MIDDLE EAST TECHNICAL UNIVERSITY

Department Of Computer Engineering

CENG 300

Summer Practice Report

METU Data Mining Research Group
Start Date: End Date:
Total Working Dates:

October 9, 2020

Student:
Onat ÖZDEMİR

Supervisors:
Prof.Dr. Pınar KARAGÖZ

Prof.Dr. İsmail Hakkı TOROSLU

Student's Signature

Organization Approval

Contents

1 Introduction

I have done my summer internship at METU Data Mining Research Group under the supervision of Prof.Dr. Pınar KARAGÖZ and Prof.Dr. İsmail Hakkı TOROSLU. The main task I have worked on was describing trust between customers and integrating the trust scores calculated by TACoRec[7] to collaborative filtering algorithm. In addition, towards to end of my internship, I implemented an additional Trust Based Recommender based on a new trust metric.

2 Project

During the internship, I implemented two trust based recommenders with different weighting methods:

1. Eigentrust Weighted Recommender
2. Inverse Distance Weighted Recommender

The details of these two recommenders can be found in section 2.4 and 2.5, respectively.

2.1 Analysis Phase

The variety and number of products are increasing day by day, which creates the problem of recommending the most appropriate products for users. One of the main approaches used in design of recommender systems is Collaborative Filtering. The approach uses prior behaviours of customers such as rating profiles, product preferences, etc. to generate recommendations. Collaborative Filtering methods can be classified according to which factor they prioritize while making suggestions. In this project we focused on Trust Based Collaborative Filtering which generate recommendations considering the trust between users.

The definition and calculation of trust may differ in many sources and researches. For instance, [10] approaches the issue from the probabilistic aspect and calculates trust through successful and unsuccessful transactions, the trust metric used in [14] is based on "Pearson Correlation Similarity"

between users while [11] stress the value of providing ratings and argued that users who give more rates are more trustworthy, even if they don't rate similarly. Such recommendation systems aim to calculate trust scores from behaviours of customers (e.g. ratings) to make good recommendations in the absence of already existing trust network.

On the other hand, there are also approaches[13] that are developed to operate on datasets containing explicit trust scores (e.g. Epinions [2]) and make suggestions by using these scores directly or by combining them with additional features such as similarity, product or user attributes (especially in hybrid recommenders), etc.

During the internship, due to lack of explicit trust information in most systems, we mainly focused on developing accurate Trust Based Recommender Systems working on datasets with no explicit trust information such as Stockmount, Amazon Food Review, etc. Some of these datasets (e.g. Stockmount) contain implicit ratings while rest of them have explicit ratings given by customers.

For the "Eigentrust Weighted Recommender", we were able to use "community detection" and "eigentrust calculation" modules provided by TACoRec. The part I worked on was integrating the "eigentrust scores" to the collaborative filtering algorithm.

For the "Inverse Distance Weighted Recommender", I experienced the difficulty of inferring trust from implicit customer behaviours which was I believe the most meaningful and challenging part of the internship.

2.2 Design Phase

2.2.1 For Eigentrust Weighted Recommender

As a as preliminary information, the Neo4j database we tested the recommender on initially contained customer, product and transaction records. After using "community detection" and "eigentrust" modules provided by TaCoRec, disjoint customer communities and Eigentrust scores between the customers belong to the same community were added to database.

Eigentrust represents how strongly connected the customers are to their communities (for the detailed information, please check section 2.4.1) and stored as a property of the relationship between the customer and his/her community in the Neo4j database. We can draw two conclusions from this information;

1. Since the Eigentrust scores only calculated between the members of a community, when making a recommendation to a customer, only the preferences of people in the same community as that customer should be taken into account.

2. Although the opinions of stereotype customers (i.e. customers with higher Eigentrust score) are important, we should also consider the similarity between the target user and the users who make suggestions to keep recommendations personalized.

Based on these conclusions, I designed two modules: Recommender and Filterer. The Recommender is responsible for writing the recommendations generated by the Filterer module to the database while the remaining weight of the project is carried by the Filterer: getting transaction and Eigentrust records from the database via Neo4j driver, measuring similarities between users based on transaction records, calculating recommendation coefficients for each product and generating recommendations based on these coefficients.

