Exploring Collaborative Filtering Recommender

System for Scratch

Roman Shaikh, B.Tech.

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Data Science)

Supervisor: Assistant Prof. Glenn Strong

August 2019

Declaration

I, the undersigned, declare that this work has not previously been submitted as an

exercise for a degree at this, or any other University, and that unless otherwise stated,

is my own work.

Roman Shaikh

August 13, 2019

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis

upon request.

Roman Shaikh

August 13, 2019

Acknowledgments

I would like to thank my supervisor Prof. Glenn Strong for his continuous guidance

and the help he has given me throughout my thesis. I am also thankful to the School

of Computer Science and Statics, Trinity College for providing a solid infrastructure

and a excellent environment for working on my thesis. I would also like to extend my

sincere gratitude to my second reader Prof. Tim Savage for his valuable input and

suggestions on improving the thesis during the presentation.

Finally, I would like to thank my parents for supporting me and my each and every

decision in life.

Roman Shaikh

University of Dublin, Trinity College August 2019

Exploring Collaborative Filtering Recommender

System for Scratch

Roman Shaikh, Master of Science in Computer Science

University of Dublin, Trinity College, 2019

Supervisor: Assistant Prof. Glenn Strong

Scratch is online learning platform developed by Lifelong Kindergarten Group at the MIT Media Labs. Scratch helps young and inexperienced students to develop program- ming skills and think creatively. With Scratch, one can program interactive stories, games, and animation. It also provides a collaborative platform through which users can share their own code and develop on others users work. It has often been observed that some users get demotivated easily because either they are unsure of where to go further after they have started or because the programming exercises are not up to their individual expectations. Thus, the concern arises that how do we keep users motivated on Scratch and improve the user experience. An effective solution is to recommend users with projects from other users according to their level of knowledge and previous experience. This intuitively is known as Recommender Systems (RSs), a system that recommends users with contents based upon their previous activities. Recommender system in an educational environment is proven to be significantly beneficial. We anal- yse the data made available by Scratch community and try to suggest an effective recommendation method. We explore the traditional recommendation techniques such as content-based filtering and collaborative filtering methods on dataset and compare the recommendation results. We explore the different methods for finding the recom- mendations from the dataset and the \*Weighted average and Multiple linear regression to evaluate the predictions.

Contents

Acknowledgments iii

Abstract iv

List of Tables viii

List of Figures ix

Chapter 1 Introduction 1 1.1 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3 1.2 Research Objective . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5 1.3 Research Challenges . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5 1.4 Dissertation Overview . . . . . . . . . . . . . . . . . . . . . . . . . . . 5 1.5 Dissertation Structure . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

Chapter 2 State of the Art 7 2.1 Recommender system overview . . . . . . . . . . . . . . . . . . . . . . 7 2.1.1 What is a recommender system? . . . . . . . . . . . . . . . . . . 7 2.1.2 Information filtering techniques . . . . . . . . . . . . . . . . . . 7 2.1.3 Challenges in user-based collaborative filtering algorithms . . . 9 2.2 Recommender systems in e-learning environments . . . . . . . . . . . . 10 2.2.1 E-learning space and recommendation . . . . . . . . . . . . . . 10 2.2.2 Recommendations in scratch . . . . . . . . . . . . . . . . . . . . 11 2.3 Evaluation of a recommendation system . . . . . . . . . . . . . . . . . 11 2.4 Other techniques in recommendation systems . . . . . . . . . . . . . . 12

v

Chapter 3 Methodology 14 3.1 Summary of the Collaborative Filtering process . . . . . . . . . . . . . 14 3.2 The item-based collaborative filtering algorithm . . . . . . . . . . . . . 16 3.2.1 Similarity computation . . . . . . . . . . . . . . . . . . . . . . . 16 3.2.2 Cosine based similarity . . . . . . . . . . . . . . . . . . . . . . . 17 3.2.3 Correlation-based similarity . . . . . . . . . . . . . . . . . . . . 18 3.2.4 Adjusted-cosine based closeness . . . . . . . . . . . . . . . . . . 18 3.3 Prediction Calculation . . . . . . . . . . . . . . . . . . . . . . . . . . . 19 3.3.1 Weighted sum method . . . . . . . . . . . . . . . . . . . . . . . 19 3.3.2 Regression model . . . . . . . . . . . . . . . . . . . . . . . . . . 20

Chapter 4 Implementation 21 4.1 Collecting the data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21 4.1.1 Data description . . . . . . . . . . . . . . . . . . . . . . . . . . 21 4.1.2 Ethical implication . . . . . . . . . . . . . . . . . . . . . . . . . 22 4.1.3 Identification of the relevant dataset . . . . . . . . . . . . . . . 22 4.2 Data Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24 4.3 Calculating similarity matrix . . . . . . . . . . . . . . . . . . . . . . . . 28 4.3.1 Preparing the data . . . . . . . . . . . . . . . . . . . . . . . . . 28 4.3.2 Imputing missing data . . . . . . . . . . . . . . . . . . . . . . . 29 4.3.3 Correlation-based similarity . . . . . . . . . . . . . . . . . . . . 30 4.3.4 Cosine-based similarity . . . . . . . . . . . . . . . . . . . . . . . 30 4.3.5 Adjusted-cosine based similarity . . . . . . . . . . . . . . . . . . 31 4.3.6 Selecting top-N recommendations . . . . . . . . . . . . . . . . . 32 4.4 Prediction Calculation . . . . . . . . . . . . . . . . . . . . . . . . . . . 33 4.4.1 Assigning the rating score for user activities . . . . . . . . . . . 33 4.4.2 Weighted Sum method . . . . . . . . . . . . . . . . . . . . . . . 33 4.4.3 Comparing the recommended items . . . . . . . . . . . . . . . . 34

Chapter 5 Method of Evaluation 36 5.1 Experiment . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36 5.2 Evaluation methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36 5.2.1 Statical accuracy metrics . . . . . . . . . . . . . . . . . . . . . . 37

vi

5.2.2 Decision support accuracy metric . . . . . . . . . . . . . . . . . 37 5.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 39 5.3.1 Comparing the similarity algorithms . . . . . . . . . . . . . . . 39

Chapter 6 Discussion and Future work 41 6.1 Limitations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41 6.1.1 Over specialisation problem . . . . . . . . . . . . . . . . . . . . 41 6.1.2 Concentration bias . . . . . . . . . . . . . . . . . . . . . . . . . 42 6.1.3 Cold Start problem . . . . . . . . . . . . . . . . . . . . . . . . . 42 6.1.4 Dataset limitation . . . . . . . . . . . . . . . . . . . . . . . . . 42 6.2 Future work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 43 6.2.1 Qualitative testing . . . . . . . . . . . . . . . . . . . . . . . . . 43 6.2.2 Mining text from the dataset . . . . . . . . . . . . . . . . . . . 44 6.2.3 Social network analysis . . . . . . . . . . . . . . . . . . . . . . . 45 6.2.4 Recommendation based on heuristics . . . . . . . . . . . . . . . 45

Chapter 7 Conclusion 46 7.1 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 46

Bibliography 49

Appendices 54 .1 A . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55

vii

List of Tables

4.1 Mean, Variance and Standard deviation for Friends and Projects data . 28 4.2 Projects recommended . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

viii

List of Figures

1.1 Projects by Scratch users . . . . . . . . . . . . . . . . . . . . . . . . . . 2

3.1 The process of a Collaborative Filtering Algorithm . . . . . . . . . . . 15 3.2 Computation of item-item similarity . . . . . . . . . . . . . . . . . . . 16

4.1 The distribution of projects in dataset over the period of 5 years . . . . 25 4.2 The frequency of the use of project blocks . . . . . . . . . . . . . . . . 26 4.3 Distribution of the user accounts created by creation date . . . . . . . . 27 4.4 Distribution of the friends over time . . . . . . . . . . . . . . . . . . . . 27 4.5 Pearson correlation output . . . . . . . . . . . . . . . . . . . . . . . . . 30 4.6 Cosine similarity computation output . . . . . . . . . . . . . . . . . . . 31 4.7 Adjusted cosine similarity output . . . . . . . . . . . . . . . . . . . . . 32 4.8 Top 5 project ids according to prediction . . . . . . . . . . . . . . . . . 34

5.1 Performance metrics calculations from Confusion metrics. . . . . . . . . 38 5.2 Performance comparison of similarity calculation method . . . . . . . . 40

ix

Chapter 1

Introduction

With the development of technology more and more focus is shifting on teaching kids the concepts of programming. Online learning is becoming an important part of our modern world. Development of Creative thinking is one of the fundamentals of early education. While kids usually get excited about learning new things, they lose interest pretty quickly. As difficult and challenging as the subject of programming could be, keeping the audience interested is even more so. Scratch by LLK group at MIT media labs is one such platform that aims at developing programming skills in kids while also keeping them interested. Scratch is not only aimed at developing programming skills but also enabling creative thinking and collaborating skills in children [1].

Scratch is a graphical programming environment that enables the user to create rich interactive projects including various media [2]. We can see various examples where users have developed comprehensive range of programs like animations, story- lines, games, music videos, online games, science projects, and sensor-driven projects (Figure 1.1).

1

Draft of 3:58 am, Tuesday, August 13, 2019 2

Figure 1.1: Projects by Scratch users

The Scratch platform lets users import various media elements like sounds and im- ages into a project or even create internally using the inbuilt editor. These objects (called Spirits) can be made to interact with each other via programming. Program- ming is done using Lego-like blocks that snap together to form a logical command. This projects can then be saved on the computer or can be shared online on the scratch website.

A key objective of Scratch is to help self-coordinated learning through tinkering and joint effort with companions [2]. Commonly users learn Scratch as they go along exploring, trying and attempting commands from the command palette or investigating code from existing undertakings. To encourage such self-directed learning, the Scratch programming environment was designed to encourage scripting, give prompt feedback to content execution, and make information obvious [2]. When a project is shared online on the scratch website, it can be seen by other peers/users. Users can comment on the project to give feedback, give a like to the project or add the project to their favourite. Peers can also use the existing project and build upon the top of it. This is known as remixing. Scratch encourages users to explore the work by other users and collaborate to build on it. At the beginning some Scratchers (Scratch users) complained about the projects remix, claiming that others were stealing their work [3]. This led to a

Draft of 3:58 am, Tuesday, August 13, 2019 3

discussion on the website forum about the idea of sharing in an open source community. Scratches goal is to make the users feel proud when their projects are remixed. Scratch consistently adds new features to the website to support this attitude. Whenever a project or task is remixed scratch adds a link to the project where it originally came from, this gives the original author full credit. Also the website’s landing page features top-remixed projects.

Some interesting use of scratch is the users forming online companies. LLk observed that some of the online users have started their own ”company” with the help of Scratch. A girl of 15 years of age from a town in England who goes by the alias BeeBop, set up a project with animations and sounds sprites and encourage users to use them in their own project or give a individual order on request for a custom build sprite. She was starting up the no-fee consulting company. Later a girl with alias MusicalMoon, who was also from the UK, really liked the project by BeeBop and inquired if she would be interested in making a custom background for her own personal project. That is how Mesh Inc. started a self declared mini company which provided remarkable games in Scratch. After some time the company was joined by a boy from US of 14, who said he could help with debugging since he was good at Scratch programming” [3].

