National College of Ireland

Product Bundle Recommendation using Collaboration of Matrix Factorization and Jaccard Similarity

MSc Research Project
Data Analytics

Anu Johny Thekadayil
x17106630

School of Computing
National College of Ireland

Supervisor: Dr. Sachin Sharma

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Anu Johny Thekadayil
Student ID:	x17106630
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Dr. Sachin Sharma
Submission Due Date:	17/09/2018
Project Title:	Product Bundle Recommendation using Collaboration of Matrix Factorization and Jaccard Similarity
Word Count:	5530

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	17th September 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Product Bundle Recommendation using Collaboration of Matrix Factorization and Jaccard Similarity

Anu Johny Thekadayil
x17106630
MSc Research Project in Data Analytics

17th September 2018

Abstract

Recommendation system is an filtering application which mainly does the work of providing recommendation based on the information collected from the users. This application is know to have been supporting the marketing and e-commerce sectors since a long time and has proved to be a critical part in it. Since recent, the recommendation systems mainly concentrated on single item recommendations but recent years has seen an increase in need for bundles. Product bundles are a boon for the enterprises as well as the users. The users tend to fall for the combination strategy and end up buying a bundle which usually includes at least two of their favorite items. Thus designing an optimal product bundle recommending system is important.

The research consists of the methodologies such as Matrix Factorization and Jaccard Similarity as they have proved to be excellent in their respective applications. This model recommends the product bundle as well as evaluates to check whether the bundle is similar in reality i.e. the bundle consisting of three products recommended by matrix Factorization is evaluated by the Jaccard score of their respective genre's.

The developed model is thoroughly evaluated using the k-fold cross validation and the performance metrics such as recall, precision and RMSE were taken into consideration. By the performance metrics we can say that this methodology provides good results which is confirmed by taking into consideration of all the metrics and moreover the Jaccard score as well as the works of other researches which suggested that a precision of 0.3 is considerably good.

1 Introduction

The onset of technology has paved way for e-commerce to a very large extent which directly led to a growth in online transactions and its associated data. Meeting the huge demand from customers in e-commerce is of utmost importance and failing to fulfil it will lead to the downfall of that e-commerce organization. According to Ge et al.; 2017, the research done by CECRC said that every day over 15 terabytes of information are generated. Even though this massive data pool gives us large meaningful information but the search cost exceeds accordingly. The same way, customers feel worn out when they search for products in a large online product catalog. Thus recommendation system

comes into existence which gives a small set of options to the users based on the analysis done on users interests. This has caught the eyes of various organizations and since then have tried to improve the recommendations, to make it better and efficient than before.

Recommendation systems try to lure customers into buying products by constantly showing what they like. There have been many attempts into improving these recommended item purchases and have also succeeded in it. One of the techniques adopted is that of the sale of group of products called as product bundles. We can see this in our day to day life e.g. buy two for the price of one, mobile phone sim purchase with an internet access at a lower price. Such technique is used for increasing the sale rate and offering discount to the customers which make them choose this product over other rival products. This could save the customers time and make their purchase cost and time efficient. The motivation behind bundling the products lies on mainly the sale of those products which would have not been purchased rather individually thus improving the overall sales margin of the enterprise and overall customer satisfaction Beladev et al. (2016).

The approach most widely used for recommendation system is collaborative filtering (CF) technique. Even though it got some drawbacks it has proved to be one of the best solution for recommendation and along with product bundling strategy it would be the best combination for an enterprise to be successful.

Designing such a combo is difficult as compared to designing a single item recommendation system. It is not just finding out which products should be recommended but also the relationship between the products in a bundle should be identified Beladev et al. (2016). Researches do not offer much insights as to how a product bundle operates when provided to customers and detailed evaluations is nowhere to be found except for the research work by Beladev et al. (2016) which includes the performance metrics of precision and recall as well.

The work described in this paper is how a bundle can be incorporated with the recommendation system. The use of state-of-the-art CF technique is considered due to its accuracy, mainly the matrix factorization. Such an approach is undertaken since most of the existing researches on recommendation systems focus on improving and optimizing single item recommendation or improving product bundle strategy and not the combination of both. Also, the collaborative filtering techniques used in various researches are not up to the mark when accuracy is considered and hence matrix factorization is considered in my research which is known to handle large datasets. Thus creating a system which recommends bundles in an efficient manner is the ultimate goal.

The research question 'How can efficient product bundles be recommended using a collaboration of matrix factorization and jaccard similarity?'can be seen explained in the following sections.

This combination is simple yet elegant that it would be able to recommend any kind of product. Jaccard similarity is known for its extravagance in handling vectors with no common features and Matrix factorization is famous for handling large and spare dataset.

The methodologies adopted and the implementation will be discussed thoroughly in the upcoming chapters. The next important section is that of literature review which gives the overview of how the evolution of recommendation system has taken place with respect to the topic in question here and what gap has been present in the existing literature that needs to be filled. Chapter 3 brings us closer to the approach undertaken for this research whereas chapter 4 specifies the techniques or the framework developed for the project. The detailed implementation is provided in chapter 5 which is followed

by the evaluation of the results i.e. chapter 6. Last but not the least, conclusion and future work is presented.

