Product Recommendation Using Teradata Aster

Pravin Dhuri

01th June 2017

Table of Contents

A. Document Information

Document Name	Product Recommendation Using Teradata Aster
Document Author	Pravin Dhuri
Document Objective	
Document Location	COMPASS
Document Version	Version 1
Release Date	01 th June, 2017

B. Document History

Version No	Date	Section No	Description of Change	Author
V1	01 th June, 2017		First Version	Pravin Dhuri

1. OVERVIEW

Recommendation System

Recommendation system is an information filtering technique, which provides users with information, which he/she may be interested in.

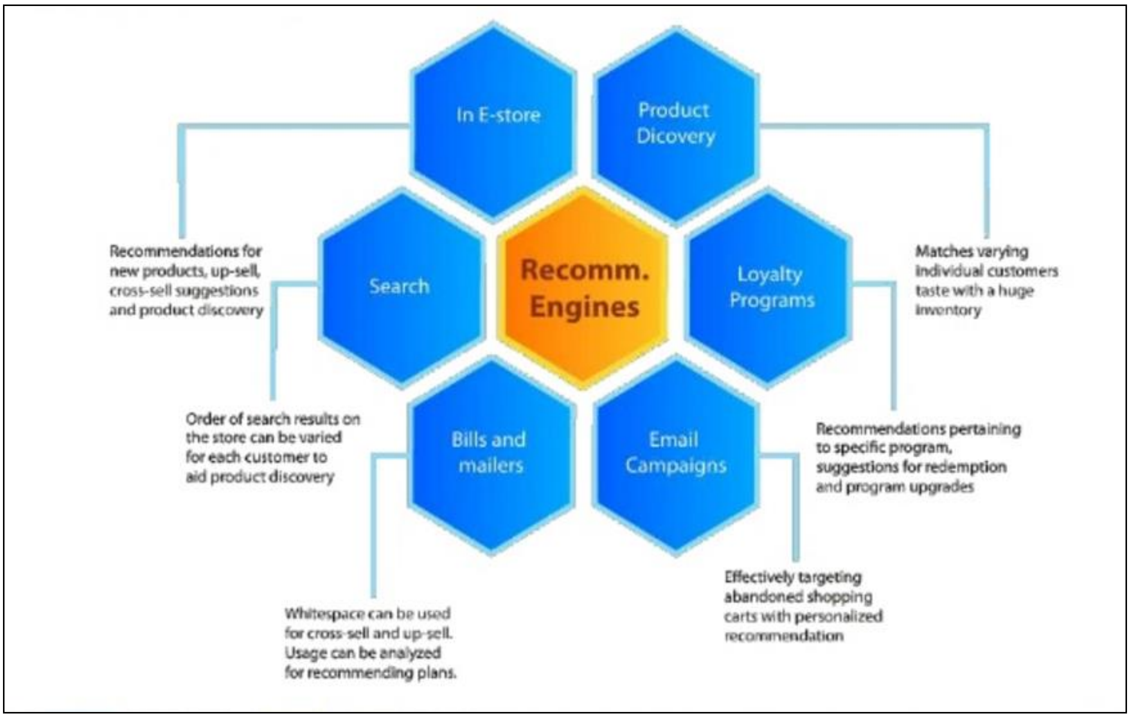


Fig: Area of Use

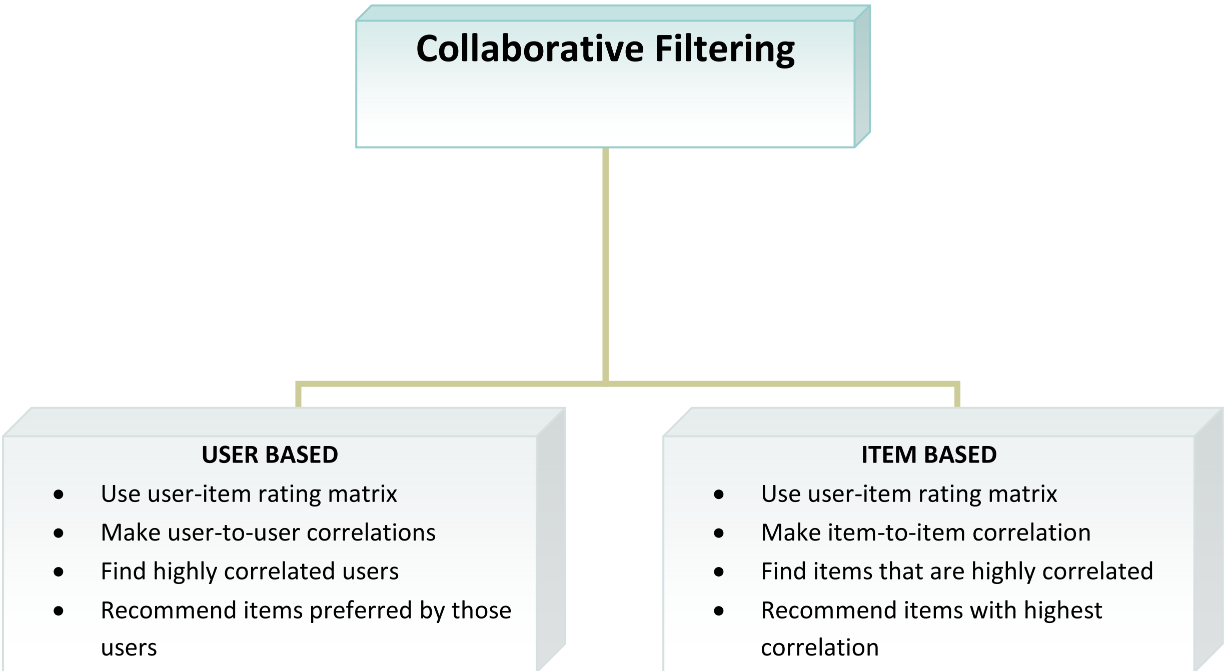
TECHNIQUES: RECOMMENDATION GENERATION

1. Collaborative Filtering
2. Content-Based Filtering

Collaborative Filtering

Collaborative filtering methods are based on collecting and analyzing a large amount of information on users' behaviors, activities or preferences and predicting what users will like based on their similarity to other users. A key advantage of the collaborative filtering approach is that it does not rely on machine analyzable content and therefore it is capable of accurately recommending complex items such as movies without requiring an "understanding" of the item itself.

Main Approaches



Content-Based Filtering

Content-based filtering approaches utilize a series of discrete characteristics of an item in order to recommend additional items with similar properties. Content-based filtering methods are based on a description of the item and a profile of the user's preference. In a content-based recommender system, keywords are used to describe the items and a user profile is built to indicate the type of item this user likes. In other words, these algorithms try to recommend items that are similar to those that a user liked in the past (or is examining in the present).

In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended. This approach has its roots in information retrieval and information filtering research.

2. DATA SOURCE

The dataset used in this case study can be downloaded from Kaggle:

https://www.kaggle.com/c/santander-product-recommendation/data

Loading Data into Aster DB:

```
 ncluster_loader -U db_superuser -w db_superuser -d beehive --skip-rows 1  -c santa.train
/home/Pravin/Santander/train_ver2.csv
```

Flags:

-U: Username

-w: Password

-d: Database Name

--skip-rows: Skipping Rows

-c: CSV file

3. DATA CLEANSING

Let's have a quick glance at the dataset

```
Select * from santa.train;
```

santa: Schema name
train: Table name

Output Table:

	fecha dato	custid	empindex	custres	sex	age	firstholder	new custi	custsenior	indrel	last date	customer	customer	residence	foreigner i	spouse in	channel	deceased
1	2016-05-28	439366	N	ES	V	79	2003-10-11	0	151	1	null	1.0	I	S	N	null	KAT	N
2	2016-05-28	439360	N	ES	V	41	2003-10-11	0	151	1	null	1.0	A	S	N	null	KFJ	N
3	2016-05-28	439358	N	ES	H	40	2003-10-11	0	151	1	null	1.0	A	S	N	null	KFJ	N
4	2016-05-28	439352	N	ES	V	61	2003-10-11	0	151	1	null	1.0	I	S	N	null	KFA	N
5	2016-05-28	439348	N	ES	H	68	2003-10-11	0	151	1	null	1.0	I	S	N	null	KFA	N
6	2016-05-28	439344	N	ES	V	69	2003-10-11	0	151	1	null	1.0	I	S	N	null	KAT	N
7	2016-05-28	439342	N	ES	H	64	2003-10-11	0	151	1	null	1.0	I	S	N	null	KAT	N
8	2016-05-28	439316	N	ES	H	65	2003-10-11	0	151	1	null	1.0	I	S	N	null	KAT	N
9	2016-05-28	439338	N	ES	V	49	2003-10-11	0	151	1	null	1.0	A	S	N	null	KAT	N
10	2016-05-28	439336	N	ES	V	48	2003-10-11	0	151	1	null	1.0	I	S	N	null	KAF	N

activity in	gross inc	segmenta	saving acc	guarantees	current ac	derivada	payroll ac	junior acc	más parti	particular	particular	short ter	medium t	long term	e account	funds	mortgage
0	222066.75	02 - PARTI...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	55704.93	02 - PARTI...	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	77098.95	02 - PARTI...	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
1	210628.92	02 - PARTI...	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
0	112368.18	02 - PARTI...	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	137477.22	02 - PARTI...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	137477.22	02 - PARTI...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	181563.03	02 - PARTI...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	73261.74	02 - PARTI...	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	null	02 - PARTI...	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0

Fig : Data Table

Checking whether Customer code has null vales or not

```
Select * from santa.train where custid is null;
```

UNPIVOTING THE TABLE:

Unpivot:

Unpivot is a SQL-MR function to convert columns into rows.

Query:

```
create table santa.train_unpivot_22 distribute by hash (ncodpers) as
SELECT * FROM Unpivot (
ON (select * from santa.train where Payroll='1')
colsToUnpivot('Payroll')
colsToAccumulate('fecha_dato','custid')
);
```

**Due to Concurrency/QoS constraint unpivoting has being done one by one on each product column and then the results are being merged

Output:

```
select * from santa.train_unpivot_final limit 20;
```

fecha_dato	custid	product
2016-05-28	17334	Pensions
2016-05-28	17318	Pensions
2016-05-28	17188	Pensions
2016-05-28	17236	Pensions
2016-05-28	17224	Pensions
2016-05-28	17214	Pensions
2016-05-28	17443	Pensions
2016-05-28	16846	Pensions
2016-05-28	16752	Pensions
2016-05-28	16728	Pensions
2016-05-28	16724	Pensions
2016-05-28	17065	Pensions

4. DATA EXPLORATION

BEHAVIOR SEGMENTATION

AppCenter Logic:

```
INSERT INTO app_center_visualizations  (json)
SELECT json FROM Visualizer (
ON "wordco" PARTITION BY 1
ColumnMap('numericValue1=cnt','label=token')
AsterFunction('custom')
Title('Word Cloud')
VizType('wordcloud')
);
```

WordCloud Chart:

WordCloud visualization displays the words in such a way that the font size represents the relative value of the word within the range of the given values.

