Affectv Data Scientist Test

Leonidas Kapsokalyvas

16 Jul 2014

Approach

- Set up a Linux VM (Centos 6.5).
- Load the dataset into a storage.
  - MongoDB
- Write analytics on top of the storage.
  - Python
- Create reports

Collaborative Filtering

- Can we predict future behaviour of a user ?
  - By establishing similarities between users.
  - Users with common preferences are likely to choose similar products in the future.
- Memory-based Algorithms (Breese et al. 98)
- Model-based Algorithms

  http://research.microsoft.com/pubs/69656/tr-98-12.pdf
- In our case
  - The campaign is equivalent to a product.
  - User activity on a campaign counts as rating that product.

Memory-based approach

- Aggregate users such that you have a summary of all campaigns per user:

```
{ "user_id" : "4fc382f297dc0938360f07ff",
"campaigns" :
{
"53207b1bc259087eb1195183" : { "retargeting" : 1 },
"5213718ec259080ec35e0e76" : { "impression" : 1 },
"532079a1c259087eb1195163" : { "retargeting" : 1 },
"51dac16cc25908069f801c70" :  { "impression" : 1 }
},
 "campaignCount" : 4
}
```

Frequency

User Similarity

- Take a user ***x*** under test.
- Leave one campaign ***j*** out from that user.
- Find all users with the same set of campaigns as ***x***.
- Predict the frequency for campaign ***j*** and activity **a** = retargeting as:

$$Freq(x, j_a) = \frac{1}{|U|} \sum_{u \in U} v_{u\, j_a}$$

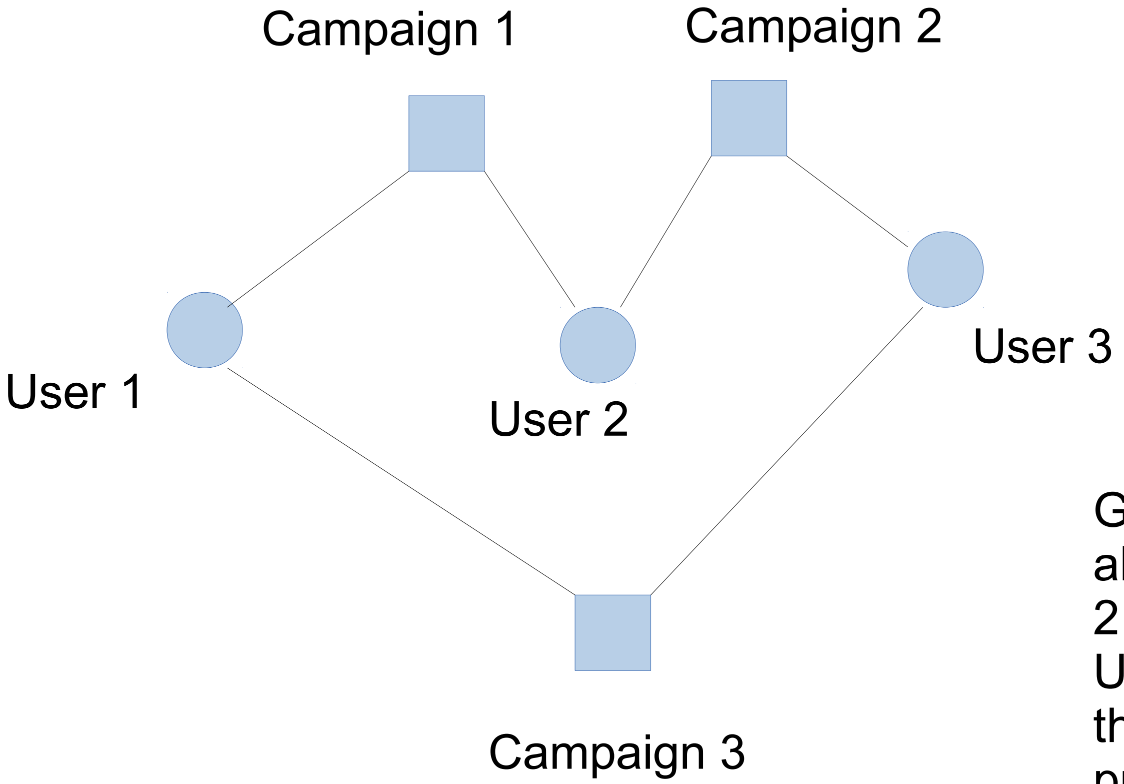
Results

- Users in the current dataset do not share a lot of common campaigns. So the similarity approach is not very useful.

Campaigns per user	Number of users
1	539,836
2	69,151
3	7,086
4	787
5	92
6	17
7	7
8	1
9	1
>=10	0

Results for 616,978 users

Graph based approach



Given the information about Campaign 1 and 2 for User 1, User 2 and User 3, can we predict that User 1 and User 3 prefer Campaign 3 ?

Do the three users form a triangle ?

Other similarity features

- Location proximity.
- Same device.
  - Both show similar preference in some products already (house and gadgets).
- Time information.
  - Can we assume that users who are active at the same time period (i.e. bank holiday weekend) and location) are likely to buy similar products ?

Summary of user behaviour

	Converted Users	Non Converted
Number of users	436	616542
Impression	42	230422
Click	0	97
Retargeting	143	768816
Conversion	480	0

Users who convert seem to be doing much less
browsing compared to those who do not convert.

Browsing Trends

- Take a particular campaign.
- Count the number of impressions, clicks, retargeting, conversions over a window of time.
- Move the window and keep counting.
- Plot the time series of counts.

- It may be of interest to:
  - Predict the time series of a particular campaign.
  - Express a time series campaign as a function of the rest campaigns.