Scratch is also being used in formal learning environments at universities to in- troduce high-school students and children to coding concepts. Studies have shown that visual programming languages have a notable influence on the understanding of computer programming concepts [4]. Some of the well-known visual programming languages like Alice [5], RAPTOR [6] and Etoys [7] have been utilized by various ed- ucational establishments. Scratch is one of the most popular and well-known visual programming language today [8],[2]. Although originally designed for kids, it was ac- cepted for introductory programming lessons by a few higher education organizations [9].1.1 Motivation

One of the main objectives of Scratch is to encourage collaboration amongst peers [2]. This is done by sharing the projects online with the community, getting feedback, tinkering with and mixing others work with their own ideas. If a user likes the project

Draft of 3:58 am, Tuesday, August 13, 2019 4

scratch provides options to FAVOURITE, LOVE or REMIX it. Users can also leave comments on the project. Users can add other users (for example the author of the project they like) as friends i.e. follow them.

The development of Scratch is always tightly associated with the development of Scratch online website[10]. For Scratch to be successful, it must be connected to a community which supports collaboration and feedback from one another and improving on others work [11]. Hence Scratch is inherently built with the idea of ”Sharing” with the community. Scratch gives a ”Share” option on the project page. Once the project is shared on the website anyone can run it , tinker with it comment on it and vote using the Love button.

Following the launch of Scratch in 27 months, users have shared more than 500,000 projects on the community website [3]. For many Scratch users the concept of sharing their work with others and getting feedback and advice from peers is the source of strong encouragement. The huge amount of work that is shared on Scratch serves as a library for exploring new ideas and learning new techniques in programming for others. Scratch website also serves as a platform for collaboration. Scratchers continuously build and adapt ideas form one anothers projects. Around 15% of the total projects shared on scratch are remixed projects meaning they were built upon some other project that some other scratcher has previously shared [3]. For instance there are a number of different versions of Tetris, where different scratchers have added new functionality and features to the game to try and improve the experience. Similarly many doll based projects which are adaptions from some original project.

There exists various different ways in which projects shared are discovered by other users. On the main home page are the sections like Featured Projects, Featured Studios, Projects Curated by, Scratch Design Studios, What the Community is Remixing, What the Community is Loving, this sections feature projects that are either hand-picked by curators, or are most trending projects in the community or the projects that are getting most attention(likes, loves, remixes etc).

However, when it comes to personalised recommendation there is no option which looks at what the user is interested in or based on users activities. Like in any online platform self-directed learning plays an important part in Scratch. For enabling suit- able self-directed learning it is important to guide the user through proper steps and resources[12]. By recommending projects based on users development we can better

Draft of 3:58 am, Tuesday, August 13, 2019 5

enable user with a self-directed learning attitude.

Thus this study focuses on developing an effective recommendation method for Scratch which would recommend users with the projects based on their preference and activities i.e. a personalised recommendation.

1.2 Research Objective

With the research goal in mind for Studying recommender system in Scratch platform, this dissertation focuses on the following objectives.

This dissertation focuses on the various content filtering techniques primarily col- laborative and content-based methods. We will study and apply some of these collab- orative technique on the Scratch data. The dataset that is made publicly available by the LLK Scratch team will be used for analysis and model building. We will try to fit an item-item collaborative model on the project datasets and compare the recommenda- tion outputs using standard metrics. We will also demonstrate how a recommendation would work on a scratch website by building a JS client/server model. Finally, we will discuss the risks and limitations of recommendations methods.

1.3 Research Challenges

• Filter the huge dataset available for finding the relevant and useful information.

• Find the best collaborative filtering algorithm that fits the available dataset.

• Finding the best metrics for Evaluating the generated recommendations for qual- ity and usefulness.

1.4 Dissertation Overview

Recommender system are all around us, from the product recommendation you get in Amazon to Spotify generating a customized playlist. This dissertation discusses the Collaborative filtering technique which is amongst one of the most common ways in recommender system research for calculating the predictions of items for a given user.

Draft of 3:58 am, Tuesday, August 13, 2019 6

There are many different techniques in collaborative filtering which can be applied depending on the requirements and the data. For predicting the Scratch projects a user might find interesting in Scratch, item - item collaborative filtering techniques is best suited. In item-item collaborative filtering, we first calculate the closeness of the project to other projects from our dataset, and then check what projects from this list the user is most likely to like.This is done by looking at the past activities of the user. The data is provided by the Scratch community website. It contains details of users, projects and the project contents that have been shared on the website. The data is collected over a period of 5 years from 2007 to 2012. The data is then filtered to search for the attributes that will be useful in our task for building a recommender system. We find that the most useful from the dataset is the users and the projects table, which contains information for our recommendation system analysis.

After fitting the item-item collaborative filtering model, for the purpose of demon- strating the recommendations system, a client/server model is built. JavaScript and Python have been used for the client-server interaction and Pandas library for data analysis and calculating the similarity matrix.

1.5 Dissertation Structure

The rest of this document is orchestrated as following. Chapter 2 discusses the current state of the art in the field of recommender systems and explorers the topic of recom- mendation system in education. It also looks at the previous studies done on Scratch with respect to Recommendation methods. Chapter 3 describes the methodology for collaborative filtering techniques and different prediction and recommendation gener- ation techniques. In Chapter 4 we apply the techniques discussed in previous chapter on the Scratch data. Chapter 5 analyses different evaluation methods for a recommen- dation system and also examines the recommendation results obtained in the previous chapter. In Chapter 6 we discuss our recommendation approach, its limitations and future work that is required. Chapter 7 concludes the study.

Chapter 2

State of the Art

2.1 Recommender system overview

2.1.1 What is a recommender system?

Often termed Recommender systems, they are simply a kind of content filtering tech- niques and tools used to provide users with the most accurate and relevant product suggestions filtered from a huge database of information. Recommendation system works by discovering information patterns from the dataset, learning user choices and yield results that are co-related with their requirement and likings.

The most common use of recommender system can be seen in the commercial applications. From recommending movies on Netflix to predicting user ratings on amazon, the use of recommender systems can be seen in various fields. Fields like financial services [13] and research articles exploration [10] and to find collaborators [14] have seen increased used recently.

Recommender systems are extensively being used in e-learning environments under the context of TEL [15] for improving the self-directed learning capabilities of students.

2.1.2 Information filtering techniques

Recommender system strategies generally use two types of techniques for generating recommendations one is content-based filtering approach and the other being collabora- tive filtering approach. Some times knowledge-based systems are also used (also known

7

Draft of 3:58 am, Tuesday, August 13, 2019 8

as rule-based). A combination of the above three techniques is also used commonly known as the hybrid approach. [16] categorises the recommender systems into 5 cate- gories: Knowledge-based, Content-based, Demographics based, Collaborative filtering, community-based and Hybrid approach.

Content-based method sometimes, also known as item based collaborative approach look at the individual items and their attributes to compute the similarity between the given item i and a list of items and then selects k most similar items [17].

Contrastively, Collaborative filtering techniques look at the users and their at- tributes. It calculates the closeness between all the users and then produces a recom- mendation based upon it.Basically, the idea of a CF-based algorithm is to look at the opinions of the like-minded users from the available pool and then recommend items based on it. The data of a users behaviour can be obtained in two ways explicitly or implicitly. Example of explicit data collection is the ratings and review a user gives to an item. Implicit data like the previous purchase history, browsing pattern of the user can also be useful in making a meaningful recommendation.

Knowledge-based techniques for recommendation takes into account the explicit knowledge about different things like the product, the user and some predefined rules to generate a recommendation. This is also known as rule-based recommendations [18]. The rule-based technique implements the relationship-rule discovery algorithm for finding the relationship amongst similar items and then produces a recommendation list based on the strength of the relationship [19].

For the purpose of this study, the focus is on the most widely used and perhaps the most effective technique [16] that is collaborative filtering.Collaborative filtering can be commonly categorised into two 1. User-based collaborative filtering and 2. Item-based collaborative filtering.

In user based collaborative filtering, the behaviour of other users is incorporated in the system to give more weights to the items which users like me have purchased or expressed and opinion about. The intuition being the more similar a user is to our target user, the more likely it is that the target user will like the items that other users with similar interests have liked.

Draft of 3:58 am, Tuesday, August 13, 2019 9

2.1.3 Challenges in user-based collaborative filtering algorithms

In spite of being amongst the most widely used algorithms in both small scale and large commercial-scale applications, the widespread use of user-based collaborative algorithms has shown some possible challenges with their use like:

• Scalability: Computational requirement for closest neighbour calculations in- creases with both the quantity of items and the users.

• Sparsity: In a large commercial use (like Amazon recommending books form large DB), even if a user has purchased ¡1% of the items(from millions of books) the nearest neighbour method won’t be able to generate recommendation of items for that user. Resulting in poor accuracy.

In conventional collaborative filtering recommender system, the quantity of work grows as the quantity of users increases. For producing accurate results quickly on a large dataset we will explorer the item-based collaborative filtering technique.

The item-based method first looks at the user vs item matrix to find out the rela- tionship between the various items present in the system and then utilize this infor- mation to implicitly calculate the recommendation for the user. [17] discussed various techniques in to compute the item-item similarity matrix (e.g. cosine closeness, Pear- son correlation etc.) and also different methods ( E.g., regression model vs weighted sum ) to obtain recommendations from them. Because the association amongst the item is relatively steady, item-based algorithms may be able to produce equivalent or sometimes superior results [17].

[17] studies collaborative filtering algorithms and their bottlenecks. Tapestry [20] shows the earliest applications of item-based collaborative recommender systems. The research in [20] relied on the direct inputs of users from the closed group, like an college working group. However, recommender systems for large groups like Scratch cannot rely upon every individual knowing each other one. In the course of time, many rating based automated recommenders were developed. The GroupLens research group at the University of Minnesota [21],[22] gives pseudonymous collaborative filtering solutions for the movies and Usenet. Video Recommender [23] and Ringo [24] are web-based music recommendation systems.

Draft of 3:58 am, Tuesday, August 13, 2019 10

2.2 Recommender systems in e-learning environ-

ments

2.2.1 E-learning space and recommendation

E-learning is a broad term that outlines an environment in which a person can learn at any time and anywhere using a computer, generally connected to a computer network. It is well established that e-learning can be as huge and valuable as the typical classroom experience or significantly more than it [25]. Using e-learning, one can learn and master the skills and knowledge just like he would in a traditional learning counterpart.

As e-learning systems begin to expand and are ever-expanding, the users need to first process a large amount of information before it meets their needs and provides them with items that are relevant to them. E-learning’s fast development has altered traditional learning conduct and presented both the students and teachers with the fresh scenario. Students often find it difficult to navigate through a huge number of exercises and courses, and teachers find it difficult to recommend students with materials. One of the answers for this information over-load issue in e-learning is a recommendation framework [26].

A personalized recommendation system in e-learning and online learning environ- ments provides learning suggestions to students. In [27] Lu discusses a framework for a personalised learning recommender system. [26] discusses a fully mature recommen- dation system architecture model for an e-learning framework.

In [28] we see the application of Hierarchical Clustering (HC) on a TEL dataset built with Coursera data known as DAJEE. Recommendation from this system was done on the basis of three educational entities, Instructors, Courses and Lessons. The main teaching context considered were Teaching preference of instructors, the course information and the lesson information.

[29] Discusses a hybrid approach for recommendation on MOOC platforms which uses social network analysis and association rule mining techniques. User contribu- tions on the forums and their social interaction and peer review activities were taken in to account to extract information for social network analysis. And then use this in- formation for collaborative filtering approach. This approach was again tested on the Coursera dataset which showed a well-performing system given the limited information

Draft of 3:58 am, Tuesday, August 13, 2019 11

available in the dataset.

2.2.2 Recommendations in scratch

In [18] the authors discuss recommender system in scratch with respect to a course framework for teaching Foundations of Computer Programming at Universidad Estatal de Milagro (State University of Milagro), Ecuador, with Scratch. consisting of various exercises that a student follows.

Benavides [20] proposes CARAMBA [30] a Scratch extension that recommends exercises, based on taste and program complexity. CARAMBA can recommend per- sonalized exercises for students. This again considers problem statements as exercises that would be recommended. It also conducts a study to evaluate the improvements in student activity as a result of the recommended exercises.