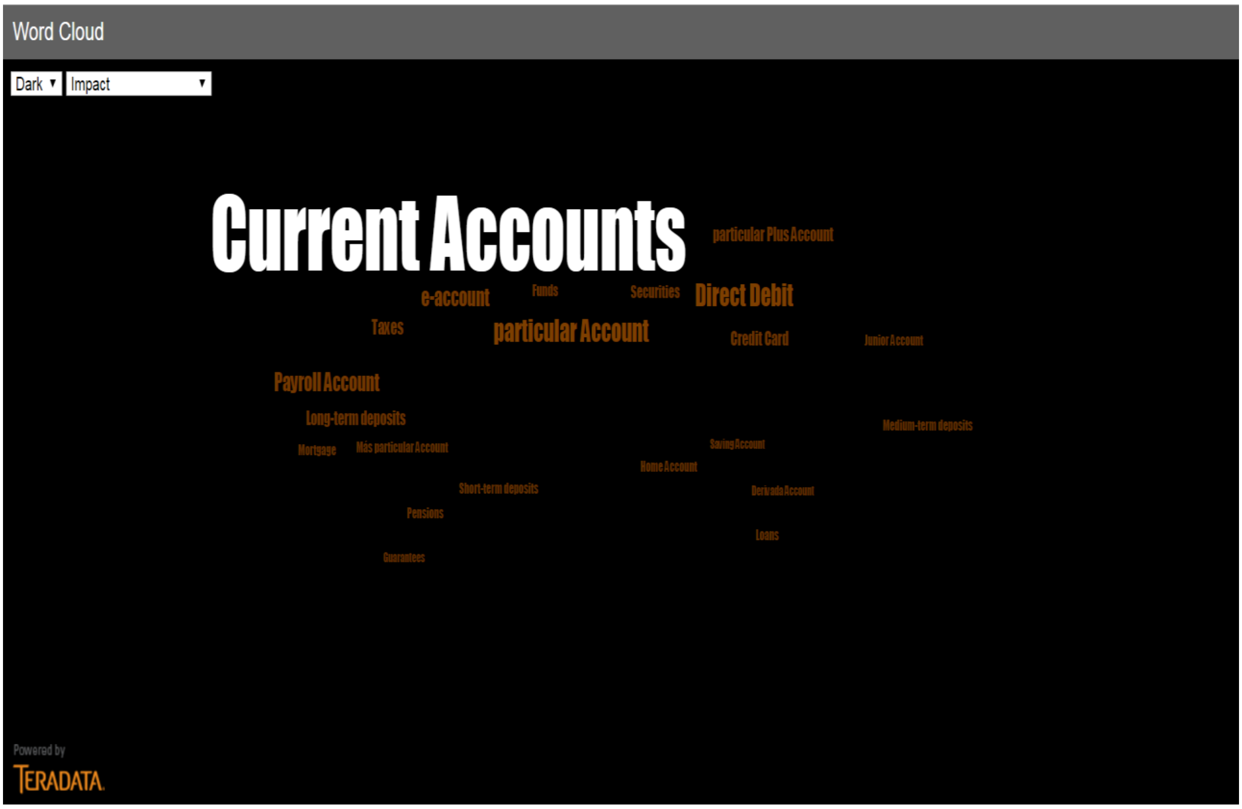


Fig:WordCloud Chart

RECOMMENDATION ALGORITHM

Algorithm CFilter (Association analysis**)**

Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users.

For example, an online store that tells a shopper, “Other shoppers who bought this item also bought these items”

Resultant:

At a customer level: Cust1-Product1 -> Cust1-Product5

Historical data duration can change after looking at initial analysis i.e. depending on the number of times product changes are seen in the recent past.

Query:

```
select *
from cfilter(
on (select 1)
PARTITION BY 1
database('beehive')
INPUTTABLE('santa.train_unpivot_final')
OUTPUTTABLE('santa.train_cfilter_test')
INPUTCOLUMNS('product')
JOINCOLUMNS('custid')
OTHERCOLUMNS('fecha_dato')
DROPTABLE('true')
);
```

