A Web Recommender System using User Logs Files with k-NN Method

Anupama Patel
PG Scholar, Department of Computer Science Engineering, GGITS, Jabalpur, M.P, India

Santosh K. Vishwakarma, PhD
Assoc. Professor, Department of Computer Science Engineering, GGITS, Jabalpur, M.P, India

ABSTRACT

In the current era there is a huge increment in the generation of web data. Internet is getting overloaded with the massive increase in data. This unstable growth in information is making the search a complicate process. This in result has given rise to a new idea to analysis the system. Web recommender system is the most efficient solution for the problem and is widely used in e-commerce websites to suggest product in reference to the user request. Giving suggestion is not an easy task, whereas it not only helps in saving time but also helps in decision making. Web servers contain log files these log files have records of events in the sequential pattern. Sequential information gives the detail information about the user's behavior. In this paper k-NN method is implemented to obtain the prediction for the new users. The dataset used in this paper is the dummy dataset.

General Terms

E-commerce websites, log files, dummy dataset.

Keywords

Web recommender system, sequential information, web server, log files, k-NN.

1. INTRODUCTION

In the today's world data has no limits. It is getting huge and overloaded. Fetching the desired information from the massive amount of data has become a big issue. Web recommender system helps in recommending the right information at the right time. Web recommender system has reduced the load for searching items and had also made the searching easier. Many data analysis tools are used for data management but the strategy behind recommending items is the importance given by the user to each platform they visit. Web recommender system is applied over variety of applications some of them are movie recommender ex. Netflix; social recommender ex. Facebook, news content recommender ex. Yahoo, Google, online course offering ex. Coursera, computational advertisement ex. Facebook, Yahoo et al. There are various other applications of web recommender system but the most widely used application is e-commerce. E-commerce organizations not only provide suggestions but it also promotes products to the user.

Recommender system has changed the way people used to find items, information, and even other peoples. Recommender system studies the pattern of user behavior in order to know what the user will prefer among the collection of data he has never experienced.

Web recommender system works on the application of web personalization. Collaborative filtering and content based filtering are the two widely used recommender techniques. Content based filtering provides recommendation to the user based on the items having similar content matches to the user profile. The content based filtering approach try to recommend items that are quite similar to those that the user has rated in the past. Content based filtering technique uses text extraction method. Content based filtering also contains some disadvantages like content mismatch which can also lead to poor performance. Another disadvantage is that if the user has not given proper rating to the web pages or to the websites then the counting of ratings will be very less over that item. Content based filtering is not only used in recommender system but it is also being used in information retrieval system which provides results based on the content entered example search engines.

Collaborative filtering, on the other hand, relies on the similar type of user's preference over the item. The web pages are recommended to a particular user only when it is being recommended by other similar types of users. It is a technique of personalized recommendation. There are algorithms that are used to measure the user similarity or the item similarity is the k-NN technique and the Pearson correlation technique. Collaborative filtering is based on one of the assumption, if the user agreed in the past will also agree in the future. Collaborative filtering plays an important role in recommending items to the user. Unlike other collaborative filtering also faces many challenges such as data sparsity, cold- start problem, scalability, synonymy, shilling attack, gray sheep, and diversity & the long tail. Figure 1 is showing the working of collaborative filtering.

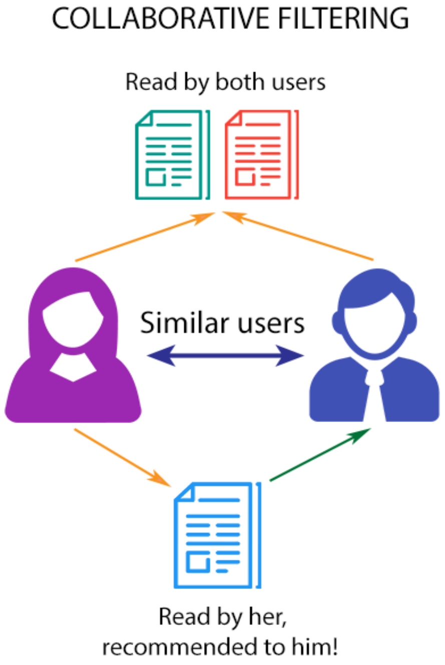


Fig 1: Collaborative filtering

Collaborative filtering plays an important role in recommending items to the user. Web server contains all the information about the user activities on the web site in the form of log files. The log files contain the browsing history of the users in a sequential pattern which helps in prediction. The user activity information can also be collected from the following sources such as web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transaction, user queries, bookmark data and much more. In this paper k-NN method is used for predicting items. Similarity metrics used here is the cosine-based filtering.

