Big Data Analytics of Movie Rating Predictive System

Neeta Choudhry¹, Jianyun Xie², Xiaoling Xia³

¹School of Computer Science & Technology, Donghua University, Shanghai, 201620, China. Email: 318062@mail.dhu.edu.cn
²Building 2, Lane 955, Jujiqiao Street Pudong New Area, China. Email: xiejianyun@beyondbit.com
³School of Computer Science & Technology, Donghua University, Shanghai, China. Email: sherlysha@dhu.edu.cn

Abstract. In every organization the data is a significant part that can be separated as structured, unstructured and semi-structured. Unstructured data cannot be administered in the real-time by RDBMS or Hadoop. PySpark can be used for realtime data analysis of movie rating data collection. To generate the modified recommendations, method is intended that is Recommender Systems. Some complications that a Recommender System encountered are data sparsity, cold start, popular bias and scalability. In this research paper collaborative technique is applied for m-ovie rating recommender system. By getting benefits of collaborative filtering technique, we conclude that we use perspective metadata that is freely accessible. Concealed suggestions occur in the middle of movie ratings to personalize recommendations. Our recommended methodology demonstrate novelty by providing the modified recommendations irrespective of research field and of customer’s capability. By the usage of MovieLens dataset a significant development above supplementary standard techniques in evaluating global enactment and capability to yield appropriate and top-ratings at the top of the recommendation list.

1. Introduction

Recommender system is used to make enhanced and supportive user understanding as it is an intelligent method. By streaming data processing, calculation performed not as a batch but in Real-Time when data appear. For various great organizations, real-time processing and analytics are becoming important part of big data approach. In the world of big data, predictive systems are flattering trendier because this automated tool unites clients toward the products that are best well-matched to them by linking product content and uttering feedback. Clients may find it hard-hitting to choose the best package that encounters their specific interest and necessity. Collaborative Filtering is to be used in our paper for the construction of recommendation system in real world as it is one of the best efficacious technique and automatic predictions about costumer’s concern can be measured by using algorithms in Collaborative Filtering. It can predict the preferences of additional indefinite customers, and offer personalize recommendations by analysing the preferences of current parts of customers and has been broadly recycled in profitable websites comprising Amazon, Netflix, eBay, Taobao, Hulu etc [1,2] [Khorasani, E. S., et al 2016] [Yang, Z., et al 2016]. Recently many efficacious fallouts attained using collaborative filtering algorithm with Hadoop. For unstructured data analysis, Hadoop is best as it doesn’t request for definite data type and inexpensive product hardware is used as data knobs for storing data and calculations of data. But the unstructured data cannot be managed by Hadoop in real-time. So that Hadoop architecture introduced Spark as its extension that is having characteristics of both Storm and Hadoop. Many computer languages such as Java, Python, R and Scala are supported by Spark and Python is one of the firmest evolving language having infinite lending library maintenance that is why we rely on this language. In my paper, for analysing movie
review dataset, the PySpark Data frame Transformation and Actions is used. For the accomplishment of proficiency, RDD is used by Spark. It focuses on commemoration of operations that caused in modern RDD and attains fault tolerance to acquire RDD. ALS (Alternating Least Squares) model is collaborative filtering model that will be used in our work. This effort offers a new instruction to the user based on Apache Spark and Spark RDD strategies combined with the mutual size reduction and clustering of the Spark MLIB algorithms. MLlib is machine learning library that is accessible by Spark and contain collective algorithms and services, these are Classification, Clustering, Collaborative filtering, Regression, size reduction and also the essential optimization primitives.

**Objectives**

Our goal is:

1) Implementation of Spark MLlib to deliver top-of-the-line film recommendations, users who are rated video based on video delivered by particular customers and others. The use case in this paper is the Recommendation Engine that we actually want to implement. The proposed recommendation engine is used to predict the unknown customer projects Association based on known customer element Associations. They can be used to make predictions. Use the user to select another project and other users to select this particular project.

2) Further our intensions to write this paper are as follows:

1 – First of all, top executed movies depending on feedback and review of costumer are analyzed.
2 – Second goal is to perform elementary Spark RDD operations on Machine learning (100k dataset) to bring top 10 trendier movies.
3 – At the end, to perform Spark Data frame operations on Machine Learning (latest dataset) to get top 10 ranked movies.

I will also focus on construction of a collaborative filtering engine. In collaborative filtering, there are characteristically subsequent tasks:

- cold start
- data sparsity
- popular bias
- scalability
- pool relationship between like-minded yet sparse users

**Problem statement**

In this research main focus is to build a Movie Predictive System that will recommend movie based on other customer’s ratings for dissimilar movies, the preferences of customer, by using Apache Spark we will analyse and assume the superlative executed movies based on customer opinion and reviews. Huge amount of data will have to be processed and input is from multiple sources. System would be easy to use and having a fast processing.

**Solution**

Use matrix factorization technique to train model to learn user-item interaction by capturing user information in user latent factors and item information in item latent factors. Meanwhile, matrix factorization technique can significantly reduce dimensionality and sparsity and it will reduce huge amount of memory footprint and make our system more scalable. This project will focus on collaborative filtering and use ALS (Alternating Least Squares) algorithm to make movie predictions.