2 Related Work

2.1 Recommendation Systems

The first ever recommendation system was developed to overcome the overloading of mail inbox of the users. Basically in 1992, Goldberg et al. (1992) developed a filtering algorithm that lets people register their reactions to each documents they read. In the years gone by, the demand for recommendation system has increased to a very large extent due to the endless amount of data being available to the online customers. A RS basically uses information from variety of sources, if not one, and provides an insight/ prediction of items to the users Bobadilla et al. (2013). Usually this information is made available by the users themselves by rating the similar items or by keeping track of their activities. Recommender system is generated if it has access to the following Bobadilla et al. (2013):

1. The dataset
2. The algorithm that does filtering
3. Model (model based or memory based)
4. Level of sparsity in the database
5. System performance (memory utilization and time consumption)
6. RS objective
7. Output quality

From the above mentioned list it is the filtering algorithm that sets the base of a recommendation system. The major types of filtering techniques are collaborative filtering, content- based filtering, demographic filtering and hybrid filtering Bobadilla et al. (2013). Collaborative filtering is mainly known as a technique that helps people identify things based on their likeness as suggested in Resnick et al. (1994). In the past a system called GroupLens was developed for the collaborative filtering of Netnews (pool of news articles) which helps people identify the news articles that they would be interested in, based on the ratings provided by other readers. The application lets the readers rate the articles on a 5 pointer scale and they in turn were given suggestions based on their ratings. If we look into the collaborative filtering algorithm that was commonly used we find it to be k - Nearest Neighbor (kNN). The pros about collaborative filtering that can be considered is that they are independent of any content and its recommendations are serendipitous Beel (n.d.) but due to its black box nature no one can specify why an item recommendation was made. Content-based filtering (CBF) as the name suggests recommends to its users based on the analysis of the text dataset. User modeling process is the central component of CBF which extracts the likes of the user from items with which the user interacted Beel (n.d.). The interaction in this context means the tags associated with each item or the downloading or buying history of the users. Even though CBF is known to provide a more personalized recommendation to its users it requires large computing

power as compared to basic RS. Another aspect where CBF is criticized is due to its low serendipity and recommending similar items to users who already know about the items. Demographic filtering suggests that individuals having certain common personal traits are given recommendations based on the similarity among their purchases Bobadilla et al. (2013).

Hybrid filtering simply means the hybrid of two filtering techniques such as collaborative filtering and content-based filtering or collaborative filtering and demographic filtering. Hybrid filtering is undertaken to use the merits found in both the filtering techniques.

No matter which technique is used, whenever there occurs a sparsity problem, the solution includes dimensionality reduction which is mainly based on matrix factorization Bobadilla et al. (2013). Matrix Factorization (MF) is known to handle large datasets with scalable approaches. As said by Luo et al. (2012) collaborative filtering recommenders prefer Matrix Factorization due to its accuracy and scalability.

According to Stremersch and Tellis (n.d.) there are two type of bundling: Product and price bundling. Product bundling is the sale of two products merged together irrespective of their individual prices. It is more dependent on the physical aspect of the products. Price bundling only becomes successful if a discount is applied to the bundle i.e. in price bundling, the product characteristics alone does not create any value to the bundle but only by applying certain discount will it attract any customers. From application point of view we can say that product bundling looks at long term use while price bundling can be applied instantaneously.

2.2 Product Bundling

There have been various data analytics techniques applied to different products available in E-Commerce in the form of recommendation systems. Recommending just a single item seems to be an incomplete work. When customers is given the privilege of variety of options in various combinations then theres a possibility of a purchase. It is said that customers tend to like few clicks to get desired products than to search the whole website Liu et al. (2017). Those few clicks can be that of bundles of products. Earliest and most common type of product bundle can be seen with the telecom industry who offers broadband deals along with telephone connection Chan-Olmsted and Guo (2011). Liu et al. (2017) points out the relation between the bundling and items associativity i.e. the relationship between the products is the aftereffect of some association between the products and not just them being in a bundle. This relationship decides what features the customers are looking into while purchasing various bundles. As product bundling is known to reduce customers searching cost by improving the efficiency of bundle promotion strategies Guo-rong and Xi-zheng (2006), its use has gradually started.

Many types of researches have been conducted revolving around product bundling such as to optimize a bundle Ye et al. (2017), to determine product and price combination of bundles to maximize the revenue Li (n.d.), bundling based on market segmentation Beheshtian-Ardakani et al. (2018), personalizing bundle recommendation Pathak et al. (2017) and many more.

Even though product bundles are thoroughly researched upon, only a handful of researches were conducted for recommending product bundles. Following are the studies:

The first research Guo-rong and Xi-zheng (2006) is based on the recommending bundles using CF method. Basically the authors used adaptive resonance theory, a clus-

tering technique for identifying customer groups and later used association mining for finding relation between two different products. This methodology was used to identify how feasible it is for the product bundle to be recommended and the high hit probability was related to the bundling strategy and not on the actual bundles.

The second research has brought forward a novel idea called Bundle Recommendation Problem such that when this problem is solved, an optimal bundle list will be recommended with respect to the business objective Zhu et al. (2014). Basically the model would select m number of items for a user from a catalog containing n number of items after analysis on the likes of the user. The final evaluation has shown a positive outcome when the experiments were conducted both offline and online. One setback mentioned here is that they havent considered price bundling and is suggested as a further work.

Craw and Preece (2002) has brought forward a travel case model to recommend travel plans based on the users profile and taste. In this model the travel bag is considered to be the bundle and also, this model is like an interactive technology which reformulates the users queries so that an appropriate solution is given. Reformulation occurs whenever the queries posed by the users gave failure as the output. Each interaction with the model leads to the plans being stored in a memory which can be used later by the recommender system.

The fourth research by Beladev et al. (2016) mainly introduces a novel technique that integrates the collaborative filtering technique, demand graph and include price of items as well. They try to optimize the recommendation by maximizing the profit obtained by price bundling. After comparison between the single product recommendation and bundle recommendation it proved that bundle improves the accuracy. Even after its success there were certain drawbacks such as they couldn't extend their bundle to have more than 2 products, online testing was not performed and items features were not taken into consideration.