2. LITERATURE REVIEW

So far many research works have been done in the field of web recommender system. So some of the studies related to web recommender system are discussed below:

Rajhans Mishra et al. [1] have proposed a model to predict user's next page visit. The next page visit of a user will provide the useful information about their likes and interest. For which, a rough set based similarity upper approximation concept is used during clustering to generate soft clusters. The S^3Msimilarity measures have been utilized which is a hybrid of content and sequential similarity measures. The generated soft clusters have been utilized to create a response matrix which is used by SVD to generate predictions. The result of the model is then compared with the random prediction and first order Markov based models. The datasets used are MSNBC web navigation dataset, simulated dataset and CTI data set.

Pradeep Kumar et al. [2] have proposed a clustering algorithm which is based on the similarity upper approximation of a rough set cluster using S^3M similarity measures.

Rajhans Mishra et al. [3] have implemented the algorithm proposed by Pradeep Kumar having similarity measures for finding the clusters and outliers. Four similarity measures have been used that are Jaccard, Sorensen-Dice, S^3M and Levenshtein. In this paper rough set upper approximation is used for finding the clusters having different classes of users and also to find the outliers. All the four results are the compared to know which one among them provides better result. S^3M measures show the better result among all four similarity measures because it measures the similarity based on content as well as sequence.

Prajyoti Lopes et al. [4] this paper focuses on providing real-time recommender to all the visitors of the website irrespective of been registered and unregistered on the website. The paper has used action based rational technique which result in the generation of lexical patterns in order to generate item recommender. The proposed system dynamically provides a recommender as per changing user behavior and traversal patterns. The system also minimizes the false positive error that occurs frequently in traditional recommender system.

The next paper entitled Applying Web Usage Mining Techniques to Design Effective Web Recommender System is case study by Maryam Jafari, Farzad soleymani Sabzchi and Amir Jalili Irani, [5] has discussed about the concepts and techniques of recommender system. In addition to this the paper has also discussed about how to apply web usage mining over the web logs for discovering access patterns. Lastly an analysis is done over the problem causing during the deployment of recommender system and has also proposed solutions which address the problems.

Shiva Nadi [6] has proposed a hybrid recommender system by combining the two filtering techniques which is the content based filtering and collaborative filtering form a hybrid combinations of both the techniques. Here the author has used web content mining as a source and fuzzy C-means techniques as an offline process.

3. METHODOLOGY

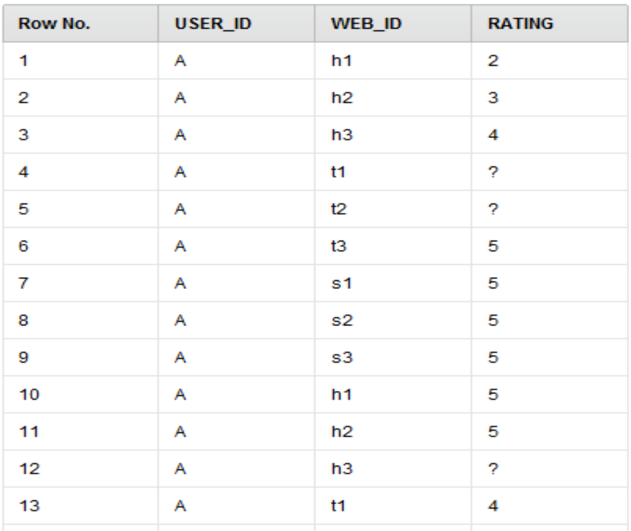
The research work has proposed a system for efficient web page recommender using sample data set coupled with User k-NN method for predicting recommender for the new user based on the previous browsing history of the users having the similar type of behavior. The web recommender system is developed with the following steps:

3.1 Data Preparation

The everyday activity of the user is maintained in the form of records in the web servers in the form of log files. The log files are in the form of plain-text files. So they are not used directly. To discover knowledge from these log files it is necessary to preprocess the data. Preprocessing of data is done in four steps:

3.1.1 Data Collection

It is the first step of preprocessing and the data is being collected from the different sources like the web servers, proxy servers, or the from the client machine. A sample dataset is shown in Figure 2 consisting of 89 example values and 3 regular attributes.