For the detailed information and the recommender structure, please check section 2.4 and Figure 1.

2.2.2 For Inverse Distance Weighted Recommender

"Inverse Distance Weighting" method is inspired by an article[12] that suggests calculating the trust scores between two customers by the reciprocal of the shortest distance between them on a trust network. Since I worked on datasets with no explicit trust information, I decided to use transaction records to build the trust network.

So with this approach, In addition to Filterer and Recommender, I implemented the Graph Module which is responsible for building the graph according to preferred method (unweighted or euclidean distance weighted)

from transaction records and calculating trust scores using the shortest distances between customers in the graph.

For the detailed information and the recommender structure, please check section 2.5 and Figure 3.

2.3 Implementation Phase

Since there are two different implementations, I have divided the implementational details of the recommenders into two subsections: section 2.4 and 2.5. Under this subsection, libraries and technologies used in implementations are explained.

2.3.1 Neo4j [3]

Neo4j is a graph database management system where data is organized as nodes, relationships, and properties. The communication between recommender and the database is maintained by Neo4j driver.

Driver Installation :

```
pip install neo4j

```

Configuration :

```
import neo4j
...

uri = self._config["database"]["neo4j"]["uri"]
user = self._config["database"]["neo4j"]["user"]
password = self._config["database"]["neo4j"]["password"]

self._driver = neo4j.Driver(uri, auth=(user, password))


```

Sample Usage :

```
import neo4j
...

def get_customer_trust(self, customer_id):

  query = (
    f"MATCH (u:Customer)-[r:BELONGS_IN]->(:Community) "
    f"WHERE u.id = {repr(customer_id)} "
    f"RETURN r.eigentrust"
  )

  with self._driver.session() as session:
    return tuple(session.run(query).single())


```

Listing 1: Neo4j driver example

2.3.2 Numpy [4]

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices. Since the core of Numpy is optimized C code, performing the calculations in the recommendation process using Numpy matrix provides serious time savings.

Installation :

```
pip install numpy

```

Sample Usage :

```
import numpy as np

class TrustBasedFilterer(object):
...

  def _create_customers_versus_products_table(self):

    self._customers_versus_products_table = np.zeros(
      (self._unique_customers.shape[0],
      self._unique_products.shape[0]),
      dtype=np.bool,
      )
```

```
13
14      self._customers_versus_products_table[
15        self._sales[:, 0],
16        self._sales[:, 1],
17      ] = True
18
```

Listing 2: Numpy example

2.3.3 Scipy [5]

SciPy is a Python library used for scientific computing. Similar to Numpy, many of the Scipy functions are written in C which provides a solution to the slowness caused by interpretation. For this reason, we prefer to use Dijkstra algorithm provided by Scipy rather than implementing by ourselves.

Installation :

```
1  pip install scipy
2
```

Sample Usage :

```
1  from scipy.sparse import csr_matrix
2  from scipy.sparse.csgraph import dijkstra
3
4  class Graph(object):
5  ...
6
7    def _create_distance_matrix(self):
8
9      self._create_adjacency_matrix()
10
11     self._adjacency_matrix = \
12       csr_matrix(self._adjacency_matrix)
13
14     self._distance_matrix = dijkstra(
15       csgraph=self._adjacency_matrix,
16       directed=False,
17       return_predecessors=False,
18       unweighted=True,
19       limit=self._max_distance)
20
21     self._distance_matrix\
```

```
22          [~np.isfinite(self._distance_matrix)] = 0
23
```

Listing 3: Scipy example

2.4 Eigentrust Weighted Trust Based Recommender

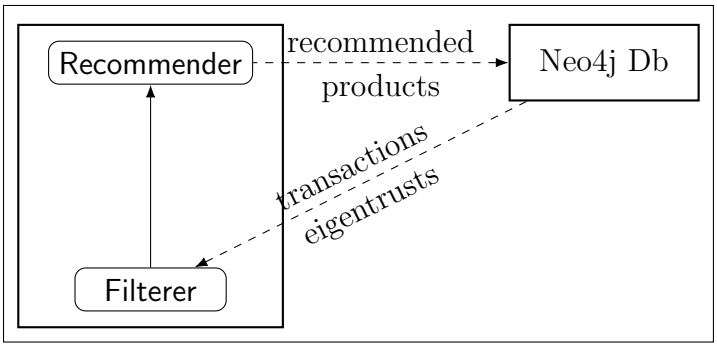

Figure 1: Recommender Structure

2.4.1 About Eigentrust

Eigentrust[10] is a reputation calculation algorithm based on the number of positive and negative transactions between customers and mainly designed for peer-to-peer networks. In our case, Eigentrust represents how strongly connected the customers are to their communities. Eigentrust values calculated by the eigentrust module provided by TACoRec[7] and stored in Neo4j database as a property of the relationship between a customer and his/her community.