2.3 Methodologies in literature

In Table 1 the basic methodologies most of the researchers used can be seen.

As we have seen in Table 1, we noticed that the combination to be used in my project has not been tried and my proposal says that this combination should ideally be optimal which recommending bundles. My claims are backed by Koren et al. (2009) who says that Matrix Factorization is known to provide best results when recommendation is considered. MF used in Netflix prize data has proved that it outperforms the most commonly used k Nearest Neighbor methodology. MF is also known to exploit the informations such a implicit and explicit feedbacks and even the temporal dynamics. Originally MF does not see to it that its values are non-negative, this is identified by Hernando et al. (2016) who proposed the Non-Negative Matrix Factorization Technique. This techniques makes sure that the vector always lies within the limit [0,1] due to its probabilistic nature. Mnih and Salakhutdinov (n.d.) brought forward the Probabilistic Matrix Factorization (PMF) which can not only handle huge datasets but also minimal ratings can be easily handled. This projects aim was to develop an algorithm that can easily handle the data imbalance situation and is linearly scalable. The basic PMF can attain acceptably good accuracy but can be improved with insertion of constraints in it. Wu (2009) and Mnih and Salakhutdinov (n.d.) both agree on the fact that MF is mostly preferred due to its capability of handling sparse data effectively and its scalability. The binomial MF is another technique proposed by Wu (2009) which calculates a lower the root mean square

Research Domain	Application	Models Used	Reference
Entertainment	Movies	Profit and Utility Maximizer	Azaria et al. (2013)
Home	Furniture	Item-based NNMF Bigram SVD	Pradel et al. (2011)
	FoodMart	Clustering Association Rule	Guo-rong and Xi-zheng (2006)
	Xbox SuperMarket Electrical Product	Collaborative filtering Jaccard Measure Demand Graph	Beladev et al. (2016)
Tourism	travel	Case Base Intelligent HMI recommender	Craw and Preece (2002)
Shopping	Online	Personalized bundle	Liu et al. (2017)

Table 1: Methodology in Literatures

error and also reduces the overfitting problem experienced by a normal MF.

The similarity measure is an important factor that plays in some recommendation systems. There are measurements such as Cosine similarity, Manhattan similarity, Jaccard similarity etc... Although all these similarity measures are efficient in their own way, there's one drawback in Manhattan and cosine that says only those attributes sets that have some common aspect are considered. This drawback is not experienced in jaccard similarity where the overlap of two vectors is considered Goebel et al. (n.d.).

Thus after seeing tonnes of literatures it is safe to assume that there is still scope for further research in the recommendation and product bundling domain. This combination is one of a kind and needs more research in it.

3 Methodology

For any project to be successful there has to be certain norms and procedure that should be followed. The research work here is backed up by the CRISP-DM methodology which suggests us how practical and beneficial it is from a business point of view [1]. The research methodology is described on the basis of phases of CRISP-DM shown in Figure 1 and can be seen as follows:

- Business Understanding: This stage determines what our business is actually in need of. This phase uncovers various important factors that lies in the large dataset and can help decide how to find an outcome to it. This stage is a very crucial point which may make or break the entire project. Hence it should be thoroughly analyzed and move forward. The recommendation system is of utmost importance in any ecommerce enterprise. If there was no business value attached to it there wouldnt have been such competition to make it the best. For the sale to be double and make the customer happy by providing various options and reducing their shopping time

[1] https://www.sv-europe.com/crisp-dm-methodology/

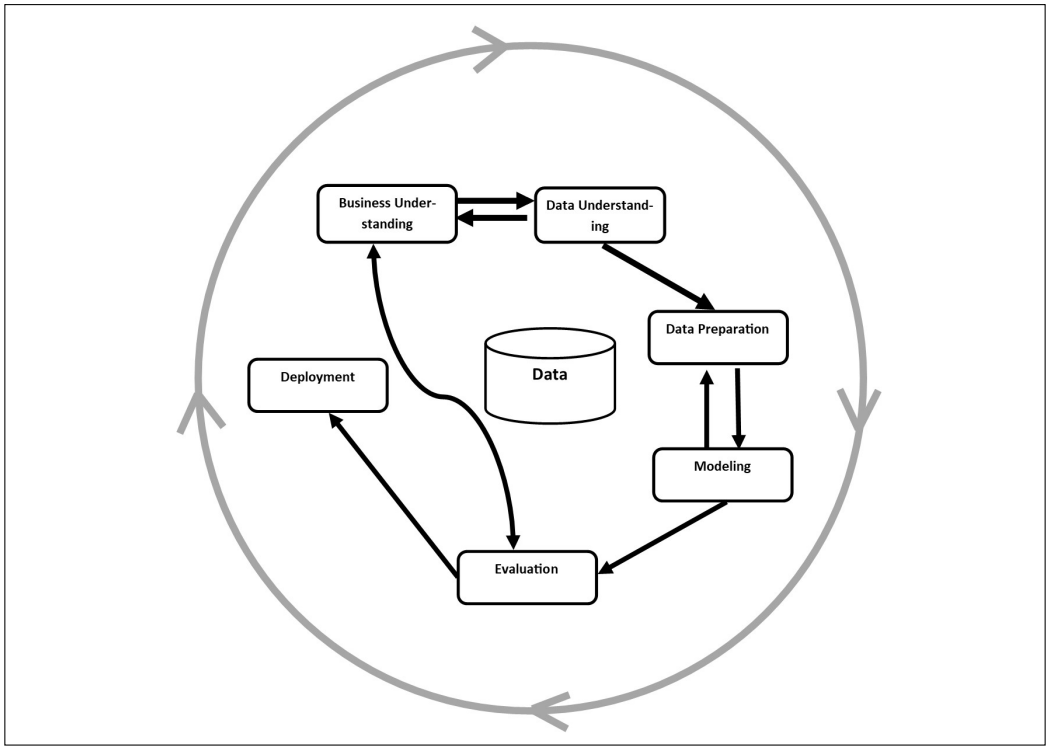

Figure 1: Process Flow of CRISP-DM

is the business objective in this project. It is this phase which decides what is to be done further such as which technology/tools to use, any constraints or obstacles that may arise, from where to get the data for testing and evaluation etc