Row No.	USER_ID	WEB_ID	RATING
1	A	h1	2
2	A	h2	3
3	A	h3	4
4	A	t1	?
5	A	t2	?
6	A	t3	5
7	A	s1	5
8	A	s2	5
9	A	s3	5
10	A	h1	5
11	A	h2	5
12	A	h3	?
13	A	t1	4

Fig 2: Sample of dummy dataset

3.1.2 Data Cleaning

It is the second step of data preprocessing and is a very important phase. In this phase the unnecessary and noisy data are removed from the log files. It has fields like data, time, client IP URL access, and Referrer Access log files [7].

3.1.3 User Identification

In this phase, of preprocessing the user identification is done. With the help of the user id and the IP address, the user, is considered as unique.

3.1.4 Session Identification

Here user session in being identified with the help of following rules [7]:

a) For every new user there is a new session.

b) For the single user session if the refer page is null, and then there will be a new session.

c) If the time limit (30 or 25 minutes) between page request exceeds, then the assumption is made that the user is starting the new session.

The above-mentioned steps transform the raw data into the readable format and it also increases the quality and efficiency of the data. Filtering in down over the data and the data is split into two parts. First part shows the rated values and other part show the unrated values.

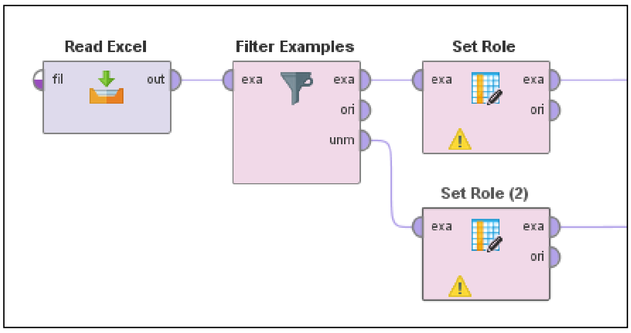
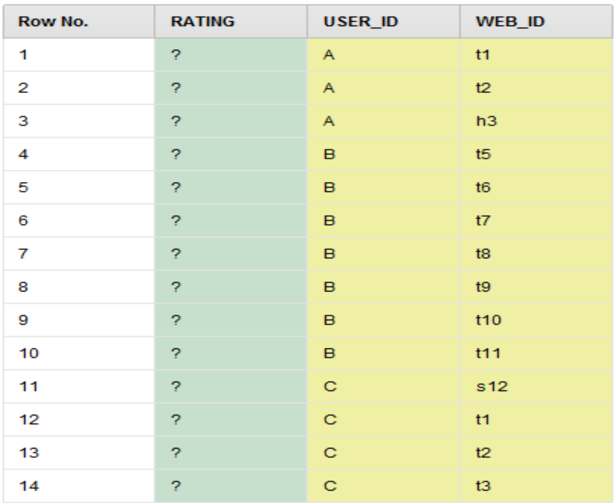


Fig 3: Use of filter operator

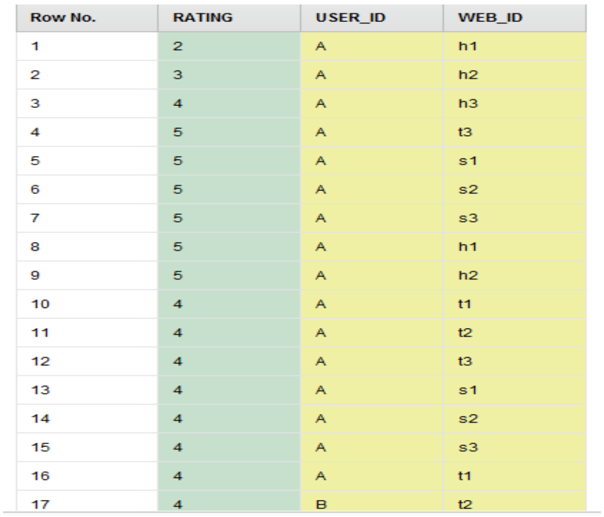
Figure 4 shows the missing values that are to be predicted for the rating based on the previous browsing history of the user. There are total 14 examples set having missing attribute values