2.3 Evaluation of a recommendation system

Recommendation algorithms typically perform differently on different recommendation tasks and domains. Therefore it is critical from a practical and research perspective to select the proper algorithm that matches the domain and tasks that is of the in- terest. The standard way to make such a choice is by comparing different algorithms offline using some assessment metric. [31] discusses different matrics for evaluating recommending algorithms.

In [32] discusses various evaluation methods for a recommender system, particularly on the top N recommendations tasks. It proposes that while assessing algorithms in the top-N suggestion task, to pick the test set cautiously otherwise precision measurements are strongly one-sided.

The area of recommendation system research has utilised diffrent types of metrics to assess the quality of the recommendation framework. Mainly these are classified into two categories [17].

• Statistical accuracy metrics evaluation computes the correctness of the rec- ommendation framework by equating the value of generated recommendation score with the user given rating in the test dataset. One of the most widely used method for evaluating the statistical accuracy is called the Mean Absolute Error

Draft of 3:58 am, Tuesday, August 13, 2019 12

(MAE). MAE is calculated between the predictions and the user-ratings. MAE gives the deviation of the recommendation score from the user given score.

• Decision support accuracy metrics evaluation evaluates how effective and useful the predictions are at helping the user find the relevant item from the set of all items. These metrics consider the process of recommendation as a set of binary task, wither the recommendation is good or bad, i.e. the item is either predicted or not considered at all. With this in mind, the prediction can be anything like 1 or 2.5, it doesn’t matter as the user only selects items with a rating higher than 4. Most common techniques in this category are the ROC sensitivity and the reversal rate [33].

This techniques are covered in detail under chapter 5 Methods of Evaluation.

2.4 Other techniques in recommendation systems

Additionally, other techniques have been applied to the recommender system area, like Clustering, Horting and Bayesian networks,

Bayesian networks work by creating models over a training dataset associated with a decision tree like structure. Each edge and node of the tree represents knowledge about the user. The model can be trained and completed in the time as less as hours or a few days. This model tends to be very small and very fast and can be as accurate as the nearest neighbour methods [34]. Bayesian can be used practically in environments in which the information about the user preferences changes steadily in reference to the time which the model building requires. But Bayesian networks have been known to fail in areas where user preferences change very rapidly.

On the other hand, clustering algorithms work by classifying users with similar preference into different groups known as clusters. After the clusters have been final- ized, predictions for a user are calculated by taking the average of the opinions of other users in the cluster. Some of the clustering techniques consider users in the form of participation across several clusters and the averaging is done by the weighted sum of the degree of the participation. Clustering is known to produce non-personal results in the recommendation, in fact, in some cases, it may give even worse accuracy than the nearest neighbour algorithms [34]. However once the clustering is done, there is a

Draft of 3:58 am, Tuesday, August 13, 2019 13

large increase in performance because the comparing needs to be done on very small clusters. Clustering techniques can also be used in conjunction with nearest neighbour algorithms to reduce the size of the dataset that must be analysed.

Horting is a diagram-based methodology in which the nodes of the graph repre- sents the users and the connections between them are edges representing the degree of closeness between the users [35]. Recommendations are given by traversing the graph to the nearest node and averaging the opinions of nearby nodes (i.e. users). Horting and nearest neighbour differ from each other in the sense that in Horting graph may be traversed through users who have not had an opinion about the item and therefore considering the changing relationships that nearest neighbour does not consider [2]. In a study with artificial datasets, Horting has shown superior performance to the nearest neighbour algorithms [2].

Chapter 3

Methodology

In the previous chapters, we investigated what are recommendations system and the different techniques applied in the recommender system and also their application in the e-learning domain. This chapter will describe the process in which an item- based collaborative filtering algorithm was applied on the Scratch data to build a recommendation system. We will look at different steps involved in an collaborative filtering methods, various item-similarity computation methods and also the prediction calculation steps.

3.1 Summary of the Collaborative Filtering process

The objective of a collaborative filtering algorithm is to generate a list of recommended items or to estimate the likeliness of the item for a particular user based on the past activities of the user. In an standard CF scheme there is a list of m number of users say U = {u1,u2,u3,u4, ..., um} and a list of n items I = {i1,i2,i3,i4, ..., in} . Every user ui has a list associated with the them consisting of items IUi , which the user has expressed interest in or has impression about. In the case of Scratch we look at the projects the user has liked and the projects user has created/published in the past. This impressions or likings can either be collected explicitly or implicitly. Notice that the item list for user IUi ⊆ I and its likely that the IUi might be an empty set (e.g. when the user newly enters the system). Also, there exists a user ua from the list of users for whom the task is been carried out. The goal for CF algorithm is to discover

14

Draft of 3:58 am, Tuesday, August 13, 2019 15

an object (in this case projects) likeliness i.e. how likely the user is to view or love this item. This can take two forms.

• Recommendation R is a list of items such that Ir ⊂ U which contains the items the user will like the most. It is to be noted that the R should avoid items that the user is previously associated with i.e., for example, the projects that the user already has on his page.

• Predictions is a number Pa,j representing the predicted likeliness of the item for the user ua. The predicted value is same as that of the normalized scale of the user ratings [17]. For the case of this study in Scratch we consider this to be the count for remixes, likes and favorites for the user-item pair.

Figure 3.1 shows a diagrammatic representation of the collaborative filtering pro- cess. The CF algorithm shows the whole m × n matrix A as a user vs item matrix of the user ratings. Each cell ai,j in the matrix holds value for every item the user has rated i.e. either liked, favourited or remixed. Every rating is on the numerical scale and can take the value 0 representing the user has not yet given the rating ( i.e. liked or favourited or remixed the project) the project or a value 3 (done all three activity) indicating user liked the project very much.

Figure 3.1: The process of a Collaborative Filtering Algorithm

Draft of 3:58 am, Tuesday, August 13, 2019 16

3.2 The item-based collaborative filtering algorithm

Unlike the user-based collaborative filtering algorithm shown above, an item-based col- laborative filtering algorithm considers the semantics of the items a user has interacted with (liked, remixed or added to the favourited) to compute the likeness between two things i.e. how closely related items {i1,i2,i3,i4, ..., in} and the item of interest are. Then it selects the k most closely related items {i1,i2,i3,i4..., ik}. Once the task of computing similarities is complete the predictions is calculated by taking the weighted sum average of the target users ratings on these similar items. This process of similarity calculation and prediction generation is detailed below.

3.2.1 Similarity computation

One of the most critical steps in an item-based collaborative filtering model is the calculation of the similarity matrix between the set of items and then selecting the most similar items. This similarity calculation step between two items i and j, basically performs two steps first it separates the users who have evaluated both the items and then applies a closeness calculation strategy to decide the similarity si,j Figure 3.2 represents this procedure. Here the matrix rows denote the user and the columns denote the items.

Figure 3.2: Computation of item-item similarity

Draft of 3:58 am, Tuesday, August 13, 2019 17

Various techniques have been proposed to compute the similarity between any two items in statics[19]. Some of the most common and well known once are Correlation- based similarity, Cosine-based closeness and Adjusted-cosine based closeness. We will examine each of this technique and apply them on our dataset to check which of them suits best for the choosen dataset.

3.2.2 Cosine based similarity

Cosine similarity is defined as the measure of similarity between any two non-zero vectors. These vectors reside in as inner product space. Cosine similarity is defined as the cosine of the angle between them. The cosine for a vector with the angle at 0 degrees is 1 and the for any angle between 0 and 90 is less than 1. Therefore it gives us information about the orientation of the vectors and not their magnitudes. Two vectors with orientation in the same direction have a cosine 1 and the cosine for two vectors that are at an angle of 90 degrees with respect to each other is 0. Similarly, two vectors within an exact opposite orientation will have a cosine similarity as -1 irrespective of their magnitude. Cosine similarity mostly is used in positive spaces where we require the output to be neatly bounded between [0,1]. Particularly data mining, information retrieval and text mining all of which have high dimensionality problem extensively use this technique. Each of the terms is assigned a new dimension to calculate the cosine similarity between them. The cosine similarity is useful because even if two vectors are very far apart in terms of Euclidean distance (due to their sizes), there is a chance that they may be similarly oriented closer to each other. The smaller the angle between the two, the higher is their cosine similarity. For this study, we consider two items as the two vectors in the m dimensional user-space. The similarity measure is calculated by computing the cosine of the angle between these two vectors. In Figure 2 we see the similarity between items i and j in an m x n matrix for user projects. The similarity is shown by sim(i, j) such that

sim(i, j) = cos( i, j) = i. j

i 2 ∗ j 2

In the equation above, dot product is represented by ”.”

Draft of 3:58 am, Tuesday, August 13, 2019 18

3.2.3 Correlation-based similarity

In statics, correlation is a statistical association between two variables, showing the degree to which the two are linearly related to one another. There are many correla- tion coefficients defined which show the degree of closeness between variables. Pearson correlation, Interclass correlation, Spearman’s rank correlation are some of the exam- ples of correlation coefficients used in statistics. Pearson correlation is one of the most widely known and used matrics to show the similarity.

Also, know as the person-r correlation coefficient, it gives us the linear correla- tion between any to variables X and Y . It can take a value between +1 and −1, where 1 denotes a total positive correlation between the variables, 0 denotes abso- lutely no correlation between the variables and -1 denotes a total negative correlation [36]. Mathematically it is defined as the co-variance of the two vectors divided by the product of their standard deviation. In a correlation-based similarity scenario, the like- ness amongst two items i and j is computed by calculating the Pearson-r correlation corr(i, j). For a correlation to be accurate the cases where the users have rated the projects same will need to be separated.

sim(i, j) =

(Ru,i − Ri√∑∑)(Ru,j − Rj)

u∈U

(u∈U

Ru,i − Ri)2√∑u∈U

(Ru,j − Rj)2

3.2.4 Adjusted-cosine based closeness

One basic contrast between the closeness calculation in a user-based collaborative fil- tering and item-based collaborative filtering is that in the item-based method the sim- ilarity is calculated based on columns of the matrix, whereas in user-based method the similarity is calculated along the rows of the matrix i.e., each pair in the co-appraised set relates to a different user(Figure 3.2). Computing similarity utilizing fundamental cosine measure in the item-based case has one significant downside - the distinctions in the rating scale between various users are not taken into consideration. The adjusted- cosine based similarity reduceses this disadvantage by subtracting the average score from every co-evaluated pair. The formula for this is given as below

Draft of 3:58 am, Tuesday, August 13, 2019 19

sim(i, j) =

(Ru,i − Ru√∑∑)(Ru,j − Ru)

u∈U

(u∈U

Ru,i − Ru)2√∑u∈U

(Ru,j − Ru)2

Where Ru is the normalized score for u th user.

3.3 Prediction Calculation

One of the most significant steps in any CF-based system is to give the output in terms of the prediction. Once we have with us the list of items that are most similar as given by the closeness measures above, the next step is to examine the user ratings and apply a predictive method to get the prediction. Various ways have been proposed to compute the prediction. Regression method and Weighted Sum method are the two such most common technique [19].

3.3.1 Weighted sum method

In mathematics, a weighted function is the method used when performing an average, a sum or an integral to assign some components more weight i.e. importance than others to increase or decrease their impact on the result. Weighted sums are most commonly used in statistic to decrease or compensate for the bias present in the data. For example, if a quantity F is measured multiple times independently with variance σ2, we can find the best approximation of this measure by taking the average of all the measurements with the weight ωi = σ12, giving us a smaller variance than each individual measurement.