- Data understanding: The second phase allows you to gather the data required for the project and have a deep understanding of how the data behaves. Data exploration is done in this phase which would help us to have deep understanding of what kind of data we are dealing with. After exploration, the data should be checked for its authenticity and also whether there are any quality issues. The data used in this project is downloaded from the movielens website, a subsidiary of grouplens. The dataset includes the users who have rated at least 20 movies and not less and includes information such as movieID, UserID, tags, ratings, timestamp, genre, movie names etc. In Table 2 we can see the data description of the datasets in consideration. The website from where the download took place is https://grouplens.org/datasets/movielens/.

Dataset	Attributes
Tags	userId,movieId,tag,timestamp
Movies	movieId,title,genres
Ratings	userId,movieId,rating,timestamp
links	movieId,imdbId,tmdbId

Table 2: Data Description

- Data Preparation: The third step is where you finalize what data to use. Based on various criteria such as feature selection, reducing the volume of data according to

the data mining goal is performed in this stage. Data cleaning and transformation is widely performed. The datasets from movielens were merged to form one common dataset on which the modeling and analysis was done. Also, it was made sure that no missing values and duplicate values were present and if any, it was taken care of. So the total ratings in the visualized form is given in Figure 2 wherein we can see that the movies were mostly rated as 4 which is considered to be a reasonably good score and which is almost 9000 more thane the rating 3.

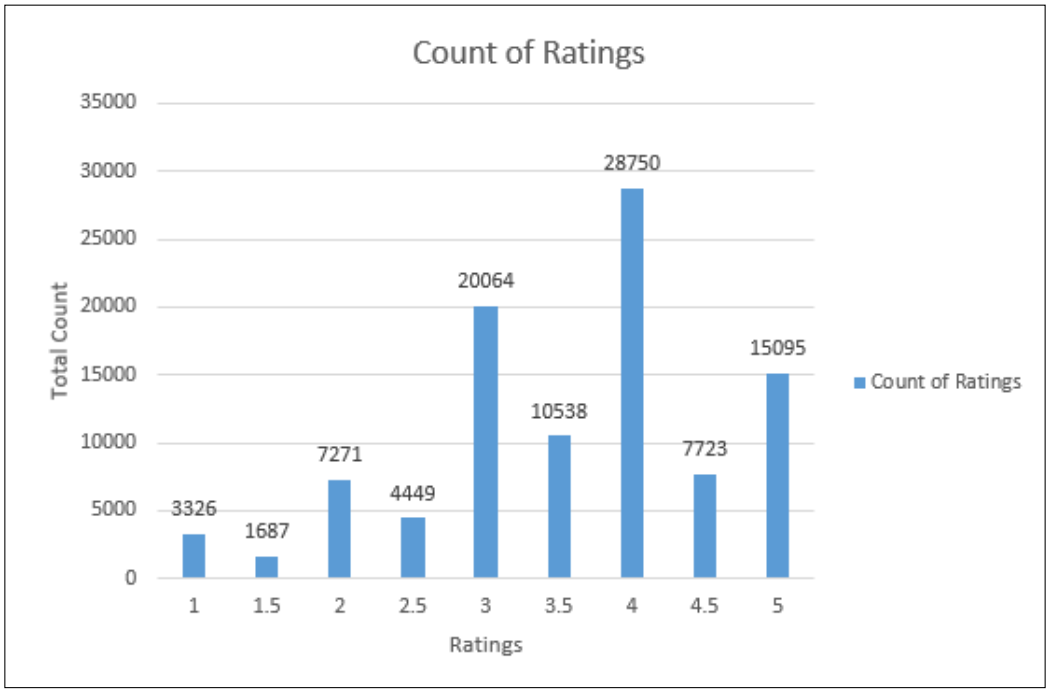

Figure 2: Total Ratings Count

- Data Modeling: This fourth stage lets us apply the modeling technique on the dataset. Even though we had selected which tool to use here in the business understanding phase, selection of the modeling technique is done here. The modeling technique decided is Matrix Factorization and jaccard similarity for finding out how much similar is each product. The model is developed using the R programming language in RStudio. After calculating the jaccard similarity score and the matrix from matrix factorization, the bundles are created. These bundles are then provided to the users based on their likes. The bundle size is decided by the enterprise and can be changed whenever required.

- Evaluation: Evaluation phase analyzes how this model tries to meet the needs of business requirements. The model is evaluated based on certain performance criteria and is checked whether all the conditions are met according to the business plan. The evaluation of my project is performed using the metric recall and precision and the final evaluation can be done by cross checking the jaccard score of the genres and the genres recommended which proves the similarity between the products in the bundle.

- Deployment: The last phase in my research is deployment wherein I will be carefully assessing the model and check its working on other dataset. Deployment also

ensures your model doesn't crash while its working.

4 Design Specification and Implementation

The implementation procedure followed for the success of this project is given in a diagrammatic form as seen in Figure 3. The following are the subsections which gives an overview of the implementation process.