Problem encountered with Eigentrust: Especially for the customers belong to communities with small size and low densities, Eigentrust scores stored in the database were either very small or equal to zero (check Figure 2). Most of the customers with zero Eigentrust were eliminated after filtering the network from customers with a small number of products. Unfortunately, even the filtering did not significantly improve the data.

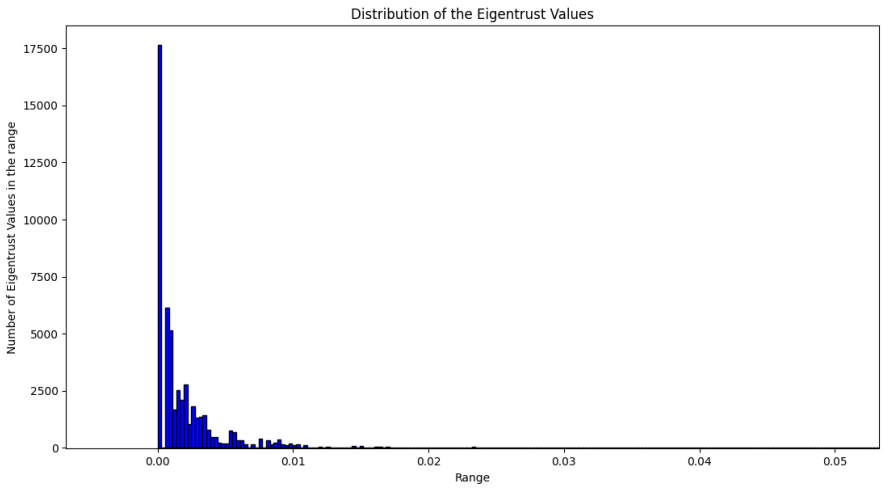

Figure 2: Distribution of the Eigentrust scores before filtering. As can be seen, nearly 17500 of the customers have zero Eigentrust.

2.4.2 Filterer Module

The Filterer Module is responsible for;

- Getting transaction/Eigentrust records for each community from the Neo4j database via Neo4j driver
- Calculating cosine similarities between customers in the same community
- Calculating weights between customers in the same community
- If the dataset consists of implicit ratings calculating recommendation coefficients otherwise making predictions for products
- Selecting k-products with the highest coefficients/predictions to recommend for each customer

2.4.2.1 Calculating Weights

$$w_{c_{target}}(c_2) = \frac{2 * sim(c_{target}, c_2) * trust(c_2)}{sim(c_{target}, c_2) + trust(c_2)}$$

where $sim(c_{target}, c_2)$ represents "cosine similarity" between customers and $trust(c_2)$ represents eigentrust belonged to c_2.

2.4.2.2 Calculating Recommendation Coefficients

$$RC(i) = \frac{\sum_{c \in C} w_{c_{target}}(c) * b_c}{\sum_{c \in C} w_{c_{target}}(c)}$$

where $RC(i)$ represents the recommendation coefficient calculated for product i and b_c is a boolean value which indicates whether the product was bought by person c or not.

2.4.2.3 Making Predictions

$$p(i) = \frac{\sum_{c \in C} w_{c_{target}}(c) * r_c}{\sum_{c \in C} w_{c_{target}}(c)}$$

where $p(i)$ represents the prediction for product i and r_c represents rating given by customer c for product i. C customer set only contains the customers who purchased product i.

Important remark: C community set used in above functions consists only of customers belonged to the same community with c_{target} while there is no such restriction in 2.5.

2.4.3 Recommender Module

Recommender Module is responsible for writing. The module has two tasks:

1. Getting the recommendation list that contains ids of the customers and corresponding recommended products from the Filterer module and writing these recommendations to the Neo4j database as a relationship between the customer and the recommended product using Neo4j driver.