This method computes prediction for project i for user u, by calculating the weighted sum of rating attributes (likes, favourites, remixes etc.) for a similar item. Formally this can be shown as the prediction Pu,i as below

Pu,i =

∑all ∑similar all similar items,N items,N (Si,N (| S∗ i,N R|) u,N) Basically, this method attempts to show how a target user u, is likely to view (favourite it, like it or remix it) based on the similar items that he has seen (liked, loved, or remixed) in past.

Draft of 3:58 am, Tuesday, August 13, 2019 20

3.3.2 Regression model

In statistics, regression modelling is a collection of procedures to estimate the depen- dency relationships between the variables. Regression analysis aids us in understanding the effect that the dependent variable has when any one or multiple independent vari- ables are changed. Number of regression techniques have been invented to support various statistical problems. Linear regression, polynomial regression, principal com- ponent analysis etc. are some of the popular techniques.

Regression model for calculating the prediction is mostly similar to the weighted sum method discussed above. The only difference is that instead of using ratings of similar items, this method approximates the ratings of items based on the regression models. When similarity is calculated using either cosine or correlation measurement, it may lead to improper conclusions in practice because the two vectors may be at distance in terms of euclidean distance, but may be highly similar to each other in terms of orientation. In this case, using the ratings from similar items can result in low-quality predictions. The general idea is the same as that of the weighted sum method, but instead of using raw ratings from the N similar items, this method uses their approximate values which are calculated using a linear regression model.

General formulation of a CF-algorithms and the necessary methods and steps re- quired for performing item-based collaborative filtering have been discussed till now. In the next chapter, all these algorithms will be applied to the actual Scratch dataset.

Chapter 4

Implementation

For implementing the item-based collaborative filtering methods on Scratch, we first look at the data that is required and the available datasets from Scratch.

4.1 Collecting the data

The Scratch Online Community has made the dataset for scratch platform publicly available on [37]. The dataset has the data for the first five years of the data from Scratch (approximately from 2007-2012). The data was collected from the MySQL database from the Scratch website. All datasets are provided in CSV file formats, which can be loaded into any compatible data analysis software. There are a total of 32 datasets which contain information about various aspects of the Scratch Online Community. These 32 datasets are divided into 3 main categories Core Datasets,Text and Code Datasets and Project Analytics Datasets. Table 4.1 shows the description of each and every table that is available in the dataset taken from [37].

4.1.1 Data description

• Core dataset consists of meta data and data tables that define major relation- ships and objects which are captured by the Scratch Online Community applica- tion.

• Text and Code datasets consists of very large tables consisting of texts sub-

21

Draft of 3:58 am, Tuesday, August 13, 2019 22

mitted by the users. This can be helpful for NLP and text-based analysis.

• Project Analytics datasets contains tables which have quantitative summaries of each project file. A project file is a .sb2 or .sb3 file that contains all of the content of a scratch project.

4.1.2 Ethical implication

The dataset that this research uses is the data from the Scratch Online Community. It has been supplied by the Lifelong Kindergarten Group at the MIT Media Lab under public domain. Although this data is available under public domain, a Scratch Research Data Sharing Agreement (SRDSA) [37] was signed to obtain access to the dataset files. The user privacy is fully protected while scraping the data. The rules that define the user privacy such as only information that was publicly visible on the site to anyone visiting the site even if it may be a web crawler software and no action was required by the user or the site administrator were used while scraping the dataset. Some of the user sensitive information like the self reported gender,age email, IP address was omitted while publishing the dataset. [38] The users of the Scratch online platform are mostly young children many below 18 years of age, from all around the globe. Therefore an ethics approval was required by the LLK group from the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). Which was granted to the team and the data was published under the specified protocol by (COUHES) MIT. All the terms of the access to the data have been adhered to and therefore as the part of this dissertation, no ethics approval was required from the university.

4.1.3 Identification of the relevant dataset

Keeping the objective of the research in mind of building a recommender system that generats recommendation on ”projects” for a user, first we need to identify the relevant data for our goal form this dataset.

For recommending projects based on user activity, we require the data related to projects and users on the Scratch platform. Looking at the list of the datasets as shown

Draft of 3:58 am, Tuesday, August 13, 2019 23

Draft of 3:58 am, Tuesday, August 13, 2019 24

in Table 4.1 we observe that projects.csv, projects blocks.csv, projects sprite.csv and users.csv have most of the data related to users and projects.

• Users - holds the data about the user and which project the user has worked on.

• Projects - contains general information about the projects and the user associ- ated with each public project.

• Projects block - holds the information about each block (the logical component of the code) used in a project, these are the extracts of the sb2 and sb3 files.

• Projects sprite - contains the information about sprite (scripts, sounds and images) used in each project.

Apart from these tables, the following datasets were also used to obtain information required for user rating calculation.

• Lovers - which has information about the projects that the user has clicked Love for.

• Friends - contains each user and its friends ( the user who follow each other).

• Favourites - shows an event where a user has added a project to their Favourites.

4.2 Data Analysis

After identification of the data required, we apply some preliminary data analysis tech- niques to check for consistency and quality of the data.

Projects - We see that the more of the projects are created in the latter half of the dataset time frame i.e. after 2010. This shows the popularity of the Scatch in the initial phases.

Draft of 3:58 am, Tuesday, August 13, 2019 25

Figure 4.1: The distribution of projects in dataset over the period of 5 years

Project Blocks - There are about 171 of these key block attributes associated with each project in this dataset.Table 2 shows each of this block. Figure 4.2 below shows the frequency of each block used in the projects. We see an even distribution of the most common operators in the projects, while the less common once are not used frequently. We can assume from this distribution that project dataset is good enough for finding the correlations between projects.

Draft of 3:58 am, Tuesday, August 13, 2019 26

Figure 4.2: The frequency of the use of project blocks

Draft of 3:58 am, Tuesday, August 13, 2019 27

Users -The following figure 4.3 shows the distribution for the accounts in scratch done by creation date over the period of 5 years.

Figure 4.3: Distribution of the user accounts created by creation date

Friends - An event where a user adds another user as a friend on the Scratch community website is recorded in this dataset. The figure 4.4 shows the distribution of this relationship over time.

Figure 4.4: Distribution of the friends over time

From table 4.1 we observe that that on an average 6.32 projects have been created per user with a SD of around 26. And on an average a user has around 11 Friends per user with a SD of 55. The variance may seem huge but considering the size of the data which contains around 1.3 million records for friends and 1.9 million records

Draft of 3:58 am, Tuesday, August 13, 2019 28

in the projects dataset, the variance is expected. The high variance values in both datasets indicate number of missing values from the data, this will be handled in the data cleaning and imputation section 4.3.2.

Observations Mean Var SD Max Min NAs Projects 6.32 721.9964 26.86999 3047 1 0 Friends 11.31493 3073.802 55.44188 4981 1 0

Table 4.1: Mean, Variance and Standard deviation for Friends and Projects data

4.3 Calculating similarity matrix

After the data analysis, we will apply the three similarity computation matrix discussed in chapter 3 on our projects dataset. All the programming and computation is done in python as programming language. Various packages and libraries have been used to facilitate the easy application of diffrent algorithms. Pythons sci-kit learn (sklearn)[39] library, which houses many machine learning packages has been used for applying the cosine- correlation and adjusted cosine correlation methods. Another library known as Pandas [40] has been used for computing the Pearson correlation matrix.

4.3.1 Preparing the data

Since our project sprite and project blocks both contains some information about the project components we can merge the two datasets to form single dataset. This has two benefits first it reduces the process of applying computation on two differ- ent datasets and secondly it gives us 3 more parameters(sounds, images and scripts) from project sprite which will add to the existing 171 parameters from project block dataset to calculate the correlations.

1 #Load projects data from csv 2 projects\_df = pd.read\_csv(’data/CSVs/project\_blocks.csv’, sep=’,’,

index\_col=0,)

34 #Imputing the others data since it has mostly the null values. 5 projects\_df = projects\_df.drop(columns="other") 6 #Load projects\_sprites data from csv

Draft of 3:58 am, Tuesday, August 13, 2019 29

7 project\_sprites\_df = pd.read\_csv(’data/CSVs/project\_sprites.csv’, sep=

’,’,index\_col=0)

89 #Merge both the panda dataframes on project\_id 10 main\_frame = pd.merge(projects\_df, galleries\_df, left\_index=True,

right\_index=True)

4.3.2 Imputing missing data

As we have seen in our previous data analysis section, there are some missing values in our dataset. For our algorithm to work correctly we need to replace this missing data. For replacing the value we need to apply some data imputation techniques which estimates the best possible replacement by looking at neighbouring values.

• Mean imputation - In the mean imputation methods, the missing values are replaced by the mean of the observed variable. For example, if the variable of sound ( number of sounds used in the project) is missing for a particular project, we take the mean of all the projects who have sound in them and replace the missing value with it.

• KNN imputation - In K nearest neighbour imputation the missing values are replaced by the approximate value based on the nearest points. The assumption being that the given K nearest neighbour point the missing value is an approx- imation of it, k being a number of closest points to look for and is supplied by the user it is the most commonly used technique.

For our purpose we use SimpleImputer and IterativeImputer from sklearn for imputing the data values.

12 #Impute the missing values using IterativeImputer from sklearn 3 MAX\_ITER = 10 4 imp = SimpleImputer(missing\_values=np.nan, strategy=’mean’) 5 imp = IterativeImputer(max\_iter=MAX\_ITER,initial\_strategy=’knn’,

random\_state=0) 6 imputed\_DF = pd.DataFrame(imp.fit\_transform(main\_frame)) 7 imputed\_DF.columns = required\_columns\_df.columns 8 imputed\_DF.index = required\_columns\_df.index

Draft of 3:58 am, Tuesday, August 13, 2019 30

4.3.3 Correlation-based similarity

For correlation-based similarity computation, Pearson-r correlation method is applied. The corr(method = pearson ) methods from the Pandas directly applies the selected correlation method on the dataset. The correlation matrix is obtained as shown in Figure 4.5. This matrix shows the item correlation between all the projects from the dataset.

Figure 4.5: Pearson correlation output

Note the numerous negative correlation values in the matrix, this suggests that many of the projects are not similar at all.

4.3.4 Cosine-based similarity

For cosine based similarity computation we apply the cosine correlation from the sklearn.metrics.pairwise module on the dataset. Figure 4.6 shows the output matrix for cosine based similarity, This matrix shows the cosine correlation values for all the projects in the dataset.

Draft of 3:58 am, Tuesday, August 13, 2019 31

Figure 4.6: Cosine similarity computation output

4.3.5 Adjusted-cosine based similarity

Similar to the previous method after appliying the adjustedcosine functions from the sklearn library, we obtain the similarity matrix shown in figure 4.7. The matrix shows correlation values of different items from dataset.

Figure 4.7: Adjusted cosine similarity output

Note here that since adjusted cosine is performed over columns rather than rows, the computation time for this method is slower compared to the other two. This computation time can be reduced by using a machine of higher computation memory or by reducing the size of the dataset

4.3.6 Selecting top-N recommendations

From the results of the similarity obtained above, for a given item i of a particular user we get the correlation score for each of the similar items from our dataset. We select the top - N values from the matrix (N=5) to be presented to the user U. This gives us the top 5 similar items to the user item I.

1 #Get Top 5 from the similarity matrix 2 top5 = final\_df.loc[project\_id].nlargest(5) 3 recommendation\_pairs = top5.to\_dict()

Draft of 3:58 am, Tuesday, August 13, 2019 32

Draft of 3:58 am, Tuesday, August 13, 2019 33

4.4 Prediction Calculation

After we have the item-item correlation matrix ready, we proceed to calculate the prediction score for each of the item based on users activity.