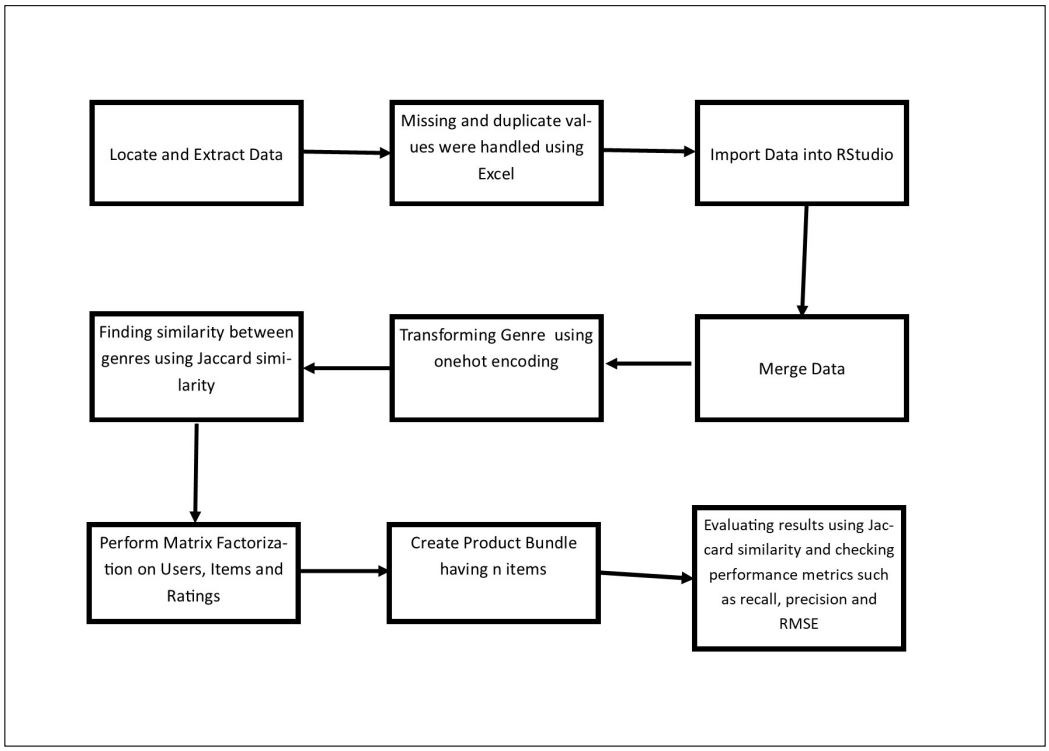

Figure 3: Process Flow Diagram

4.1 Dataset

The datasets downloaded from Movielens.com has four parts in it viz. movies.csv, links.csv, ratings.csv and tags.csv where movies and links have 9125 rows, tags have 1296 rows and ratings has 100004 rows. After the merging of movies and ratings file the total row count was 100004. The users in the dataset have at least rated 20 movies or more and there are unique 9125 movies in the dataset.

These datasets were initially checked for any missing values which was corrected instantly using the Microsoft Excel tool as well the duplicates were checked and eliminated as it emerged. The genre column in movies.csv had values separated by a separator i.e. each column in genre consisted of more than one type of genre which was then separated from each other using the encoding technique similar to onehot. The genre's were identified and individual genre were assigned to each columns which looked like a matrix. The genre's which were present alongside the movies were then assigned the number '1' if present or else '0' was assigned to that cell in the individual genre column. The Figure 4 clearly shows the encoding output of the genre's.

Even though the genre's were separated, in order to check which genre is the most popular i.e. mostly preferred watching is shown in the Figure 5. From the image we can see that it is the genre drama that is mostly opted to watch which is then followed by the genre comedy.

```
> genreMatrix
    Action Children Comedy Adventure Documentary Drama Fantasy Film-Noir Animation
1        0        1      1         1           0     0       1         0         1
2        0        1      0         1           0     0       1         0         0
3        0        0      1         0           0     0       0         0         0
4        0        0      1         0           0     1       0         0         0
5        0        0      1         0           0     0       0         0         0
6        1        0      0         0           0     0       0         0         0
7        0        0      1         0           0     0       0         0         0
8        0        1      0         1           0     0       0         0         0
9        1        0      0         0           0     0       0         0         0
10       1        0      0         1           0     0       0         0         0
11       0        0      1         0           0     1       0         0         0
12       0        0      1         0           0     0       0         0         0
13       0        1      0         1           0     0       0         0         1
14       0        0      0         0           0     1       0         0         0
15       1        0      0         1           0     0       0         0         0
16       0        0      0         0           0     1       0         0         0
17       0        0      0         0           0     1       0         0         0
18       0        0      1         0           0     0       0         0         0
19       0        0      1         0           0     0       0         0         0
20       1        0      1         0           0     1       0         0         0
21       0        0      1         0           0     0       0         0         0
22       0        0      0         0           0     1       0         0         0
23       1        0      0         0           0     0       0         0         0
24       0        0      0         0           0     1       0         0         0
25       0        0      0         0           0     1       0         0         0
26       0        0      0         0           0     1       0         0         0
```

Figure 4: Encoding of Genre

4.2 Process

The main implementation of the model starts here. After the data gathering and transformation is performed, the next stage is to formulate the methodology/technique to be used. The models are then fed with the dataset to produce outputs i.e. product bundles for each users. The process of finding the similarity to identifying the missing ratings in matrix factorization to creating the bundles is explained in the following sections.

4.2.1 Jaccard Similarity

The Jaccard similarity is basically the similarity measurement of variables that have asymmetric information. The basic formula for understanding the jaccard score is the intersection of two set divided by the union of those same sets. Jaccard similarity has values that lie between 0 and 1. It is known to overcome the problem of not having any similarity score when there is nothing in common. Whenever two users who watch a same movie having different opinions are considered they are still considered to be similar as per jaccard similarity. The onehot encoding performed on the genres are used specifically for the calculation of jaccard similarity. These binary genres were applied to the jaccard method in R and provided with a large matrix which showed the individuals scores that each genre possesses. It directly tells us which genres go together well.