2.5 Inverse Distance Weighted Trust Based Recommender

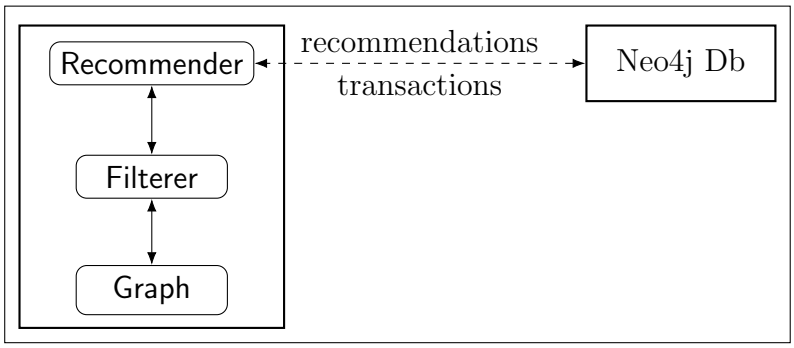

Figure 3: Recommender Structure

Inverse Distance Weighted Trust Based Recommender consists of three modules:

2.5.1 Graph Module

Graph Module is responsible for three tasks:

1. Constructing "adjacency matrix" from "customer versus product table" provided by Filterer module
2. Constructing "distance matrix" using "adjacency matrix"
3. Constructing "trust matrix" using "distance matrix"

2.5.1.1 Constructing Adjacency Matrix and Distance Matrix

Since the recommender is tested in both the datasets with implicit ratings and explicit ratings, to construct the "adjacency matrix" from customer versus products table, I propose two methods:

Proposed Method 1: Unweighted Graph

In this method, the "adjacency matrix" is constructed based on whether customers have a common product or not. In other words, edge between two customers can exist if and only if the intersection of the set of products they

purchased is not the empty set. This method is proposed for especially the datasets with **implicit ratings**.

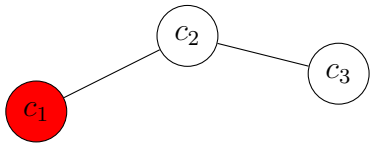

Figure 4: c_1 and c_2 have at least one common product while c_1 and c_3 do not have a common product

Problem encountered with Proposed Method 1: During the tests applied on the Movielens100k dataset, I observed that although the dataset is sparse, the maximum distance between two users was calculated as 2. Since the distances were distributed in such small range, trust values calculated with this method were not so meaningful. With this observation, I decided to propose another method and use the "Unweighted Graph" method in extremely sparse datasets.

Proposed Method 2: Euclidean Distance Weighted Graph
In this method, the "adjacency matrix" is constructed based on the "euclidean distances"1 between customers. This method is proposed for especially the datasets with **explicit ratings**.

$$adj[c_1]][c_2] = \frac{\sqrt{\sum_{i \in I_1 \cap I_2} (r1_i - r2_i)^2}}{|I_1 \cap I_2|} \tag{1}$$

where $r1_i$ and $r2_i$ represents ratings given by c_1 and c_2 for product i. Unlike the commonly used "euclidean distance" calculation, in this method, only ratings given to common products are included in the calculation.

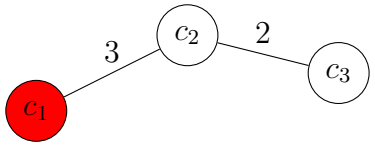

Figure 5: euclidean distance between c_1 and c_2 equals to 3, and c_1 and c_3 do not have a common product

Dijkstra's Algorithm
To construct the "distance matrix", the graph module uses "Dijkstra's Algorithm"[8] which takes the adjacency matrix as a parameter and returns the distance matrix.

2.5.1.2 Trust Calculation

After calculating the shortest distance between each pair of customers using "Dijkstra's Algorithm", to calculate the trust scores between customers Graph module uses

$$T(c_1, c_2) = \begin{cases} \frac{1}{d(c_1,c_2)} & d(c_1, c_2) \neq np.inf \\ 0 & d(c_1, c_2) = np.inf \end{cases}$$

function where $d(c_1, c_2)$ represents the shortest distance between the $customer_1$ and $customer_2$. If $d(c_1, c_2)$ equals $np.inf$ that means either there is no path connecting the customers or the shortest distance between the customers exceeds the distance limit specified in the config file.