4.4.1 Assigning the rating score for user activities

In a conventional personalised recommendation, a users feedback is captured through the likes and the rating given to an item. For example in case of Netflix a user rates a movie on a scale of 1 to 5, 1 being the lowest and 5 being the highest. Some times decimal ratings are also allowed. Since there is no rating system in scratch and a user can either Like, favorite or remix a project and these three details are publicly available we will use this information to develop a rating system. While viewing a project , a user has a choice of performing various actions on the project page. He/She can either click on Love/Like to show the appreciation or add the project to his favorites or click on remix to copy the project to his own profile and adapt on its code. User can perform one or two or all three of this activities.For each of the user activity (love, favorite and remix) we assign a rating score of 1 if the user has performed the activity or 0 if not. We then take the sum of all three activities to find a user rating on a scale of 3. So for example, if a user has liked and favorited a project but not remixed the total user rating is love (1) + favourite (1) + remixed (0) = 2. And If a user has only remixed the project then the total rating for that project is love (0) + favourite (0) + remixed (1) = 1. Henceforth in our weighted sum method we use this rating to find the most relevant projects for the user.

4.4.2 Weighted Sum method

Weighted sum method has been used for prediction calculation because of its simplicity to implement. For an item I for user U, this method calculates the prediction score by evaluating the sum of the ratings given by the user on other similar items to item I. Each rating by the user is assigned a weight si,j based on the similarity output of the item.

1 #Calculates the prediction score for a user given recommendation as

the list of similar recommended projects and target project\_id

Draft of 3:58 am, Tuesday, August 13, 2019 34

2 def calculate\_prediction\_score(user\_id,project\_id,

similar\_recommendation): 3 \_w = [get\_user\_ratings(user\_id)] 4 \_x = similar\_recommendation 5 score = sum( \_x \* \_w for \_x, \_w in zip( x, w ) )/ calculate\_score(\_x

) 6 return score

Figure 4.8: Top 5 project ids according to prediction

The figure shows the top 5 most similar items computed after applying the weighted sum method. For the supplied user id: xxxxxxx and project id: 2437714 all the other are the recommendation scores for the top 5 similar items.

4.4.3 Comparing the recommended items

On comparing the results that we get from the recommendation method we observe the following. We follow the link of each project ID to see the structure of the project. Every project comprises of a similar set coding blocks as of the original project by the user U. For instance every project has one forever loop, one wait timer one costume switch. Also, note the the two most similar projects have the sound block explicitly specified and while the next two have an implicit sound build in requiring no sound block. This gives us confidence about the similarity matrix outputs in the N dimen- sional user space. It is also observed the second most similar project (id=2437719) is given priority because it was liked and remixed by our test user.

Draft of 3:58 am, Tuesday, August 13, 2019 35

Project URL Code Snip https://scratch.mit.edu/projects/2437714/

Likes: 0, Favourite: 0, Remixes: 0

Table 4.2: Projects recommended

Next chapter will deal with evaluating and comparing the quality of these recom- mendations methods and prediction scores more formally. It discusses various different matrics for evaluating the recommendation methods and prediction scores.

https://scratch.mit.edu/projects/2437719/

Likes: 2, Favourite: 4, Remixes: 2

https://scratch.mit.edu/projects/2437771/

Likes: 3, Favourite: 3, Remixes: 0

https://scratch.mit.edu/projects/2437735/

Likes: 0, Favourite: 0, Remixes: 0

Chapter 5

Method of Evaluation

This section discusses various evaluation methods that exists in the recommender sys- tem research paradigm to determine the quality and accuracy of the recommendation method discussed in previous chapter. We also compare the recommendation system based on the three similarity computation algorithms discussed previously.

5.1 Experiment

After building the model and applying the methods on our training dataset, we apply the methods on our test dataset for comparing the accuracy. From the dataset of over 1.9 million observations of projects and around 1 million users, the data set was divided into 80% as training data and 20% was used as test data for the model.

5.2 Evaluation methods

Various methods of evaluation have been studied in the paradigm of recommender system research. These methods are mainly used to evaluate the accuracy of a recom- mender system. These metrics can be broadly categorised into 2 groups as described below .

36

Draft of 3:58 am, Tuesday, August 13, 2019 37

5.2.1 Statical accuracy metrics

A statistical accuracy measure is the one which evaluates the accuracy of a recom- mender system by calculating the difference in the actual user given score in the test data and the score calculated by the recommender system.

• MAE(Mean Absolute Error) is one of the most popular metrics that is widely used to compare the actual and expected values. Basically, MAE is the measure of the deviation of the predicted recommendation rating from the actual user given ratings. In the scratch dataset for each pair of {predicted, user − rating}, MAE handles the absolute difference between them equally i.e. it is given by | redicted−user rating |. MAE is computed by summing the absolute difference for all the predicted, user-rating pairs and then averaging it over the total number of pairs. Mathematically it can be written as,

MAE =

∑(predicted ratings − user ratings) Total number of pairs

• Root Mean Squared Error (RMSE) is another statistical accuracy matric that is used for measuring model performance. It was first used in the areas of environmental changes and climate research [41]. While MAE assigns the same weight to all the errors, RMSE penalizes large variance in the reading as it assigns errors with large absolute value more weight compared to the smaller values.

MAE =

√√√√n

1∑i=1

ne2i

5.2.2 Decision support accuracy metric

Decision support accuracy metrics are used to measure the usefulness of a recommen- dation engine at assisting the user in selecting the top-quality items from a list of all available items. They work by assuming the predictions to be a binary process. Either an item is a good one or a bad one, i.e. it is either predicted or excluded from predic- tion. If on a rating scale of 5 if a user only chooses items with a rating more than 4, it becomes irreverent if the rating is 2.5 or 3. Some of the most widely utilised support

Draft of 3:58 am, Tuesday, August 13, 2019 38

accuracy matrices are ROC sensitivity, weighted error and reversal rate [33].

• ROC sensitivity - Receiver operating characteristics curve is a graphical plot that is used to visualise and select classifiers based on their performance. ROC graphs are being used since long time in the area of signal detection system for showing variance in a hit and false rates of the classifiers [42]. The medical community has a huge literature on the use of ROC for decision making on diag- nostic testing. ROC curves are adopted in machine learning for comparing and evaluating various algorithms. ROC graphs have recently gained popularity in the machine learning community due to the understanding that simple accuracy metrics are a weak measure of the classifiers performance [43].Besides of being useful as a performance mapping technique, they have additional qualities that make them suitable for problems with very skewed data distribution.

We begin with building a confusion matrix for the classifier, counting the number of correctly classified and incorrectly classified values into the matrix. As shown in figure 5.1 the true positive and true negatives are the correctly classified values by the classifier and false positive and false negative are the incorrectly classified values.

Figure 5.1: Performance metrics calculations from Confusion metrics.

Figure 5.1 shows various standard metrics that can be calculated from the confu- sion matrix. Formally it is defined as the area under the curve of the ROC curve.

Draft of 3:58 am, Tuesday, August 13, 2019 39

ROC is plotted as sensitivity v/s (1- specificity) of the test. In case of a rec- ommender system Sensitivity is defined as the probability of a random selection of item is accepted by the filtering method (a good prediction) and specificity is defined as the probability of a random selection of item being rejected by the filtering method (a bad prediction). A good filtering method would enable the user to select from 90% of the good recommendation and 10% of the bad ones.

• Reversal rate is defined as the frequency at which the recommendation by the system is wrong or incorrect. On a scale of 5, it is commonly given by the percentage of the recommendation where the prediction was varied by 3 or more points.

• Weighted error is the metrics which gives extra weight to the errors due to the strong outlook about the recommended item. For instance, if a user considers an item to be his favourite (i.e. a rating of 5 out of 5) the error for this could be considered double or more.

Some of the new and novel approaches have been discussed in [44] which are based on coverage and serendipity of the recommendations. It discusses the matrics beyond accuracy of the recommendations and takes into consideration the quality and useful- ness of the system

For the simplicity and ease of interpretation of MAE, we select it as our preferred choice of evaluation. In the experiment [33] Sarwar notes that MAE and ROC give the same arrangement of different experimental procedures with respect to prediction quality.

5.3 Results

5.3.1 Comparing the similarity algorithms

After running the three similarity computation methods namely cosine, adjusted cosine and Pearson-r correlation as discussed in previous chapters on our training data and using the weighted sum method to get the predictions score, we calculate the MAE of

Draft of 3:58 am, Tuesday, August 13, 2019 40

for each of the methods. Figure 5.2 shows the results of the three algorithms on the dataset.

Figure 5.2: Performance comparison of similarity calculation method

Since lower value of MAE denotes a better performance, we observe the best perfor- mance is given by the Adjusted cosine similarity method. However since the computa- tion time required for the adjusted-cosine matrix calculation is high and the difference between Cosine and Adjusted cosine is relatively less we can say Cosine similarity is best suited for the Scratch dataset. Note that since the Person-correlation matrix gave us many negative values its MAE score is higher compared to other two methods.

The client server model in appendix A which was built to demonstrate the working of recommender system uses cosine similarity implementation

Chapter 6

Discussion and Future work

Although we have successfully displayed the working of the proposed recommenda- tion methods, there are many more ways in which the recommendation system for a complex platform like Scratch could be developed. In the design of this item-based recommendation, we assume the item-item similarity and user rating are the defining characteristic of the user preference, however that may not be necessarily the case. In this chapter, we discuss the limitations of the current system and the possible im- provements that could be made to improve the performance of current recommender system.

6.1 Limitations

6.1.1 Over specialisation problem

One of the major problem faced by collaborative filtering algorithms is the problem of over specialisation. A classic example of this is when one buys a lawn mover on amazon, he only sees recommendations for lawn mover. A good recommender should provide suggestions on a diverse set of items, rather than suggesting the same type of items that the user is already acquired or liked in past. The recommendation system proposed in this study suffers from this problem i.e. if a user has liked a certain type of project in Scratch, the recommended are generated for the project that has a similar characteristic. It prevents the user from discovering new items that might help them. Researchers have tried to solve the problem by neighbourhood based CF and hybrid

41

Draft of 3:58 am, Tuesday, August 13, 2019 42

recommender systems [45].

6.1.2 Concentration bias

A concentration bias arises when a recommender system proposes the items with higher ratings more than with low ratings. Specially in CF algorithms which use user rat- ings to recommend the items, they often create the ”rich gets richer” effect for more well-known products, this concentration bias limits the reach of lesser known items to user. Various researchers have tried to eliminate concentration bias problem from classical recommendation algorithms like k-NN filtering [45]. [46] suggests a a proba- bilistic neighbourhood-based method popularly known as k-PN as an alternative which consistently outperforms k-NN in terms of concentration bias.

6.1.3 Cold Start problem

A cold start problem occurs when a new user or item enters the system and there is no information about it. Cold start problem can be broadly divided into three types new item problem, new user problem and new system problem [47]. In the case of a new user, it is very difficult to give a recommendation as there is no information available about the user. In the case of projects in scratch when a new item enters the system there is some information about the project, hence we can at least compute the similarity matrix for the item. But without a user rating, we cannot calculate the prediction score. Content bases systems succeed in case of the new item as they do not depend on the user for generating the recommendations.

6.1.4 Dataset limitation

Over the years scratch has gone through many iterations of development. The first version (Scratch 0.x to 1.x) of the scratch was implemented in Squeak [48]. Later when Scratch 2.0 was launched in 2013 it shifted to ActionScript. The current version of scratch launched on January 2nd, 2019 i.e. Scratch 3.0 uses javaScript for its imple- mentation [49]. This dataset was scraped at the initial phase of Scratch from 2007 to 2012, hence it only contains projects from Scratch 1.x. The development has led to many UI changes to the scratch development workspace

Draft of 3:58 am, Tuesday, August 13, 2019 43

over time. UI may have a significant influence on user activity and skills development [50]. The quality of projects and user data might have been affected by this effect. Also , many new attributes and features have been added to the newer version of the project files that give us the parameters for item similarity computation. Hence the more recent data collections effort is needed in order to accurately calculate the item similarity between projects. A more recent dataset would give a better understanding of the user’s mindset and may probably give a different result for different similarity algorithms. The dataset also provides very little information on the user rating part. Some implicit information like the clicks the user makes and the changes a user makes to different projects may help in gaining more knowledge about users preference.