4.2.2 Matrix Factorization

The matrix factorization in this project is implemented using the columns movies, ratings and user. Before moving forward with this process, the ratings given to each movies by

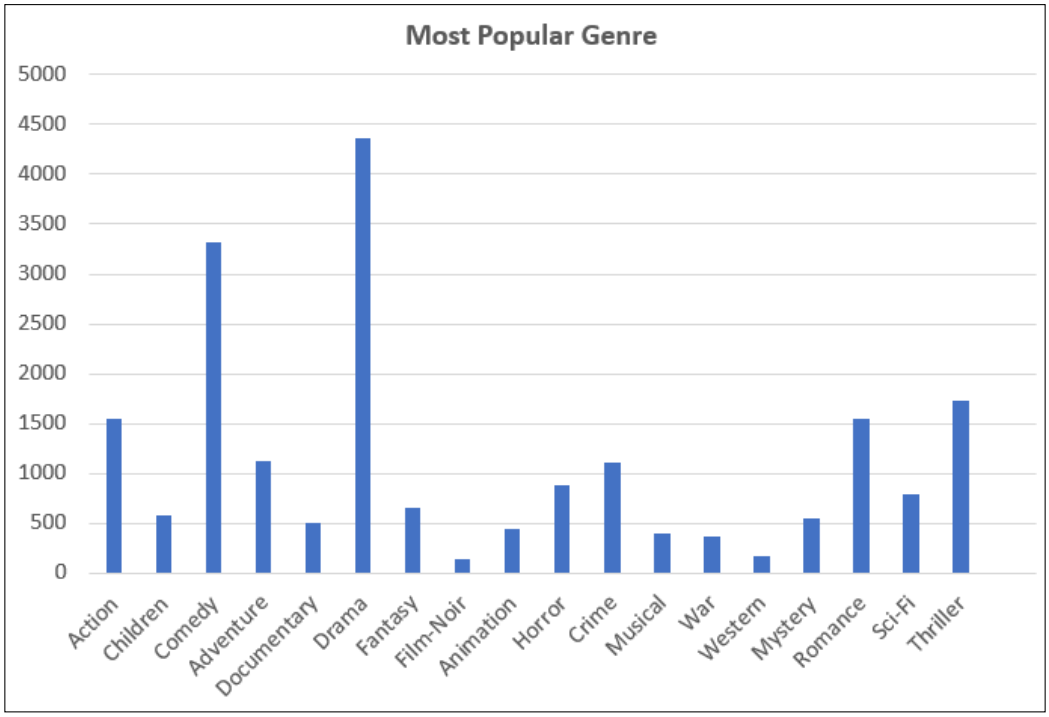

Figure 5: Most Popular Genre

each users is shown in the Figure 6. The sparsity of data can be evidently seen in this figure. By using the predefined libraries such as recosystem and recommenderlab available in R the implementation of MF was performed. Two models were created using the libraries but the recosystem library model performed poorly in terms of predicting the ratings hence had to be discontinued from the bundling phase. As the dataset consisted of around 100004 rows there was a huge possibility that the data would be sparse and the only way to tackle is to use MF. The users, ratings and movies are fed into the Matrix Factorization algorithm which then creates a matrix capable of predicting the missing ratings. Matrix factorization is basically converting a large matrix into multiple small matrices. The basic formula for matrix factorization is as given below

$$R = XY^T$$

where R stands for the large matrix and X and Y stands for the smaller matrices (X*K) and (Y*K) respectively. In this case the Matrix formed is decomposed into User affinity matrix and Movie affinity matrix which are the matrices of a large matrix. The result obtained can be seen in the Figure 7 which presents us with the movies that are bundled together for the userId 2.

5 Evaluation

The evaluation phase of a project is of utmost importance as it is in this phase that the accuracy and reliability of the model is tested. The most commonly used metric for evaluation is the use of precision and recall and then the use of K-fold mechanism which checks whether the model is being over-fitted.

userId	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	NA	NA	NA	NA	NA	NA	NA	NA	NA	4.0	NA	NA	NA	NA	NA	NA	5.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
4	NA	NA	NA	NA	NA	NA	NA	NA	NA	4.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5	NA	NA	4.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
7	3.0	NA	NA	NA	NA	NA	NA	NA	NA	3.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3.0	NA	NA	NA	NA	NA	NA
8	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9	4.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4.0	NA	NA	NA	NA	NA	NA	NA	NA	3	NA
10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
13	5.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
15	2.0	2.0	NA	NA	4.5	4.0	NA	NA	NA	3.0	2.5	NA	NA	2.5	NA	3.5	3.0	NA	1.0	NA	4.5	2.5	NA	NA	3.0	NA	NA
16	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
17	NA	NA	NA	NA	NA	4.5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4.5	NA	NA
18	NA	NA	NA	NA	3.0	4.0	3.0	NA	3	NA	NA	NA	NA	2.0	NA	NA	4.0	3.0	NA	NA	NA	NA	NA	NA	5.0	NA	NA
19	3.0	3.0	3.0	3	NA	3.0	3.0	NA	3	3.0	3.0	NA	NA	5.0	NA	5.0	NA	NA	NA	NA	3.0	3.0	1.0	NA	3.0	NA	NA
20	3.5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
21	NA	NA	NA	NA	NA	NA	NA	NA	NA	3.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3.0	NA	NA	NA	NA	NA	NA
22	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
23	3.0	NA	NA	NA	NA	3.5	NA	NA	NA	NA	3.5	NA	NA	NA	NA	4.0	NA	NA	2.0	1.5	NA	NA	NA	3.5	NA	NA	NA
24	NA	NA	NA	NA	NA	5.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
25	NA	NA	3.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
26	5.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Figure 6: Rating matrix