A benefit of the method: Especially for excessively sparse datasets, recommenders using euclidean distance-based similarity fails since they cannot calculate a similarity score for the customer pairs with no common products. Since the "Dijkstra's Algorithm" propagates weights even for the customer pairs with no common products, we are able to calculate trust scores between the customers.

2.5.2 Filterer Module

The Filterer Module is responsible for;

- Cleaning/Filtering the provided transaction list (optional)
- Calculating cosine similarities between customers (optional)
- Calculating weights between customers
- If the dataset consists of implicit ratings calculating recommendation coefficients otherwise making predictions for products
- Selecting k-products with the highest coefficients/predictions to recommend for each customer

2.5.2.1 Calculating Weights

If 2.5.1.1 being used,

$$w(c_1, c_2) = \alpha * sim(c_1, c_2) + (1 - \alpha) * trust(c_1, c_2)$$

where $sim(c_1, c_2)$ represents "cosine similarity" between customers, $trust(c_1, c_2)$ represents trust calculated by the Graph Module and α is a weight ratio that changes according to the dataset .

Otherwise, $w(c_1, c_2)$ directly equals to $trust(c_1, c_2)$ since 2.5.1.1 method is already based on similarity.

2.5.2.2 Calculating Recommendation Coefficients

$$RC(i) = \frac{\sum_{c \in C} w(c_{target}, c) * b_c}{\sum_{c \in C} w(c_{target}, c)}$$

where $RC(i)$ represents the recommendation coefficient calculated for product i and b_c is a boolean value which indicates whether the product was bought by person c or not.

2.5.2.3 Making Predictions

$$p(i) = \frac{\sum_{c \in C} w(c_{target}, c) * r_c}{\sum_{c \in C} w(c_{target}, c)}$$

where $p(i)$ represents the prediction for product i and r_c represents rating given by customer c for product i. C customer set only contains the customers who purchased product i.

2.5.3 Recommender Module

Recommender Module is responsible for reading/writing. The module has two tasks:

1. Getting transaction list which contains customer id product id pairs from the Neo4j database using neo4j driver and sending the list to the Filterer module as parameter.

2. Getting the recommendation list which contains ids of the customers and corresponding recommended products from the Filterer module and writing these recommendations to Neo4j database as a relationship between the customer and the recommended product using neo4j driver.

2.6 Testing Phase

Although there are many factors such as diversity, coverage, serendipity that determine the efficiency of recommendation systems, in this project we focused on accuracy and decided to left these factors beyond the scope.

2.6.1 Methods

2.6.1.1 Leave-one-out Cross Validation

2.6.1.2 K-Fold Cross Validation

2.6.1.3 Shuffle Split Cross Validation

2.6.2 Results

2.6.2.1 Stockmount

About the dataset: Dataset origanally consists of 142501 customers and 73482 products with implicit ratings (i.e. purchased or not). However, to make the customer versus product matrix a little denser, customers with less than 2 products and products with less than 2 customers are eliminated. After filtering,

	filtering threshold = 2
Number of Customers	
Number of Products	
Sparsity	99.808124 %

Testing method: Since the dataset contains implicit ratings, I preferred to use "Hit Rate" rather than RMSE or MAE as the test metric and "Leave-one-out CV" as the cross validation iterator. Basically, for each deleted

transaction, test module checked whether the product of the deleted transaction is within the recommended products.
Tested Recommender: Eigentrust Weighted Trust Based Recommender 2.4

Results:

	5 Recommendations	10 Recommendations
Number of Tests	4160	4160
Number of Hits	1367	1983
Hit Rate	0.328605	0.476682

2.6.2.2 Movielens100k [9]

About the dataset:

	Movielens 100k
Number of ratings	100.000
Number of users	943
Number of movies	1682
Rating range	1-5
Sparsity	93.6953 %

Testing method: I preferred to use RMSE and MAE as the test metric and "K-Fold CV" as the cross validation iterator. Tests were done using Surprise 2.6.3.1.
Tested Recommender: Inverse Distance Weighted Trust Based Recommender 2.5

Results:

2.6.2.3 Amazon Food Reviews [1]

About the dataset:

	Amazon Food Reviews (threshold = 10)
Number of Users	4276
Number of Products	1140
Rating Range	1-5

Testing method: For this dataset, I wanted to see how efficiently the recommender works on extremely sparse datasets, as a result I prefered to use "Shuffle Split CV" as the cross validation iterator since it is easy to set up the test and train set ratios. For the test metric, RMSE and MAE were used. Tests were done using Surprise 2.6.3.1.
Tested Recommender: Inverse Distance Weighted Trust Based Recommender 2.5

Benchmark:
Number of Splits: 3, Trainset: 0.5, Testset: 0.5, Sparsity: 0.9931

	Trust Based	MSD	SVD	Slope One	KNN	NMF
RMSE	0.7411	0.7898	0.8152	0.7553	0.7679	0.7516
MAE	0.3996	0.4037	0.61073	0.4264	0.4902	0.4832

Number of Splits: 3, Trainset: 0.2, Testset: 0.8, Sparsity: 0.997

	Trust Based	MSD	SVD	Slope One	KNN	NMF
RMSE	0.9766	1.551	1.0252	0.9045	1.0501	0.9659
MAE	0.6321	0.9870	0.8062	0.5104	0.7172	0.6932

Number of Splits: 3, Trainset: 0.1, Testset: 0.9, Sparsity: 0.998

	Trust Based	MSD	SVD	Slope One	KNN	NMF
RMSE	1.275	2.2077	1.1096	1.015	1.2562	1.1271
MAE	0.8909	1.6566	0.8806	0.6178	0.9067	0.8613

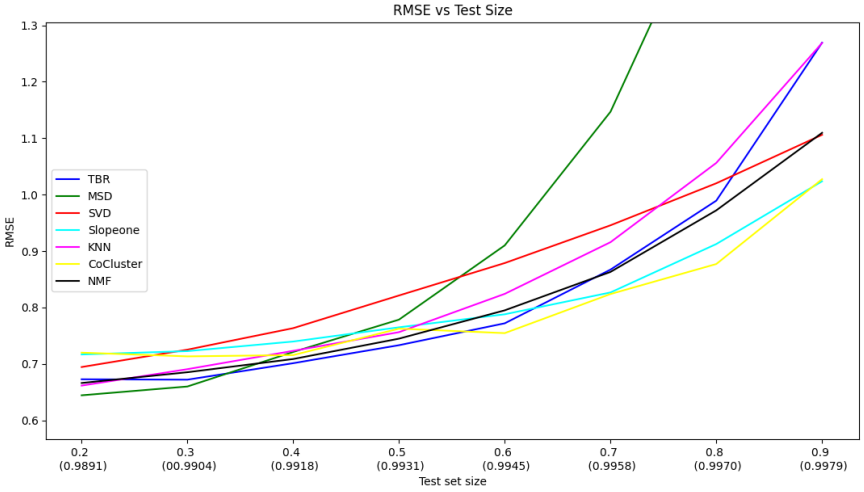

Figure 6: RMSE versus Testset Size. For instance, 0.2 means testset-trainset ratio is $20\% - 80\%$. Additionally, floats in parentheses represent the sparsity of trainset for corresponding test set size