Row No.	RATING	USER_ID	WEB_ID
1	?	A	t1
2	?	A	t2
3	?	A	h3
4	?	B	t5
5	?	B	t6
6	?	B	t7
7	?	B	t8
8	?	B	t9
9	?	B	t10
10	?	B	t11
11	?	C	s12
12	?	C	t1
13	?	C	t2
14	?	C	t3

Fig 4: Attribute showing unrated values

Figure 5 workflow shows the example set having no missing values. This consists of the remaining 75 example set.



Row No.	RATING	USER_ID	WEB_ID
1	2	A	h1
2	3	A	h2
3	4	A	h3
4	5	A	t3
5	5	A	s1
6	5	A	s2
7	5	A	s3
8	5	A	h1
9	5	A	h2
10	4	A	t1
11	4	A	t2
12	4	A	t3
13	4	A	s1
14	4	A	s2
15	4	A	s3
16	4	A	t1
17	4	B	t2

Fig 5: Filtered rated values

3.2 Data Mining

Data mining is a step of the knowledge discovery in database process. It is mainly concern with the algorithmic means by which patterns or structures are extracted from the data under computational efficiency limitations. The user k-NN technique is used for predicting the rating. The cosines similarity measure is used in the technique.

Cosine similarity is one of the most commonly used distance metric which is used in text analysis. For a given vector A and B of length n, here cosine similarity is calculated as the dot product of two vectors. When the angle is 0, then the result of the cosine function is equal to 1. And when the angle is of any other value the resultant of the cosine function is less than 1.

$$\text{Similarity} = \cos(\theta) = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_j \times B_j}{\sqrt{\sum_{i=1}^{n} (A_j)^2} \times \sqrt{\sum_{i=1}^{n} (B_j)^2}}$$

Where A and B are the dot product the two vector.

The user k-NN operater is applised along with filter example operator to build the model. The filter operater is used to reduce the number of observations. In this model the filter operater has divided the input data into two part. Then the apply model opertor is the applied to combine the two sub model to construt the original model. Flow diagram of which is shown below.

In the Figure 6 below, the first operator is used to load data from Microsoft Excel and read the example set from the specified excel file using Read Excel operator. Following, the filtering of given example set is done on the basis of no-missing-attribute parameter this is done by filter operator. Next the appropriate rules are set to attribute using Set Role operator. The user identification role is set to user id attribute and item identification role is set to item id attribute in the both Set Role operators. Also, a target role is given as label to the attribute whose values are to be predicted. Next the output with appropriate set rules is trained for predicting the rating by using User k-NN operator. Now the output is ready for the implementation of algorithm by using the Apply Model operator. Apply Model consist information about the data he has been trained