	Recommended movie	Genre
1	Fargo (1996)	Comedy|Crime|Drama|Thriller
2	Dark Knight, The (2008)	Action|Crime|Drama|IMAX
3	Blues Brothers, The (1980)	Action|Comedy|Musical

Figure 7: Bundle of three movies

The k-fold validation is used to overcome the issue caused by validation set. In k-fold validation the dataset is divided into k small sets which in my case is 5 folds. The k value is selected to be 5 because it is an ideal value for recommendation system. Once the k value is decided, the training data with k-1 folds is used to train the model and the remaining is used to evaluate the model. This action provides us with the evaluation metrics in each fold. The k-fold validation provides us with the values such as true positives, true negatives, false positives, false negatives, recall and precision as well. The terms recall and precision gives us not just mere metrics but many metrics are related to it. Recall is a models ability to find all the cases that are relevant within a dataset while precision is the models ability to capture only relevant data points.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegative}$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

The precision and recall obtained in Table 3 shows us that as the number of iteration is increasing the precision value is decreasing slightly, whereas it is exactly opposite in the case of recall. When we take into consideration the recall value we see that it is too low to correctly recommend so a need to improve it should be considered. But for three recommendations in a bundle the precision obtained is comparatively good and the precision when compared to previous results achieved by other researchers such as Beladev et al. (2016) and Liu et al. (2017) this result came out to be better.

	TP	**FP**	**FN**	**TN**	**precision**	**recall**	**TPR**	**FPR**
1	0.2740741	0.60741	22.68889	9039.43	0.310924	0.027677	0.027677	6.71E-05
3	0.5925926	2.05185	22.37037	9037.985	0.224089	0.058166	0.05817	2.27E-04
5	0.8148148	3.59259	22.14815	9036.444	0.184874	0.071585	0.071585	3.97E-04
10	1.3481481	7.46667	21.61481	9032.57	0.152941	0.110467	0.110467	8.26E-04
15	1.8296296	11.39259	21.13333	9028.644	0.138375	0.141299	0.141299	1.26E-03

Table 3: Evaluation Metrics

The jaccard similarity is a score that has a very important role in this project. The bundles recommended by using matrix factorization returns the movie names along with its genre. To check how similar these bundles are on the basis of genre we use jaccard similarity correlation matrix. This matrix clearly shows the scores of correlation between each other as we can see in Table 4 and Table 5. The score between genres animation and drama can be seen as 99% which shows that they are highly correlated to each other. On similar basis using this table we can cross check the genres displayed in the bundle.

The highlighted portions in Table 4 and Table 5 shows the genre's identified in the product bundle as given in Figure 7 marked in red box and we can see individual similarities between each genre that occurred at least twice which is highlighted in grey.

6 Conclusion and Future Work

In this work a novel yet explainable product recommendation model is described which can contribute to an enterprise's positive growth. The methodlogy included jaccard similarity and the collaborative filtering method, matrix factorization which recommended

	Action	Child	Com	Adv	Doc	Drama	Fantasy	Film	Anim
Children	0.97								
Comedy	0.92	0.92							
Adventure	0.75	0.81	0.92						
Doc	1	1	0.99	1					
Drama	0.91	0.98	0.86	0.94	1				
Fantasy	0.92	0.83	0.93	0.81	1	0.96			
Film-Noir	1	1	1	1	1	0.98	1		
Animation	0.95	0.68	0.95	0.84	1	0.99	0.85	1	
Horror	0.95	1	0.96	0.98	1	0.97	0.95	1	0.99
Crime	0.85	0.99	0.93	0.97	1	0.87	0.99	0.94	0.99
Musical	1	0.9	0.95	0.97	0.97	0.97	0.94	1	0.91
War	0.93	1	0.99	0.96	0.98	0.94	0.99	1	0.99
Western	0.98	0.99	0.99	0.97	1	0.99	1	1	0.99
Mystery	0.96	0.99	0.98	0.98	1	0.94	0.96	0.95	0.99
Romance	0.97	0.98	0.8	0.96	1	0.81	0.95	0.99	0.98
Sci-Fi	0.81	0.97	0.96	0.85	1	0.96	0.93	1	0.93
Thriller	0.79	1	0.97	0.93	1	0.85	0.97	0.97	0.99

Table 4: Jaccard Similarity - Part I

	Horror	Crime	Musical	War	Western	Mystery	Romance	Sci-Fi
Crime	0.98							
Musical	0.99	0.99						
War	1	0.99	0.99					
Western	1	0.99	0.99	0.98				
Mystery	0.9	0.89	0.99	0.99	1			
Romance	0.99	0.97	0.94	0.97	0.99	0.97		
Sci-Fi	0.89	0.99	0.99	0.99	0.99	0.95	0.98	
Thriller	0.82	0.77	1	0.98	0.99	0.83	0.97	0.88

Table 5: Jaccard Similarity - Part II

the top N bundles respective to each user. An important contribution of my project is that when everyone is trying to find optimal solution in single recommendation system domain, I've done a bundle recommendation and that too using jaccard similarity and the Matrix Factorization.

In the evaluation section we see that the recommended bundles are once more thoroughly evaluated using the jaccard scores which proved that the product in the bundles are in fact similar to each other. Another methodology used is by checking the precision and recall which showed that the precision was better as compared to recall. Thus further research has to be carried out to overcome this issue.