6.2 Future work

6.2.1 Qualitative testing

As discussed in the chapter 5 we evaluated and validated the prediction accuracy for the proposed recommendation techniques. Research in the recommendation system space mainly focuses on the prediction rating accuracy. And since accuracy only partly comprises the user experience of the recommender system, more qualitative testing is required for evaluating the usefulness of the technique. It needs to be verified against the need of the user who is getting the recommendation. One of the goals of developing this recommendation system for Scratch was to aid the user in their learning journey by suggesting projects that might improve their skills and the techniques they are learning. Some of the methods that could be used are an A/B testing amongst the users, collecting user data about the recommendations or by surveying the users for their opinion about the recommended items.

A/B testing

Also known as the bucket test or split-run test is a method to compare two versions, A and B of the same technique, commonly done on users response to the different versions to determine which one is more effective. Using the A/B testing methods different versions of recommendation can be served to users to see which one is more useful to

Draft of 3:58 am, Tuesday, August 13, 2019 44

the user. For example, if the user chooses to view a more complex project instead of the more similar one to the original, this could give an insight of how to develop a recommendation which will help the user in discovering more complex projects.

User Surveys

A user survey could be conducted to get an understanding of what the user thinks of the recommendations given to him/her. This could help us understand the user needs and performance parameters upon which we could tune the recommendation system. As noted in [49] preference elicitation is a difficult problem since many time a user is unaware of his own personal preference, especially when he is beginning to use the system or is unaware of all the available choices. [49] also suggests asking users a small set of goals in the signup process to understand the user’s recommendation parameters for the individual.

Many novel methods have been proposed in [51] which focuses on the user-centric de- velopment of the recommendation systems. It proposes a framework that considers how the objective components of the system e.g. algorithms, are perceived subjectively and how these perceptions along with situational and personal characteristics, gives a dif- ferent user experience leading to different interactions with the system. For example, if the user perceives the recommendation quality of these different algorithms differently, a higher perceived quality of the recommendation leads to higher interaction with the system.

6.2.2 Mining text from the dataset

While exploring the available dataset, we found that the text and code dataset con- tains much interesting information that might be useful in building a stronger rec- ommender system. The comments posted on a project and galleries are available in pcomments text and gcomments text datasets. Projects strings contains information about all the user-generated strings that have been used in the project. All these text- based information could yield some meaningful information if explored correctly. A

Draft of 3:58 am, Tuesday, August 13, 2019 45

careful sentiment analysis of the comments might help us understand the kind of re- sponses a project is getting. Based on which we could infer project usefulness to various users. In addition to this, comments on different project by the same user could help us understand users personal taste which could help in predicting user choices. Strings inside projects could help in categorising the project into various groups and sections, which in turn could help in building a cluster-based recommendation. Also, there is information available about the tag text which is the tags assigned to the project by different users. This could also help in categorising the projects. TFIDF algorithms could be used to find the frequency of terms and then classify the projects into different categories.

6.2.3 Social network analysis

Scratch also has a large community website and a community forum where users inter- act with one another, asks for help and post useful stuff. Users can follow one another if they find the profile interesting. A social network graph analysis can be done on this information to find users with similar interests and user groups that are closely tight to identify similar interest groups. Followers can be recommended based on this information. Projects can also be recommended amongst the groups themselves based on the idea of being from the same circle.

6.2.4 Recommendation based on heuristics

CF-based and content-based recommender systems are well known in the field. How- ever, they may not be perfect for every recommendation problem. Many different approaches need to be explored for building a good recommendation for scratch. One of the more simple and easy to implement approach is the heuristics based. In a heuristics-based system, one could implement a different set of rules to recommend projects to users. For example, if the user is new in the system, we might recommend projects that guide the user in getting started and simpler projects. If the user is an experienced one and has built many complex projects, recommendations for projects from other users who have more complex projects in their profile may be recommended.

Chapter 7

Conclusion

7.1 Summary

This dissertation explores the concept of a recommender systems for Scratch. It briefly describes the motivation for the study and focuses on the problem of content discovery and self-directed learning in the Scratch environment. The idea of a recommender system might look like a simple problem from top but it requires an understanding of the underlying data analysis concepts. Various different recommender systems used in different fields have been discussed. Specifically the once in the area of e-learning and online teaching, which have gained popularity and are used extensively. Different studies have attempted to develop a recommender system for Scratch in their own teaching context.

Over the period, various research has been done in the area of a recommender system. Various techniques have been proposed for building recommend systems for different applications. This thesis discusses the standard techniques in recommender system space and their uses. Content-based filtering, collaborative filtering and hybrid approach are most widely used and accepted as standard amongst the community. A Collaborative filtering process is used and applied on the Scratch data to try and build an item-based collaborative filtering recommender system. Basic tasks in a recom- mender system like generating a recommendation list and calculating the prediction are detailed for the dataset.

For generating a recommendation list in an item-based collaborative technique, the

46

Draft of 3:58 am, Tuesday, August 13, 2019 47

first step is similarity computation. Three different techniques are used for similarity computation. Cosine based similarity, correlation-based similarity and adjusted cosine based similarity are the three well-known techniques for computing the item similarity. These three techniques were applied on the dataset to get the item similarity matrix for all the project a user. Once the item similarity matrix was generated prediction calculation was done based on the weighted sum method. Prediction output gave us the predicted recommendation rating for the user. Once the prediction was generated. the projects were recommended to the user according to the score.

Different methods of evaluation for the accuracy of recommendation are discussed. For measuring the accuracy of the recommendation we use MAE as the evaluation metrics. On comparing the three similarity calculation methods, it is observed that the cosine similarity is the best performing one on the Scratch dataset.

When we look at the top 5 generated recommendation of Scratch projects similar to the supplied item i.e our seed project, we observe some similarities between the recommended items. All the items had a similar set of script blocks and sprite usages. This shows the correctness of the similarity computation blocks in N dimensional user space.Although this thesis demonstrates a working model of a recommendation system for scratch, much more work is needed to validate the usefulness of the proposed methods. Many more useful techniques could be developed by incorporating the user’s behaviour over time. This method of recommendation is more prone to general problems faced by recommender systems. Limitations like over specialisation problem, cold start problem and concentration bias limit the work of the proposed method. Moreover, the quality of the data available also affects largely on the prediction results. If the data has missing values for the user ratings, predictions cannot be calculated correctly by the algorithms.

Future work for this study would involve qualitatively testing the proposed method for the usefulness of the system. Various testing methodologies could be adopted like a/b testing, user surveys or other novel techniques as discussed in [51]. Also, more data could be extracted from the datasets to gain more information about the user, like a text analysis or sentiment analysis could be done on the iext data of the projects for understanding the users preference and taste. A more complex analysis could be a social network graph analysis on the follower-followee friendship to understand the

Draft of 3:58 am, Tuesday, August 13, 2019 48

user interactions and taste. A heuristics-based recommendation could be a fast and effective alternative if developed correctly.

To demonstrate this a client/server model was developed (appendix A). In the scripts, we use standard Scratch API and Python data analysis libraries and ML li- braries for model building.

Bibliography

[1] M. Resnick and K. Robinson, Lifelong kindergarten: Cultivating creativity through

projects, passion, peers, and play. MIT press, 2017.

[2] J. Maloney, M. Resnick, N. Rusk, B. Silverman, and E. Eastmond, “The scratch programming language and environment,” ACM Transactions on Computing Ed- ucation (TOCE), vol. 10, no. 4, p. 16, 2010.

[3] M. Resnick, J. Maloney, A. Monroy-Hernández, N. Rusk, E. Eastmond, K. Bren- nan, A. Millner, E. Rosenbaum, J. S. Silver, B. Silverman, et al., “Scratch: Pro- gramming for all.,” Commun. Acm, vol. 52, no. 11, pp. 60–67, 2009.

[4] J.-M. Sáez-López, M. Román-González, and E. Vázquez-Cano, “Visual program- ming languages integrated across the curriculum in elementary school: A two year case study using scratch in five schools,” Computers & Education, vol. 97, pp. 129–141, 2016.

[5] S. Cooper, W. Dann, and R. Pausch, “Teaching objects-first in introductory com-

puter science,” in ACM SIGCSE Bulletin, vol. 35, pp. 191–195, ACM, 2003.

[6] M. C. Carlisle, T. A. Wilson, J. W. Humphries, and S. M. Hadfield, “Raptor: introducing programming to non-majors with flowcharts,” Journal of Computing Sciences in Colleges, vol. 19, no. 4, pp. 52–60, 2004.

[7] C. J. Bouras, V. Poulopoulos, and V. Tsogkas, “Squeak etoys: Interactive and collaborative learning environments,” in Handbook of research on social interaction technologies and collaboration software: Concepts and trends, pp. 417–427, IGI Global, 2010.

49

Draft of 3:58 am, Tuesday, August 13, 2019 50

[8] D. J. Malan and H. H. Leitner, “Scratch for budding computer scientists,” ACM

Sigcse Bulletin, vol. 39, no. 1, pp. 223–227, 2007.

[9] I. F. de Kereki, “Scratch: Applications in computer science 1,” in 2008 38th

Annual Frontiers in Education Conference, pp. T3B–7, IEEE, 2008.

[10] H.-H. Chen, I. Ororbia, G. Alexander, and C. L. Giles, “Expertseer: A keyphrase based expert recommender for digital libraries,” arXiv preprint arXiv:1511.02058, 2015.

[11] N. R. Council et al., How people learn: Brain, mind, experience, and school:

Expanded edition. National Academies Press, 2000.

[12] L. Song and J. R. Hill, “A conceptual model for understanding self-directed learn- ing in online environments,” Journal of Interactive Online Learning, vol. 6, no. 1, pp. 27–42, 2007.

[13] A. Felfernig, K. Isak, K. Szabo, and P. Zachar, “The vita financial services sales support environment,” in Proceedings of the national conference on artificial intel- ligence, vol. 22, p. 1692, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

[14] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles, “Collabseer: a search en- gine for collaboration discovery,” in Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, pp. 231–240, ACM, 2011.

[15] N. Manouselis, H. Drachsler, R. Vuorikari, H. Hummel, and R. Koper, “Rec- ommender systems in technology enhanced learning,” in Recommender systems handbook, Springer, 2011.

[16] F. Ricci, L. Rokach, and B. Shapira, “Introduction to recommender systems hand-

book,” in Recommender systems handbook, pp. 1–35, Springer, 2011.

[17] B. M. Sarwar, G. Karypis, J. A. Konstan, J. Riedl, et al., “Item-based collaborative

filtering recommendation algorithms.,” Www, vol. 1, pp. 285–295, 2001.

Draft of 3:58 am, Tuesday, August 13, 2019 51

[18] A. Zielinski, “A utility-based semantic recommender for technology-enhanced learning,” in 2015 IEEE 15th International Conference on Advanced Learning Technologies, pp. 394–396, IEEE, 2015.

[19] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “analysis of recommendation algorithms for e-commerce, in proceeding of the 2th acm, e-commerce conference (ec00), october 2000,” 2000.

[20] C. Jesennia, A. Puris, D. Benavides, et al., “Recommender systems and scratch: An integrated approach for enhancing computer programming learning,” IEEE Transactions on Learning Technologies, 2019.