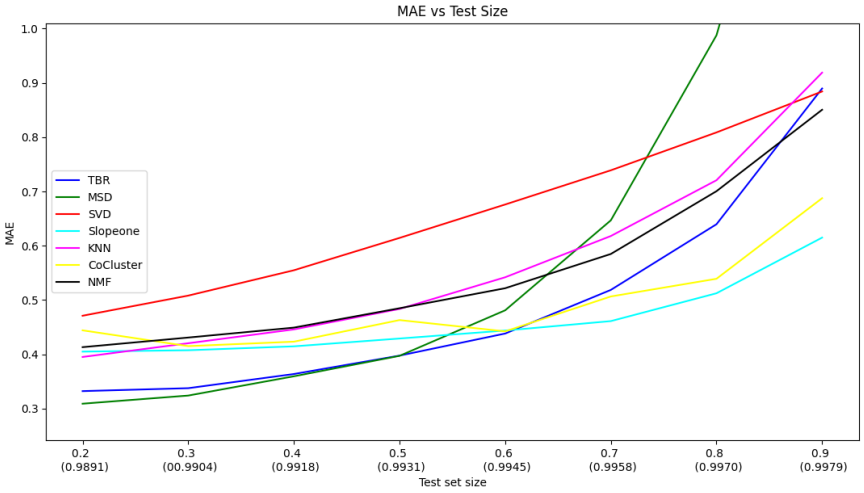

Figure 7: MAE versus Testset Size

2.6.3 Libraries I Used

2.6.3.1 Surprise [6]

Surprise is a Python library for building and analyzing recommender systems. We use it for evaluating the performance of the trust based recommenders on various datasets and comparing them with the provided built-in recommender systems.

Installation :

```
pip install scikit-surprise

```

Sample Usage :

```
from surprise import AlgoBase, PredictionImpossible,
 Dataset
from surprise.model_selection import cross_validate

class Inverse_distance_weighted_tbr(AlgoBase):
...

reader = reader = Reader(line_format='user item rating
 timestamp', sep=';', rating_scale=(1, 5))

data = Dataset.load_from_file('./dataset.csv', reader=
 reader)
algo = Inverse_distance_weighted_tbr()

cross_validate(algo, data, cv=5, verbose=True)

```

Listing 4: Surprise example

2.6.3.2 Matplotlib

Installation :

```
pip install matplotlib

```

Sample Usage :

```
import matplotlib.pyplot as plt
...
plt.hist(eigentrust_list,
color = 'blue',
edgecolor = 'black',
bins = bins)
plt.title('Distribution of the Eigentrust Values')
plt.xlabel('Range')
plt.ylabel('Number of Eigentrust Values in the range')
plt.show()

```

Listing 5: Matplotlib example

3 Organization

3.1 METU Data Mining Research Group

4 Conclusion

I had the chance to,

- recognize the main challenges such as sparsity, scalability, cold start etc. experienced by recommenders
- experience firsthand which types of algorithms are effective and ineffective in solving which problems
- use evaluation methods and understanding which method should be used where

For the future,

1. Not only customers and products records but extra attributes such as product supplier, platform, product category, seller, store, location, date, etc. can be made part of the system either adding them as nodes to the Graph of Inverse Distance Recommender or using them as an extra filtering layer as in "Hybrid Recommender Systems".

Discussions, propagating trust

References

[1] Amazon food reviews dataset. https://snap.stanford.edu/data/web-FineFoods.html.

[2] Epinions dataset. https://snap.stanford.edu/data/soc-Epinions1.html.

[3] Neo4j. https://neo4j.com/.

[4] Numpy library. https://numpy.org/.

[5] Scipy library. https://www.scipy.org/.

[6] Surprise library. http://surpriselib.com/.

[7] Aksoy, K., Odabas, M., Bozdogan, I., and Temur, A. Tacorec. https://senior.ceng.metu.edu.tr/2020/tacorec/, 2020.

[8] Dijkstra, E. W. Dijkstra's algorithm. https://en.wikipedia.org/wiki/Dijkstra%27s_algorithm.

[9] Harper, F. M., and Konstan, J. A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst. 5*, 4 (Dec. 2015).

[10] Kamvar, S. D., Schlosser, M. T., and Garcia-Molina, H. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th International Conference on World Wide Web* (New York, NY, USA, 2003), WWW '03, Association for Computing Machinery, p. 640–651.

[11] Lathia, N., Hailes, S., and Capra, L. Trust-based collaborative filtering. vol. 263.

[12] Li, Y., Liu, J., Ren, J., and Chang, Y. A novel implicit trust recommendation approach for rating prediction. *IEEE Access 8* (2020), 98305–98315.

[13] Massa, P., and Avesani, P. Trust-aware recommender systems. In *Proceedings of the 2007 ACM Conference on Recommender Systems* (New York, NY, USA, 2007), RecSys '07, Association for Computing Machinery, p. 17–24.

[14] Papagelis, M., Plexousakis, D., and Kutsuras, T. Alleviating the sparsity problem of collaborative filtering using trust inferences. In *Trust Management* (Berlin, Heidelberg, 2005), P. Herrmann, V. Issarny, and S. Shiu, Eds., Springer Berlin Heidelberg, pp. 224–239.