This work has had certain drawbacks which needs to be extended and addressed in the future such as:

The online testing: All the evaluations in this project was done offline. When users are with online application, their response could really help in this research domain.

Same item Bundle: The bundles that were created had different movies in it. If a customer is in need of two same copies, such recommendations need to be made.

Even though there are drawbacks, the results obtained are comparatively better as compared to the previous researches. There is still scope for further research in this domain by considering the real time data and recommending on a real time basis.

References

Azaria, A., Hassidim, A., Kraus, S., Eshkol, A., Weintraub, O. and Netanely, I. (2013). Movie recommender system for profit maximization, ACM Press, pp. 121–128.
URL: *http://dl.acm.org/citation.cfm?doid=2507157.2507162*

Beel, J. (n.d.). Towards Effective Research-Paper Recommender Systems and User Modeling based on Mind Maps, p. 329.

Beheshtian-Ardakani, A., Fathian, M. and Gholamian, M. (2018). A novel model for product bundling and direct marketing in e-commerce based on market segmentation, *Decision Science Letters* pp. 39–54.
URL: *http://www.growingscience.com/dsl/Vol7/dsl*$_2$*017*$_1$*4.pdf*

Beladev, M., Rokach, L. and Shapira, B. (2016). Recommender systems for product bundling, *Knowledge-Based Systems* **111**: 193–206.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0950705116302751*

Bobadilla, J., Ortega, F., Hernando, A. and Gutirrez, A. (2013). Recommender systems survey, *Knowledge-Based Systems* **46**: 109–132.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0950705113001044*

Chan-Olmsted, S. M. and Guo, M. (2011). Strategic Bundling of Telecommunications Services: Triple-Play Strategies in The Cable TV and Telephone Industries, *Journal of Media Business Studies* **8**(2): 63–81.
URL: *http://www.tandfonline.com/doi/abs/10.1080/16522354.2011.11073523*

Craw, S. and Preece, A. (eds) (2002). *Advances in case-based reasoning: 6th European conference, ECCBR 2002, Aberdeen, Scotland, UK, September 4-7, 2002: proceedings*, number 2416 in *Lecture notes in computer science ; Lecture notes in artificial intelligence*, Springer, Berlin ; New York.

Ge, X., Zhang, Y., Qian, Y. and Yuan, H. (2017). Effects of product characteristics on the bundling strategy implemented by recommendation systems, IEEE, pp. 1–6.
URL: *http://ieeexplore.ieee.org/document/7996297/*

Goebel, E. R., Siekmann, J. and Wahlster, W. (n.d.). Lecture Notes in Artificial Intelligence, p. 454.

Goldberg, D., Nichols, D., Oki, B. M. and Terry, D. (1992). Using collaborative filtering to weave an information tapestry, *Communications of the ACM* **35**(12): 61–70.
URL: *http://portal.acm.org/citation.cfm?doid=138859.138867*

Guo-rong, L. and Xi-zheng, Z. (2006). Collaborative Filtering Based Recommendation System for Product Bundling, IEEE, pp. 251–254.
URL: *http://ieeexplore.ieee.org/document/4104904/*

Hernando, A., Bobadilla, J. and Ortega, F. (2016). A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model, *Knowledge-Based Systems* **97**: 188–202.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0950705115005006*

Koren, Y., Bell, R. and Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems, *Computer* **42**(8): 30–37.
URL: *http://ieeexplore.ieee.org/document/5197422/*

Li, Y. (n.d.). A DATA MINING FRAMEWORK FOR PRODUCT BUNDLE DESIGN AND PRICING, p. 56.

Liu, G., Fu, Y., Chen, G., Xiong, H. and Chen, C. (2017). Modeling Buying Motives for Personalized Product Bundle Recommendation, *ACM Transactions on Knowledge Discovery from Data* **11**(3): 1–26.
URL: *http://dl.acm.org/citation.cfm?doid=3058790.3022185*

Luo, X., Xia, Y. and Zhu, Q. (2012). Incremental Collaborative Filtering recommender based on Regularized Matrix Factorization, *Knowledge-Based Systems* **27**: 271–280.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0950705111002073*

Mnih, A. and Salakhutdinov, R. R. (n.d.). Probabilistic Matrix Factorization, p. 8.

Pathak, A., Gupta, K. and McAuley, J. (2017). Generating and Personalizing Bundle Recommendations on *Steam*, ACM Press, pp. 1073–1076.
URL: *http://dl.acm.org/citation.cfm?doid=3077136.3080724*

Pradel, B., Sean, S., Delporte, J., Gurif, S., Rouveirol, C., Usunier, N., Fogelman-Souli, F. and Dufau-Joel, F. (2011). A case study in a recommender system based on purchase data, ACM Press, p. 377.
URL: *http://dl.acm.org/citation.cfm?doid=2020408.2020470*

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews, ACM Press, pp. 175–186.
URL: *http://portal.acm.org/citation.cfm?doid=192844.192905*

Stremersch, S. and Tellis, G. J. (n.d.). Strategic Bundling of Products and Prices: A New Synthesis fbr i\/lariceting, p. 19.

Wu, J. (2009). Binomial Matrix Factorization for Discrete Collaborative Filtering, IEEE, pp. 1046–1051.
URL: *http://ieeexplore.ieee.org/document/5360354/*

Ye, L., Xie, H., Wu, W. and Lui, J. C. (2017). Mining Customer Valuations to Optimize Product Bundling Strategy, IEEE, pp. 555–564.
URL: *http://ieeexplore.ieee.org/document/8215528/*

Zhu, T., Harrington, P., Li, J. and Tang, L. (2014). Bundle recommendation in ecommerce, ACM Press, pp. 657–666.
URL: *http://dl.acm.org/citation.cfm?doid=2600428.2609603*