[21] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “Grouplens: an open architecture for collaborative filtering of netnews,” in Proceedings of the 1994 ACM conference on Computer supported cooperative work, pp. 175–186, ACM, 1994.

[22] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, “Grouplens: applying collaborative filtering to usenet news,” Communications of the ACM, vol. 40, no. 3, pp. 77–87, 1997.

[23] U. Shardanand and P. Maes, “Social information filtering: algorithms for automat-

ing” word of mouth”,” in Chi, vol. 95, pp. 210–217, Citeseer, 1995.

[24] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, “Recommending and evaluating choices in a virtual community of use,” in Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 194–201, ACM Press/Addison-Wesley Publishing Co., 1995.

[25] H. Tan, J. Guo, and Y. Li, “E-learning recommendation system,” in 2008 In- ternational Conference on Computer Science and Software Engineering, vol. 5, pp. 430–433, IEEE, 2008.

[26] A. Pongpech, M. E. Orlowska, and S. W. Sadiq, “Personalized courses recom- mendation functionality for flex-el,” in Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007), pp. 631–633, IEEE, 2007.

Draft of 3:58 am, Tuesday, August 13, 2019 52

[27] J. Lu, “A personalized e-learning material recommender system,” in Interna- tional Conference on Information Technology and Applications, Macquarie Sci- entific Publishing, 2004.

[28] C. Limongelli, M. Lombardi, and A. Marani, “Towards the recommendation of resources in coursera,” in Intelligent Tutoring Systems: 13th International Con- ference, ITS 2016. Proceedings, vol. 9684, p. 461, Springer, 2016.

[29] A. Kardan, A. Narimani, and F. Ataiefard, “A hybrid approach for thread recom- mendation in mooc forums,” International Journal of Social, Behavioral, Educa- tional, Economic, Business and Industrial Engineering, vol. 11, no. 10, pp. 2195– 2201, 2017.

[30] J. Cardenas, “Scratch and caramba video.”

[31] A. Gunawardana and G. Shani, “A survey of accuracy evaluation metrics of rec- ommendation tasks,” Journal of Machine Learning Research, vol. 10, no. Dec, pp. 2935–2962, 2009.

[32] P. Cremonesi, Y. Koren, and R. Turrin, “Performance of recommender algorithms on top-n recommendation tasks,” in Proceedings of the fourth ACM conference on Recommender systems, pp. 39–46, ACM, 2010.

[33] B. M. Sarwar, J. A. Konstan, A. Borchers, J. Herlocker, B. Miller, and J. Riedl, “Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system,” in in the GroupLens Research Collaborative Fil- tering System???. Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW, 1998.

[34] J. S. Breese, D. Heckerman, and C. Kadie, “Empirical analysis of predictive al- gorithms for collaborative filtering,” in Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, pp. 43–52, Morgan Kaufmann Publishers Inc., 1998.

[35] C. C. Aggarwal, J. L. Wolf, K.-L. Wu, and P. S. Yu, “Horting hatches an egg: A new graph-theoretic approach to collaborative filtering,” in Proceedings of the fifth

Draft of 3:58 am, Tuesday, August 13, 2019 53

ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 201–212, ACM, 1999.

[36] C. Zaiontz, “real-statistics.com basic concepts of correlation,” 2019.

[37] “Scratch research data.”

[38] “nature.com.”

[39] “scikit-learn.”

[40] “Python data analysis library.”

[41] T. Chai and R. R. Draxler, “Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature,” Geoscientific model development, vol. 7, no. 3, pp. 1247–1250, 2014.

[42] J. P. Egan, Signal detection theory and ROC-analysis. Academic press, 1975.

[43] F. J. Provost, T. Fawcett, et al., “Analysis and visualization of classifier per- formance: Comparison under imprecise class and cost distributions.,” in KDD, vol. 97, pp. 43–48, 1997.

[44] M. Ge, C. Delgado-Battenfeld, and D. Jannach, “Beyond accuracy: evaluating recommender systems by coverage and serendipity,” in Proceedings of the fourth ACM conference on Recommender systems, pp. 257–260, ACM, 2010.

[45] S. Jain, A. Grover, P. S. Thakur, and S. K. Choudhary, “Trends, problems and solutions of recommender system,” in International Conference on Computing, Communication & Automation, pp. 955–958, IEEE, 2015.

[46] P. Adamopoulos and A. Tuzhilin, “On over-specialization and concentration bias of recommendations: Probabilistic neighborhood selection in collaborative filtering systems,” RecSys 2014 - Proceedings of the 8th ACM Conference on Recommender Systems, pp. 153–160, 10 2014.

[47] L. Sharma and A. Gera, “A survey of recommendation system: Research chal- lenges,” International Journal of Engineering Trends and Technology (IJETT), vol. 4, no. 5, pp. 1989–1992, 2013.

Draft of 3:58 am, Tuesday, August 13, 2019 54

[48] “Development of scratch 1.0.”

[49] “Scratch3.0.”

[50] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl, “Is seeing believing?: how recommender system interfaces affect users’ opinions,” in Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 585–592, ACM, 2003.

[51] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, “Ex- plaining the user experience of recommender systems,” User Modeling and User- Adapted Interaction, vol. 22, no. 4-5, pp. 441–504, 2012.

Appendix

.1 A

1 #!/usr/bin/python 2 # 3 # Created on Sat Jul 4 20:31:00 2019 4 # Copyright (C) 2019, shaikhr@tcd.ie 5 # @author: shaikhr 6 # 7 # API reference guide : https://github.com/LLK/scratch-rest-api/wiki/

Projects 8 # Scratch data guide : https://communitydata.science/scratch-data/

910 #Libs for Serving 11 from http.server import HTTPServer, BaseHTTPRequestHandler 12 from urllib.parse import urlparse,parse\_qs 13 from io import BytesIO 14 import urllib.request, json

1516 #not required for now can be commented 17 from selenium import webdriver 18 from selenium.webdriver.chrome.options import Options 19 import time

2021 #Libs for data analysis 22 import pandas as pd 23 from sklearn.impute import SimpleImputer 24 from sklearn.experimental import enable\_iterative\_imputer 25 from sklearn.impute import IterativeImputer 26 from sklearn import preprocessing

55

Draft of 3:58 am, Tuesday, August 13, 2019 56

27 from sklearn.metrics.pairwise import euclidean\_distances 28 from sklearn.metrics.pairwise import cosine\_similarity 29 from scipy.spatial.distance import correlation

303132 #Returns a list of the users that the specified user has followed. 33 def get\_user\_following(username, LIMIT): 34 if not LIMIT: 35 LIMIT = 10 36 OFFSET = 0 37 with urllib.request.urlopen("https://api.scratch.mit.edu/users/"+

username+"/following?limit="+str(LIMIT)+"&offset="+str(OFFSET)) as

url: 38 data = json.loads(url.read().decode()) 39 return data

404142 #Returns an array of details regarding the projects that a given user

has favourited on the website. 43 def get\_user\_favriots(username, LIMIT): 44 if not LIMIT: 45 LIMIT = 3 46 OFFSET = 0 47 with urllib.request.urlopen("https://api.scratch.mit.edu/users/"+

username+"/favorites?limit="+str(LIMIT)+"&offset="+str(OFFSET)) as

url: 48 data = json.loads(url.read().decode()) 49 return data

505152 #Returns an array with information regarding the projects that a given

user has shared on the Scratch website. 53 def get\_user\_projects(username, LIMIT): 54 if not LIMIT: 55 LIMIT = 10 56 OFFSET = 0 57 with urllib.request.urlopen("https://api.scratch.mit.edu/users/"+

username+"/projects?limit="+str(LIMIT)+"&offset="+str(OFFSET)) as url:

Draft of 3:58 am, Tuesday, August 13, 2019 57

58 data = json.loads(url.read().decode()) 59 return data

606162 #Returns an array with information regarding the project with provided

ID of particular username (dosent work , requires auth) 63 def get\_user\_project\_details(username,project\_id): 64 with urllib.request.urlopen("https://api.scratch.mit.edu/users/"+

username+"/projects/"+project\_id) as url: 65 data = json.loads(url.read().decode()) 66 return data

676869 #Gets the project name from given project ID (api deprecated , using

selenium to render JS) 70 def get\_project\_title(project\_id): 71 url = ’https://scratch.mit.edu/projects/’+str(project\_id) 72 #rendered page is a JS script, no use of lxml #alternative is using

selenium 73 chrome\_options = Options() 74 chrome\_options.add\_argument("--headless") 75 driver = webdriver.Chrome(’./selenium/chromedriver’,chrome\_options= chrome\_options) # Optional argument, if not specified will search

path. 76 #driver = webdriver.Chrome(’./selenium/chromedriver.exe’) 77 driver.get(url) 78 time.sleep(5) 79 title = driver.title 80 driver.quit() 81 return title

828384 # ( No working API, therefor using lxml html parsing) 85 def get\_author\_from\_project(project\_id): 86 url = ’https://scratch.mit.edu/projects/’+str(project\_id) 87 #rendered page is a JS script, no use of lxml #alternative is using

selenium 88 chrome\_options = Options() 89 chrome\_options.add\_argument("--headless")

Draft of 3:58 am, Tuesday, August 13, 2019 58

90 driver = webdriver.Chrome(’./selenium/chromedriver.exe’,

chrome\_options=chrome\_options) # Optional argument, if not specified will search path. 91 #driver = webdriver.Chrome(’./selenium/chromedriver.exe’) 92 driver.get(url) 93 time.sleep(5) 94 author = driver.find\_element\_by\_xpath("//div[@class=’title’]/a").

text 95 driver.quit() 96 return author

979899 #Takes project json object as input and returns the stats if present. 100 def get\_project\_stats(project\_obj): 101 #Difficult to retrive the below values from current APIs therefor

assigning mock values. 102 sprites\_website = 5 103 scripts\_website = 10 104 blocks = 6 105 block\_types = 7 106 images = 2 107 sounds = 2 108 ugstrings = 11 109 stats = {’sprites\_website’:sprites\_website,’scripts\_website’:

scripts\_website,’blocks’:blocks,’block\_types’:block\_types,’images’ :images,’sounds’:sounds,’ugstrings’:ugstrings} 110 for key in project\_obj: 111 if key == ’id’: 112 stats[’id’] = project\_obj[key] 113 if key == ’stats’: 114 stats.update(project\_obj[key]) 115 return stats 116 return stats

117

118 119 #Checks if the supplied project has a remix associated with it. If yes

returns the remix object containing the remixes 120 def is\_project\_remix(project\_obj): 121 for key in project\_obj:

Draft of 3:58 am, Tuesday, August 13, 2019 59

122 if key == ’remix’: 123 #Nested for not too heavy because only 2 remix has only two

components 124 for x in project\_obj[key]: 125 if project\_obj[key][x] != None: 126 return project\_obj[key] 127 else: 128 return False

129

130 131 #Gets from the CSV database against the supplies stats of the project 132 def generate\_recommmendation\_from\_db(stats\_obj): 133 #Load data from the csv 134 projects\_df = pd.read\_csv(’data/CSVs/projects.csv’, sep=’,’,

index\_col=0, nrows=100) 135 #index = projects\_df.index 136 #columns = projects\_df.columns 137 #values = projects\_df.values

138 139 #Filter columns 140 required\_columns\_df = projects\_df[[’viewers\_website’,’lovers\_website

’,’downloaders\_website’,’sprites\_website’,’scripts\_website’, ’ blocks’, ’block\_types’, ’images’, ’sounds’, ’ugstrings’]]

141 142 project\_id = stats\_obj.pop(’id’)

143 144 new\_stats\_row = pd.Series(stats\_obj) 145 new\_stats\_row.name = project\_id 146 required\_columns\_df = required\_columns\_df.append(new\_stats\_row)