Output:

fecha_dato	col1_item1	col1_item2	cntb	cnt1	cnt2	score	support	confidence	lift	z_score
2015-11-28	Direct Debit	Derivada Account	13	104990	319	5.0460097...	1.9120121...	1.2382131...	0.2639109...	-0.456490...
2015-07-28	Direct Debit	Loans	31	99981	2024	4.7489260...	4.9568751...	3.1005891...	0.0958048...	-0.433849...
2015-07-28	Credit Card	Medium-term deposits	22	37159	1396	9.3303068...	3.5177823...	5.9205037...	0.2652326...	-0.446161...
2016-04-28	Pensions	Loans	6	7365	1987	2.4599817...	8.6340004...	8.1466395...	0.2849179...	-0.416065...
2015-03-28	Medium-t...	Pensions	3	1560	7384	7.8131510...	4.9015683...	0.0019230...	0.1594010...	-0.477215...
2015-06-28	Taxes	Pensions	125	42888	7363	4.9479970...	2.0293031...	0.0029145...	0.2438273...	-0.322706...
2016-01-28	Derivada ...	Taxes	2	323	43829	2.8255038...	2.9080757...	0.0061919...	0.0971606...	-0.449730...
2016-04-28	Loans	Credit Card	20	1987	34807	5.7835638...	2.8780001...	0.0100654...	0.2009577...	-0.403483...
2016-04-28	Medium-t...	Credit Card	17	1035	34807	8.0221522...	2.4463001...	0.0164251...	0.3279300...	-0.406179...
2016-04-28	Credit Card	Taxes	1217	34807	45275	9.3984471...	0.0017512...	0.0349642...	0.5366667...	0.6722521...

The output table contains these columns:

- Col1_item1: Name of item1.
- Col2_item2: Name of item2.
- Cntb: Count of the co-occurrence of both items (situations when people buy the two items together).
- Cnt1: Count of the occurrence of item1 within the partition formed by the 'OTHERCOLUMNS' argument.
- Cnt2: Count of the occurrence of item2 within the partition formed by the 'OTHERCOLUMNS' argument.
- Score: The product of two conditional probabilities.
- Lift: The ratio of the observed support value to the expected support value if item1 and item2 were independent.
    - Lift > 1 - The occurrence of item1 or item2 has a positive effect on the occurrence of the other items.
    - Lift < 1 - The occurrence of item1 or item2 has a negative effect on the occurrence of the other items.
    - Lift = 1 - The occurrence of item1 or item2 has a no effect on the occurrence of the other items.
- Z_score: It is a way to measure how significant the co-occurrence is.
- Support: The percentage, among all the transactions, that the two items co-occur.
- Confidence: The percentage of item2 occurrence in all the transactions in which item1 occurs.

VISUALIZATION

Output of Association analysis i.e. Cfilter with assigned probability would be visualized using **Aster Appcenter**.

VISUALIZATION TYPE: SIGMA

The Sigma visualization is appropriate for depicting data networks, intuitively portraying items and how they relate to each other. AppCenter's Sigma graph provides options to filter, search, navigate, and customize the layout of the data displayed.

AppCenter Logic:

```
INSERT INTO app_center_visualizations  (json)
SELECT json FROM Visualizer (
ON santa.train_cfilter_test  PARTITION BY 1
AsterFunction('cfilter')
Title('My Viz')
VizType('sigma')
);
```

Sigma Chart:

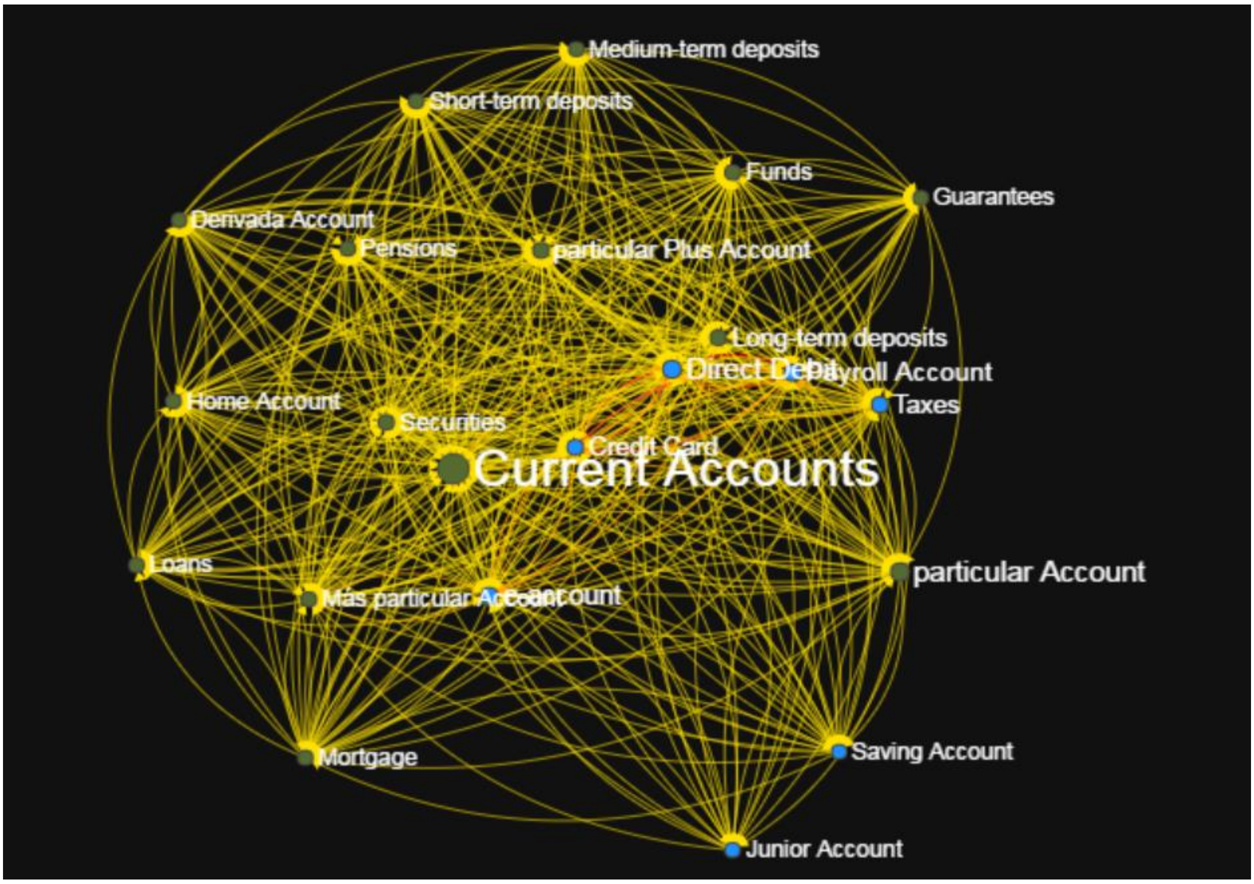


Fig : Sigma Chart

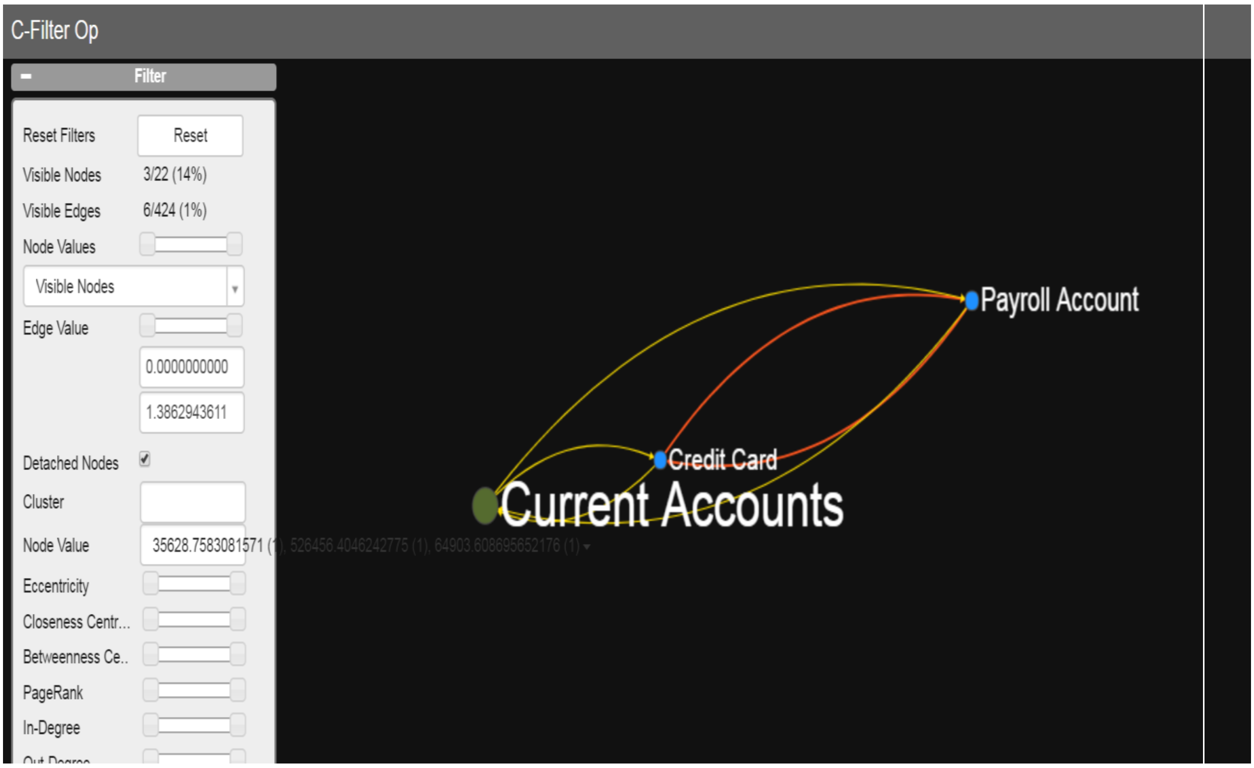


Fig: Filtered Sigma Chart

- The direction of the arrows indicates what customer will buy in addition to what they already have.
- The width and color of the arrow (edge) indicates the relative value of the edge.
- The color range goes from yellow to red. The higher the value of the edge, the wider and closer to red in color it will be.

My App Link:

https://192.168.100.100:444/appserver/portal/#app/15

C-FILTER RECOMMENDER

Query:

```
SELECT * FROM cfilterRecommender(
              ON (select 1)
              PARTITION BY 1
              transaction_table('santa.train_unpivot_final_limiteddata')
              cfilter_table('santa.train_cfilter_test_limiteddata')
              recommendation_table('recommendation')
              purchased_item_column('product')
              user_column('custid')
              userid('db_superuser')
              password('db_superuser')
              database('beehive')
              drop_table('true')  );
```

Final Output:

```
select * from recommendation;

(select * from recommendation where custid='15892')
```

	custid	col1_item2	purchase_probability
1	15892	Direct Debit	0.0636758034369441
2	15892	Current Accounts	0.0697717273627715
3	15892	Payroll Account	0.0650496395089029
4	15894	Direct Debit	0.0636758034369441
5	15894	Current Accounts	0.0697717273627715
6	15894	Payroll Account	0.0650496395089029
7	15900	Direct Debit	0.0636758034369441
8	15900	Current Accounts	0.0697717273627715
9	15900	Payroll Account	0.0650496395089029
10	15920	Direct Debit	0.0636758034369441
11	15920	Current Accounts	0.0697717273627715
12	15920	Payroll Account	0.0650496395089029
13	15944	Direct Debit	0.0636758034369441
14	15944	Current Accounts	0.0697717273627715
15	15944	Payroll Account	0.0650496395089029
16	16002	Direct Debit	0.0636758034369441
17	16002	Current Accounts	0.0697717273627715
18	16002	Payroll Account	0.0650496395089029
19	16022	Direct Debit	0.0636758034369441
20	16022	Current Accounts	0.0697717273627715

Eg: Customer with custid=15892 will purchase or can be recommended Direct Debt entity having probability of 0.063.