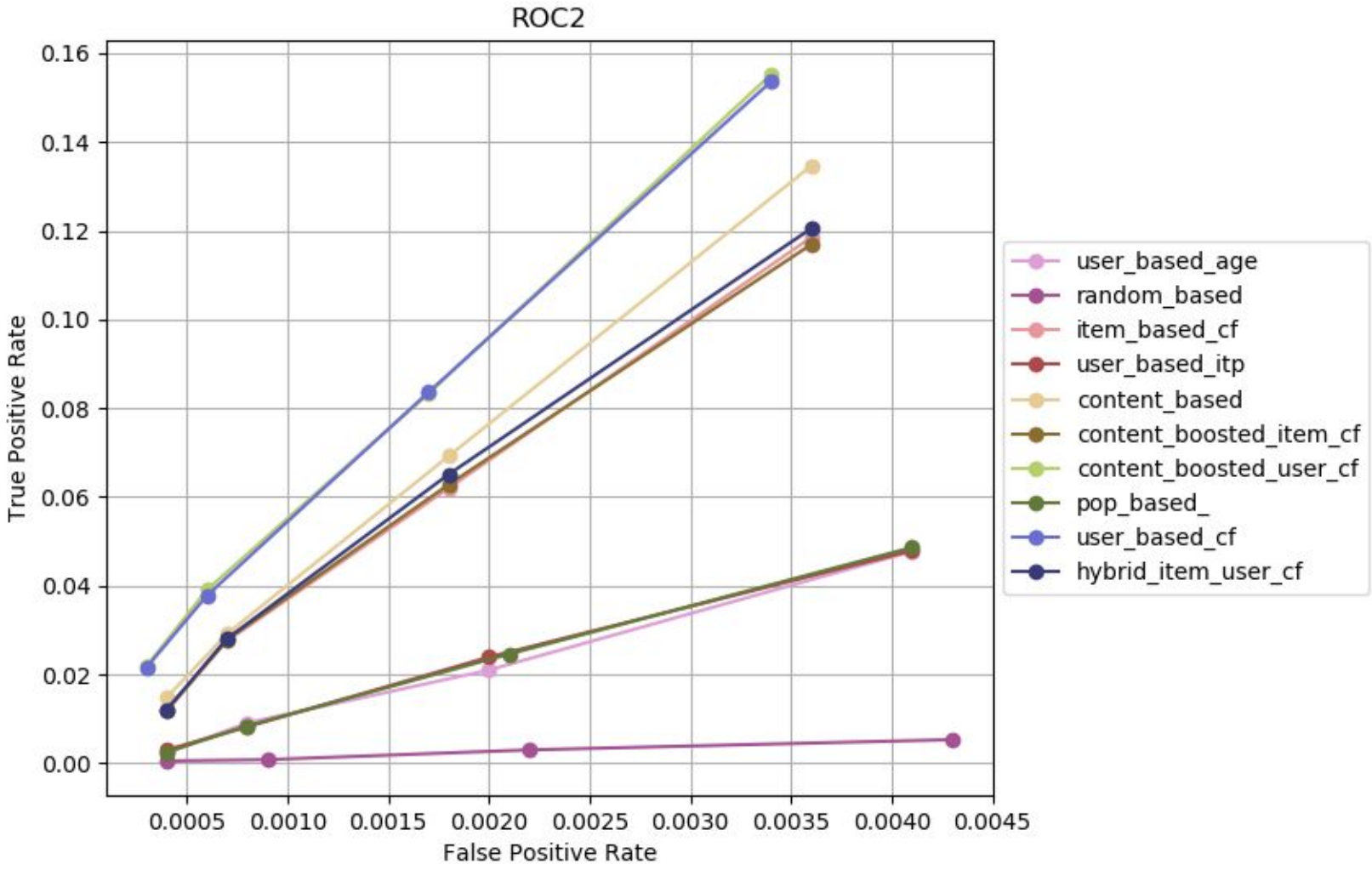


Customized User-Based Collaborative Filtering



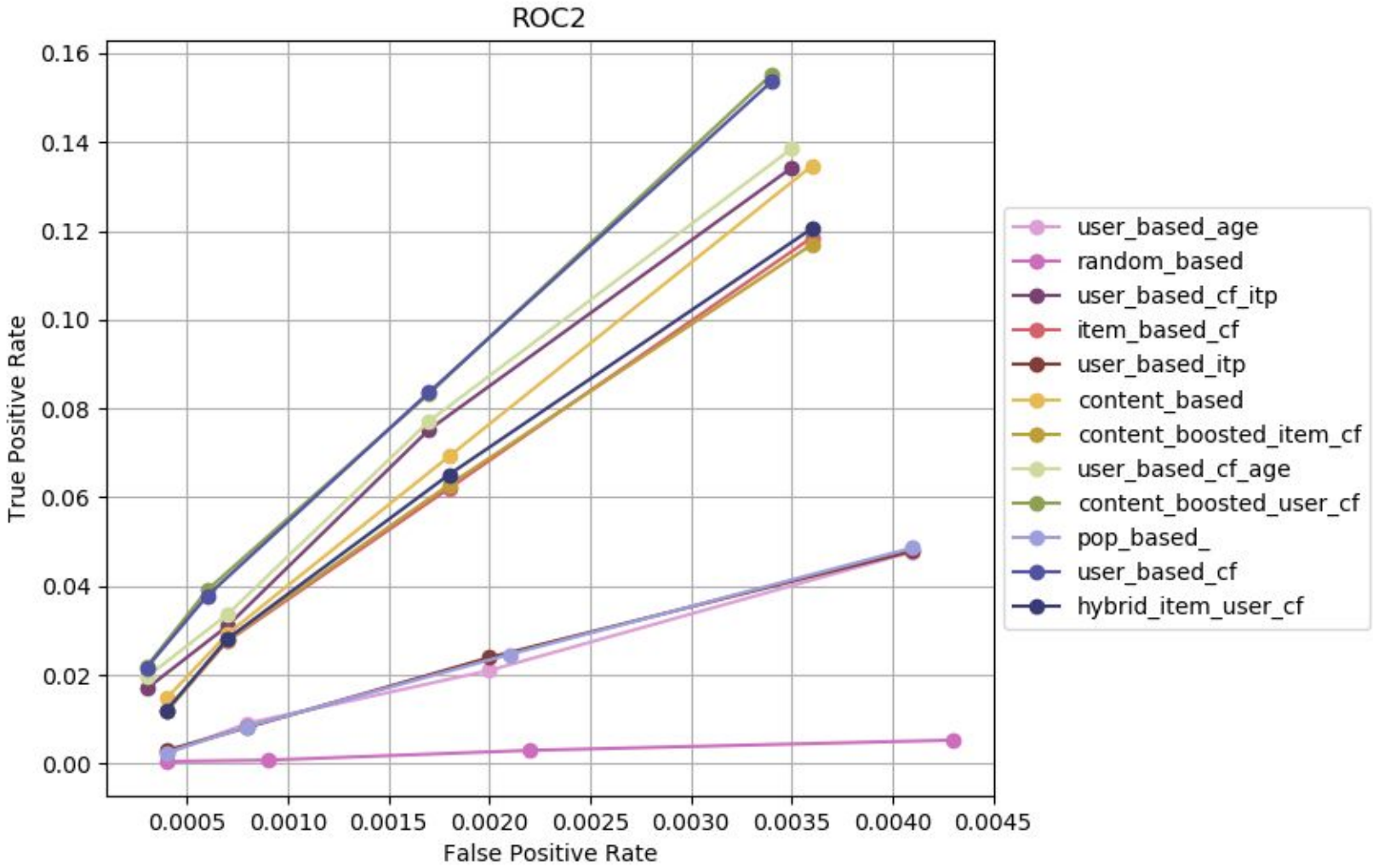
kidaptive

Generic Hybrid Recommender

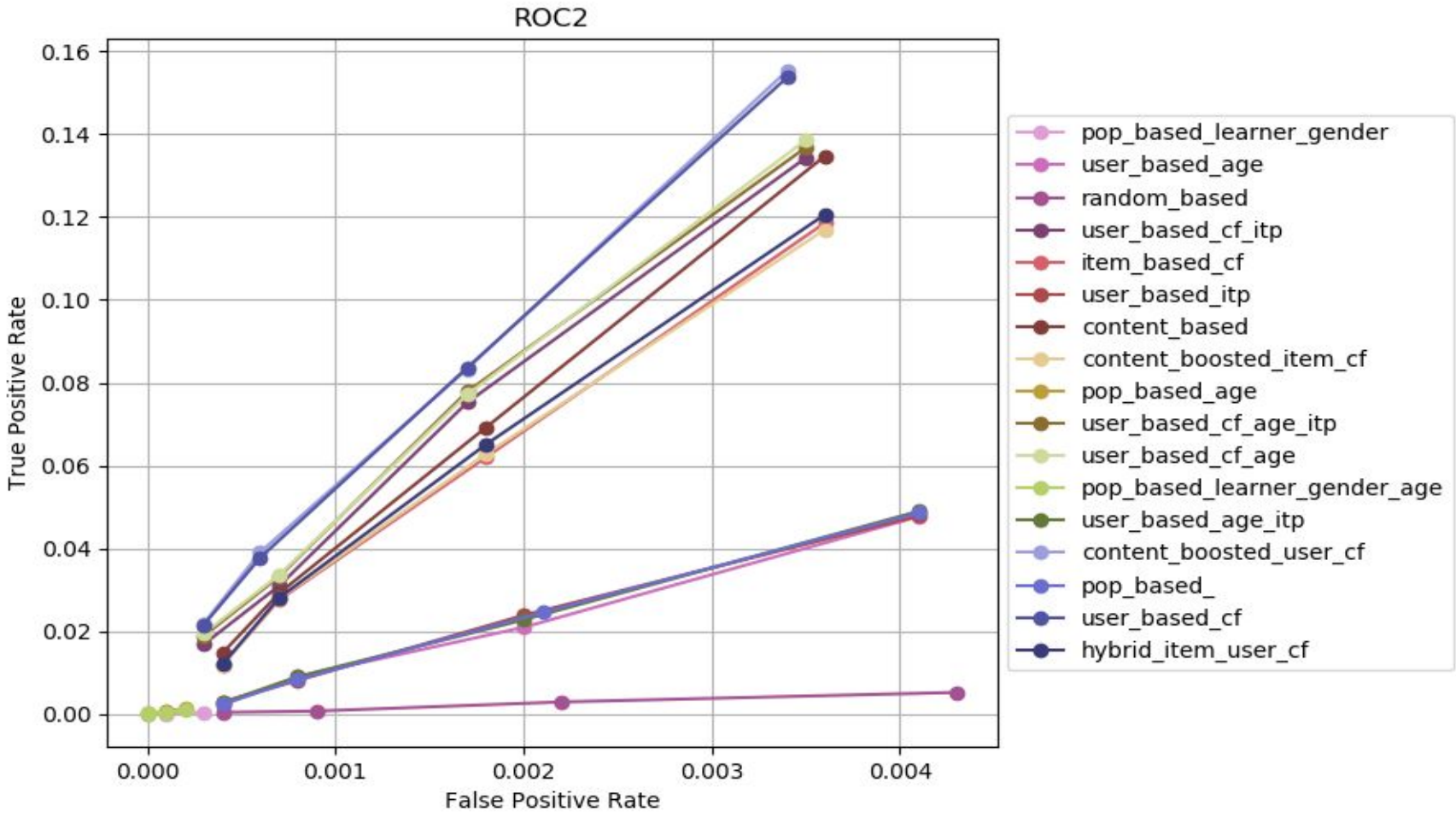


kidaptive

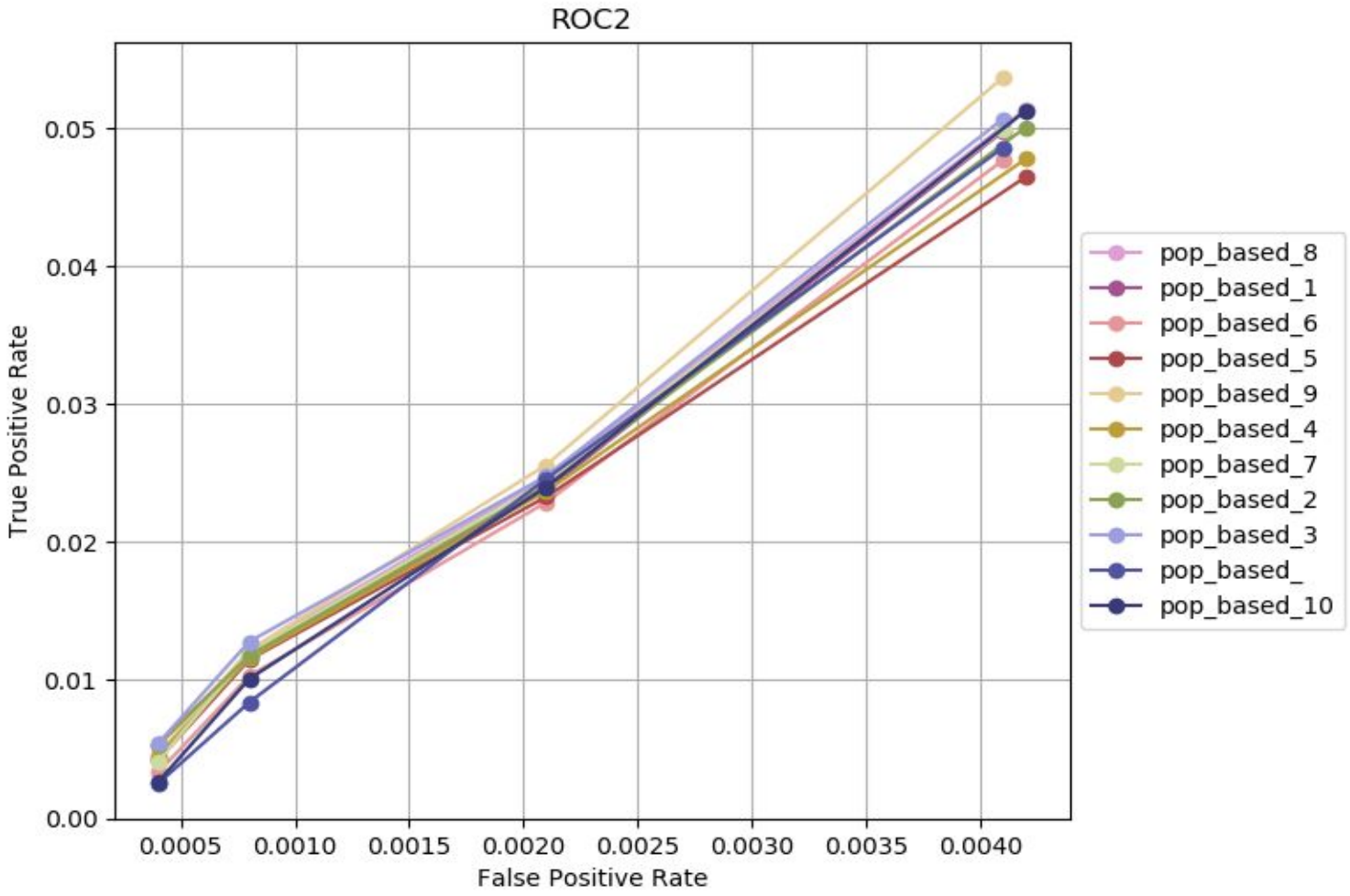
Custom Hybrid Recommender



All Recommender Models



Impact of Varying Number of Views



kidaptive

Future Enhancements

1. Use different recommender model as per no of views
2. Convert Binary->Non-Binary and leverage evaluation metrics from open source libraries
3. Recommend a sequence: Instead of focusing on the generation of a single recommendation, the idea is to recommend a sequence of items that is pleasing as a whole.

Using Additional Data from BookClub Platform

4. Searching habits can be leveraged along with reading habits
5. Time spent on each resource, useful to derive ratings
6. Similarity between free-form notes and content can be heuristic to gauge engagement with content, and hence preference/rating for resource

kidaptive

kidaptive

Personalized, Playful, Lifelong Learning