147 148 #Impute the missing values using IterativeImputer from sklearn 149 #imp = SimpleImputer(missing\_values=np.nan, strategy=’mean’) 150 imp = IterativeImputer(max\_iter=10,initial\_strategy=’most\_frequent’,

random\_state=0) 151 imputed\_DF = pd.DataFrame(imp.fit\_transform(required\_columns\_df)) 152 imputed\_DF.columns = required\_columns\_df.columns 153 imputed\_DF.index = required\_columns\_df.index

154 155 #Normalize the data for further calculation

Draft of 3:58 am, Tuesday, August 13, 2019 60

156 x = imputed\_DF.values 157 min\_max\_scaler = preprocessing.MinMaxScaler() 158 x\_scaled = min\_max\_scaler.fit\_transform(imputed\_DF) 159 normalised\_df = pd.DataFrame(x\_scaled) 160 normalised\_df.columns = required\_columns\_df.columns 161 normalised\_df.index = required\_columns\_df.index

162 163 #Calculate the RMSE score 164 normalised\_df[’RMSE’] = pd.Series((normalised\_df.iloc[:,1:]\*\*2).sum

(1).pow(1/2)) 165 print(normalised\_df)

166 167 #Generate the similarity matrix based on the the new project stats. 168 sim = cosine\_similarity(normalised\_df) 169 final\_df = pd.DataFrame(sim) 170 final\_df.columns = required\_columns\_df.index 171 final\_df.index = required\_columns\_df.index 172 print(final\_df)

173 174 #Get Top 5 from the similarity matrix 175 top5 = final\_df.loc[project\_id].nlargest(5) 176 recommendation\_pairs = top5.to\_dict()

177 178 return recommendation\_pairs

179

180 181 #Calculates recommendation score by taking the mean of the generated

recommendation matrix based on provided input stats of the project . 182 #Need more optimized method to calculate more accurate score.

Currently needs work.( Probably apply K nearest neighbor technique

to get nearest neighbor and calculate further score. 183 def calculate\_recommendation\_score(stats\_obj): 184 MAX\_ITER = 10 185 RECOMMENDATION\_NUMBER = 5 186 projects\_df = pd.read\_csv(’data/CSVs/projects.csv’, sep=’,’,

index\_col=0, nrows=100)

187 188 #Filter columns

Draft of 3:58 am, Tuesday, August 13, 2019 61

189 required\_columns\_df = projects\_df[[’viewers\_website’,’lovers\_website

’,’downloaders\_website’,’sprites\_website’,’scripts\_website’, ’ blocks’, ’block\_types’, ’images’, ’sounds’, ’ugstrings’]]

190 191 #project\_id = stats\_obj.pop(’id’) 192 project\_id = ’10000001’

193 194 new\_stats\_row = pd.Series(stats\_obj) 195 new\_stats\_row.name = project\_id 196 required\_columns\_df = required\_columns\_df.append(new\_stats\_row)

197 198 #Impute the missing values using IterativeImputer from sklearn 199 #imp = SimpleImputer(missing\_values=np.nan, strategy=’mean’) 200 imp = IterativeImputer(max\_iter=MAX\_ITER,initial\_strategy=’

most\_frequent’, random\_state=0) 201 imputed\_DF = pd.DataFrame(imp.fit\_transform(required\_columns\_df)) 202 imputed\_DF.columns = required\_columns\_df.columns 203 imputed\_DF.index = required\_columns\_df.index

204 205 #Normalize the data for further calculation 206 x = imputed\_DF.values 207 min\_max\_scaler = preprocessing.MinMaxScaler() 208 x\_scaled = min\_max\_scaler.fit\_transform(imputed\_DF) 209 normalised\_df = pd.DataFrame(x\_scaled) 210 normalised\_df.columns = required\_columns\_df.columns 211 normalised\_df.index = required\_columns\_df.index

212 213 #Calculate the RMSE score 214 normalised\_df[’RMSE’] = pd.Series((normalised\_df.iloc[:,1:]\*\*2).sum

(1).pow(1/2))

215 216 #Generate the similarity matrix based on the the new project stats. 217 sim = cosine\_similarity(normalised\_df) 218 final\_df = pd.DataFrame(sim) 219 final\_df.columns = required\_columns\_df.index 220 final\_df.index = required\_columns\_df.index

221 222 score = final\_df.loc[project\_id].nlargest(RECOMMENDATION\_NUMBER) 223 final\_score = 0

Draft of 3:58 am, Tuesday, August 13, 2019 62

224 for x in score: 225 final\_score += x

226 227 return final\_score/RECOMMENDATION\_NUMBER

228

229 230 #Reformats the stats object in the format required for the

recommendation method 231 #DICT FORMAT: {’sprites\_website’: int64(xxx), ’scripts\_website’: int64

(xxx), ’blocks’: int64(xxx), ’block\_types’: int64(xxx), ’images’: int64(xxx), ’sounds’: int64(xxx), ’ugstrings’: int64(xxx), ’ viewers\_website’: int64(xxx), ’lovers\_website’: int64(xxx), ’ downloaders\_website’: int64(xxx)} 232 def reformat\_stats(stats\_obj): 233 stats\_obj[’viewers\_website’] = stats\_obj.pop(’views’) 234 stats\_obj[’lovers\_website’] = stats\_obj.pop(’loves’) 235 stats\_obj[’downloaders\_website’] = stats\_obj.pop(’favorites’) #Need

to fix this 236 stats\_obj.pop(’comments’) 237 stats\_obj.pop(’remixes’) 238 return stats\_obj

239

240 241 #Generates recommendations for a given username, and returns a json

object 242 def get\_recommended\_projects(username): 243 RECOMMENDATION\_REASON\_1 = "<i> Recommended because you follow: <a

href=/users/" 244 RECOMMENDATION\_REASON\_2 = "<i> Recommended because you have project:

<a href=/projects/" 245 RECOMMENDATION\_REASON\_3 = "<i> Recommended because user: <a href=/

projects/" 246 RECOMMENDATION\_REASON\_4 = "<i> Recommended because its a remix of <a

href=/projects/"

247 248 all\_followers = get\_user\_following(username,10) 249 all\_recommendation = []

250 251 #Gets recommendation from the csv database against the given user

Draft of 3:58 am, Tuesday, August 13, 2019 63

data 252 for projects in get\_user\_projects(username,5): 253 project\_stats = get\_project\_stats(projects) 254 project\_id = project\_stats["id"]

255 256 recommendations\_from\_db = generate\_recommmendation\_from\_db(

reformat\_stats(project\_stats)) 257 recommended\_project\_ids = recommendations\_from\_db.keys()

258 259 for recom in recommended\_project\_ids: 260 recommendations={} 261 recommendations[’id’] = recom 262 recommendations[’title’] = recom #get\_project\_title(recom)----->

replace for getting actual name of the project insted of the ID but very slow because of selenium call 263 recommendations[’stats’] = project\_stats 264 recommendations[’score’] = recommendations\_from\_db[recom] 265 recommendations[’reason’] = RECOMMENDATION\_REASON\_2 + str(

project\_id) + ">" + str(project\_id) + "</a>" 266 recommendations[’reason\_id’] = username 267 all\_recommendation.append(recommendations)

268 269 #Get details of the users who the user is following ( projects of

following , favorites of following) 270 for user in all\_followers:

271 272 #Projects from followers 273 all\_follower\_projects = get\_user\_projects(user["username"],5) 274 for project in all\_follower\_projects: 275 recommendations={} 276 recommendations[’id’] = project["id"] 277 recommendations[’title’] = project["title"] 278 recommendations[’stats’] = project["stats"] 279 recommendations[’score’] = calculate\_recommendation\_score(

project["stats"]) 280 recommendations[’reason’] = RECOMMENDATION\_REASON\_1 + user["

username"] + ">" + user["username"] + "</a>" 281 recommendations[’reason\_id’] = user["id"] 282 all\_recommendation.append(recommendations)

Draft of 3:58 am, Tuesday, August 13, 2019 64

283 284 #Projects that the followers have liked 285 all\_followers\_favourite = get\_user\_favriots(user["username"],5) 286 for favourite in all\_followers\_favourite: 287 recommendations={} 288 recommendations[’id’]= favourite["id"] 289 recommendations[’title’]= favourite["title"] 290 recommendations[’stats’] = favourite["stats"] 291 recommendations[’score’] = calculate\_recommendation\_score(

favourite["stats"]) 292 recommendations[’reason’] = RECOMMENDATION\_REASON\_3 + user["

username"] + ">" + user["username"] + "</a> has liked it." 293 recommendations[’reason\_id’] = user["id"] 294 all\_recommendation.append(recommendations)

295 296 #Check if project has remix of the project, and add that to the

recommendation list. 297 for project in get\_user\_projects(user["username"],5): 298 is\_remix = is\_project\_remix(project) 299 if is\_remix: 300 recommendations={} 301 recommendations[’id’]= is\_remix["parent"] 302 recommendations[’title’]= project["title"] 303 recommendations[’stats’] = project["stats"] 304 recommendations[’score’] = calculate\_recommendation\_score(

project["stats"]) 305 recommendations[’reason’] = RECOMMENDATION\_REASON\_4 + str(

is\_remix["root"]) + ">" + str(is\_remix["root"]) + "</a>" 306 recommendations[’reason\_id’] = user["id"] 307 all\_recommendation.append(recommendations)

308 309 #Get details of the favorite project of the user ( No working API,

therefor using selenium for parsing). 310 #Too slow needs,an alternative 311 all\_favriots = get\_user\_favriots(username,5) 312 for favriots in all\_favriots: 313 project\_id = favriots["id"] 314 #get\_author\_from\_project(project\_id)

315

Draft of 3:58 am, Tuesday, August 13, 2019 65

316 #print(all\_recommendation) 317 return json.dumps(all\_recommendation)

318

319 320 #Extends pythons simple http server to handle GET requests locally 321 class SimpleHTTPRequestHandler(BaseHTTPRequestHandler): 322 def end\_headers (self): 323 #Required for running locally on chrome ( allows CORS for all

request to the server ) 324 self.send\_header(’Access-Control-Allow-Origin’, ’\*’) 325 BaseHTTPRequestHandler.end\_headers(self)

326 327 #Gets the username in the GET request url as username and returns a

list of recommendations 328 def do\_GET(self): 329 self.send\_response(200) 330 self.end\_headers() 331 query\_components = parse\_qs(urlparse(self.path).query) 332 username = query\_components["username"][0] 333 response = BytesIO() 334 response.write(b’The username recived in URL is ’) 335 response.write(bytes(username, ’utf-8’)) 336 self.wfile.write(bytes(get\_recommended\_projects(username), ’utf-8’

))

337

338 339 #Handels post request, wrote this to get data via POST, not used

currently 340 def do\_POST(self): 341 content\_length = int(self.headers[’Content-Length’]) 342 body = self.rfile.read(content\_length) 343 self.send\_response(200) 344 self.end\_headers() 345 response = BytesIO() 346 response.write(b’This is POST request. ’) 347 response.write(body) 348 self.wfile.write(response.getvalue())

349

350

Draft of 3:58 am, Tuesday, August 13, 2019 66

351 if \_\_name\_\_ == "\_\_main\_\_": 352 try: 353 #NOTE: If you update port here, remember to update in the JS

extension as well. 354 PORT = 8000 355 HOST = ’localhost’ 356 httpd = HTTPServer((HOST, PORT), SimpleHTTPRequestHandler) 357 print(’Started httpserver on port ’ + str(PORT)) 358 httpd.serve\_forever()

359 360 except KeyboardInterrupt: 361 print( ’^C received, shutting down the web server’) 362 httpd.socket.close()