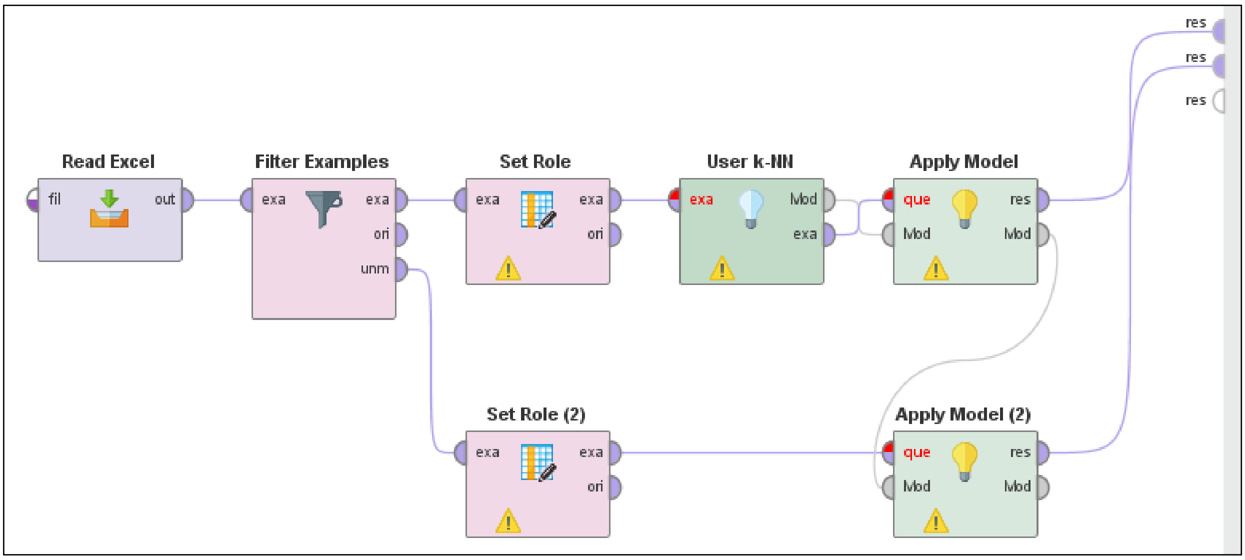
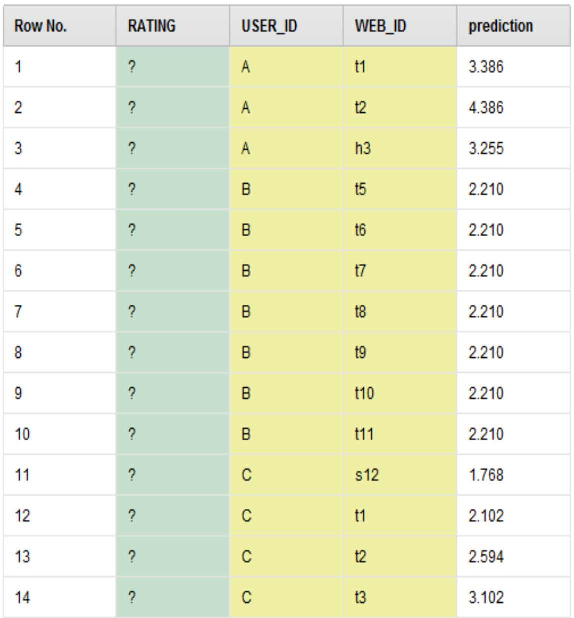


Fig 6: Workflow to find the prediction for the unrated values



Row No.	RATING	USER_ID	WEB_ID	prediction
1	?	A	t1	3.386
2	?	A	t2	4.386
3	?	A	h3	3.255
4	?	B	t5	2.210
5	?	B	t6	2.210
6	?	B	t7	2.210
7	?	B	t8	2.210
8	?	B	t9	2.210
9	?	B	t10	2.210
10	?	B	t11	2.210
11	?	C	s12	1.768
12	?	C	t1	2.102
13	?	C	t2	2.594
14	?	C	t3	3.102

Fig 7: Result showing the predicted rating which the user has not rated

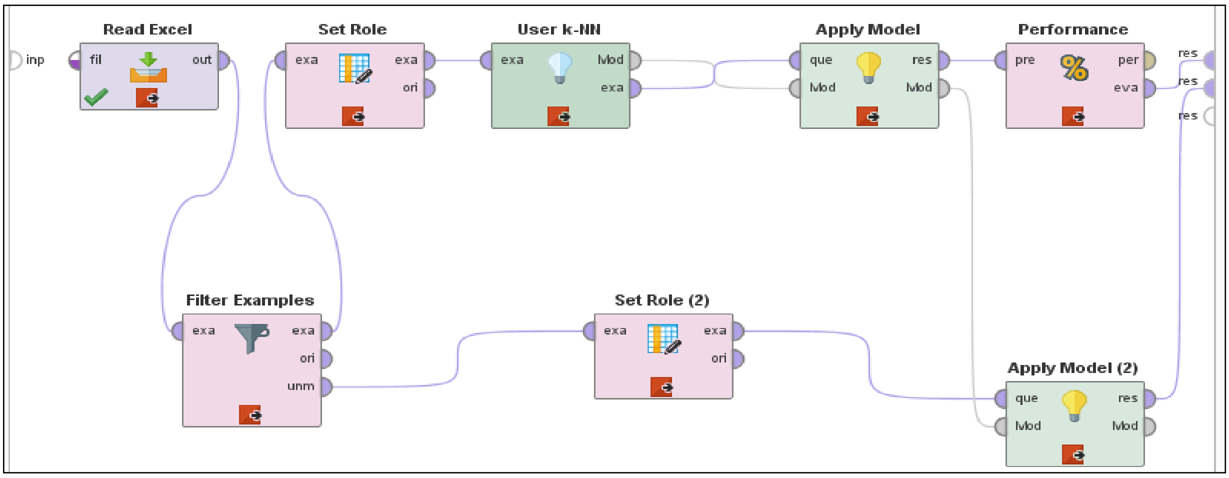
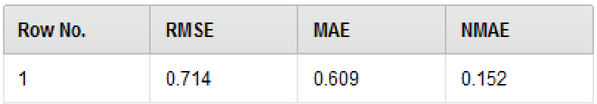


Fig 8: Workflow showing complete model with user k-NN

3.3 Evaluation of data

Now the performance of the built recommendation model is to be calculated and this is done by the Performance operator. The result of Performance operator is shown in the form of (RMSE), (MAE) and (NMAE).



Row No.	RMSE	MAE	NMAE
1	0.714	0.609	0.152

Fig 9: Result showing the performance of rating prediction model

The result of performance operator are dicussed below.

3.3.1 RMSE

The regression lines predict the average of ‘y’ value in association with an ‘x’ value. To predict the value of ‘y’ root-mean-square error (r.m.s.e.) is used. RMSE is very commonly used for an excellent general purpose error metric for numerical predictions.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \overline{y_j}\right)^2}$$

It is not necessary that root-mean-square error increase with the variance of the errors. It increases with the variance of the frequency distribution of error magnitudes [9].

3.3.2 MAE

Mean absolute error (MAE) is a standard error for measuring rating prediction or it is measuring of difference between the two variables. Suppose y and x are the two variables. MAE is the average vertical along with average horizontal distance between each point and y=x line.

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \overline{y_j}\right|$$

Where n is no. of observations.

3.3.3 NMAE

The normalized mean absolute error (NMAE) is additionally normalized to make it independent of the rating scale.

4. CONCLUSION AND FUTURE WORK

Recommender system is a part of machine learning, which automatically learns from the experience rather than the predefined data. The study makes use of the user k-NN method to predict the ranking for the pages. The k-NN technique is a type of supervised classification method which is based on prediction. In this paper prediction is generated based on the previous browsing history of the similar type of user’s. After this the performance of the model is also measured to know the efficiency of the model. As for the part of future work, the real dataset can be used that will provide more accuracy for the real time result. In this paper dummy dataset is used. In future other operators can be used to calculate the accuracy of the model.

5. REFERENCES

[1] Rajhans Mishra, Pradeep Kumar and Bharat Bhasker, “A Web Recommender System Considering sequential information”, Decision Support Systems 75(2015) 1-10.

[2] P. Kumar, P.R. Krishna, R. S. Bapi and S.K. De, “Clustering using Similarity Upper Approximation”, IEEE International Conference on Fuzzy Systems, Vancouver, 2006, pp. 893-844.

[3] Rajhans Mishra, Pradeep Kumar, “Clustering Web Logs Using Similarity Upper Approximation with Different Similarity Measures”, International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2 012.

[4] Prajyoti Lopes, Bidisha Roy, “Dynamic Recommender System using Web Usage Mining for E-commerce Users”, International Conference on Advanced Computing Technologies and applications, Procedia Computer Science 45(2015) 60-69.

[5] Maryam Jafari, Farzad soleymani Sabzchi and Amir Jalili Irani, “Applying web usage mining Techniques to design effective web Recommender systems: A case study”. ACSIJ Advances in Computer Science: an International Journal, Vol. 3, Issue 2, No. 8, March 2014.

[6] Shiva Nadi, Mohammad Hossein Saraee, Ayoub Bagheri, “A Hybrid Recommender System for Dynamic Web Users, International Journal Multimedia and Image Processing (IIJMIP), 2011, 1(1).

[7] B. Sarwar, G. Karypis, J. Kostan, and J. Riedl. “Analysis of recommendation algorithms for e-commerce”. In EC, pages 158–167, 2000.

[8] Rapid miner Studio, Documentation //docs.rapidminer.com/studio

[9] Google www.google.com