**Significance**

As this research show big data flow exploration to user-based collaborative filtering technique depending on Apache Spark stage having combination of Spark data frames and spark RDD through Spark ML algorithms. Hadoop processes the data stored over a period of time. This work will be beneficial for big data analysts, data scientists, statisticians, predictive modellers as well as other analytic professionals to study growing volumes of transaction data.
2. Related Work

Similar to my research, there are also many previous researches that relate to movie rating appeal recommendation system by using Apache Spark. Here previous research work will be discussed.

[Ruining He, & Julian McAuley. 2016][3] They proposed that to construct a recommender system, the understanding about client’s performances and emotions are mandatory. To construct such type of system, the dataset is rare. With the help of previous feedback, a new modified recommender system is established. Over a period of time the journalist correspondingly search the changing fashion tendency.

[Li Zhuang, Feng Jing, & Xiao-Yan Zhu. 2006][4] They proposed that, for the companies, reviewers and readers of review, the reviews are very advantageous. The companies attract the clients by the reviews and reviewers are given the incentives and the readers will be able to read the reviews and distinguish that the product that they want to buy is comparatively better than other product or not.

[Lina L Dhande and Girish K Patnaik. 2014][5] In their paper they discussed to know about movie is worthy or ruthless, simple classifier along with neural network is applied and for the grouping, the unigram feature is recycled.

[Callen Rain. 2013][6] The author said, when 15 products of Amazon gets 50000 reviews then It can be classified that if the clients loved the products or not. In this research simple classifier is taken for grouping. Kindle is a product that is the focus of study.

[Neelu Rani, Nishant Singh, SujayPawar, et al. 2017][7] They classified positive or negative opinions of people by SVM based classifier on movie review data.

[Xiaomeng Su.] [8] Showed a diagrammatic explanation in his paper.

[J. Sangeetha and Dr. V. Sinthu Janita Prakash. 2019][13] In proposed research method MoCFCR Modified Collaborative Filtering and Clustering with Regression is presented for the accuracy of movie review classification. K means algorithm is used for clustering process. In this way the movie review is done in an easy way. The overall methods performed in Hadoop so proving that the research is giving better outcome than previous research paper’s outcome.

[G. Bathla, 2017][14] the researchers now-a-days are arguing on the societal systems and also about the big data. By the usage of societal system graphs, the journalist deliberated content based and collaborative filtering (CF) in encouragement or guidance of societal reliance.New methods of recommendations are projected using item rating matrix, by use of Pearson correlation coefficient the comparison between sets are to be intended.

| S# | Paper Title                                                                 | Advantage(s)                                                                 | Disadvantage(s)                                                                 |
|----|------------------------------------------------------------------------------|------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| 1  | A Collaborative filtering recommendation Engine in a Distributed Environment  | This model is more robust to data sparsity.                                 | This model is less scalable Data mining or machine learning techniques can improve the prediction quality. |
| 2  | A Hybrid Recommendation algorithm based on Hadoop.                           | This model improves the speed and achieves better performance for large data set. | It has low prediction quality.                                                  |
| 3  | A Hybrid Distribution Collaborative filtering Recommendation Engine Using Apache Spark. | This model improves the speed and provides good prediction quality.           | Less Scalable as it fails for the large dataset.                                |
| 4  | Design and implementation of an Intelligent System for Tourist Routes Recommendation based on Hadoop. | This model improves the scalability.                                         | It has poor prediction quality.                                                 |

3. Methodology

To bring top 10 trendier movies we do calculation by performing spark RDD tasks on ML (100k dataset) and then top 10 classified movies are to be brought by using spark Data frames tasks on ML. By using ratings data, ALS algorithm is used to train the model. A research method is taken out form MovieLens data as dataset that is for the recommendation of movie for user-based on preference. Apache spark is a technique or platform that used both Spark RDD operations and data frames in my research and as Spark supports programming languages, here in this paper I will use the Python
programing language for coding our program and Apache Spark Machine Learning (ML) Algorithm MLLib (machine learning library) will be used.

3.1. Project Flow
Our Project contains the following phases to get at the top movie recommendations for customer.

- Load Data
- Spark SQL and OLAP
- Spark ALS based approach for training model
- ALS Model Selection and Evaluation
- Model testing
- Make movie recommendation to myself

![Figure 1. Sample Application Architecture](image)

3.2. Big Data Analytics
It is not accord here as how big data is described, as various popular “hyped” expressions are now available in markets. These expressions are frequently used by correlated impressions like data mining and BI (Business Intelligence). These expressions are truly about unconventional analytic circumstances and analysing data. However from these two expressions, the big data theory is diverse as there is much bigger and multifarious volume of data, number of transactions and also number of data sources and to pull vision out of data, distinctive techniques and skills are obligatory. For example when allocating big data may be the outmoded warehouse clarifications collapse undersized [Xiaomeng Su.] [8].

3.2.1. The Four V’s of Big Data
There are certain precise characteristics to describe big data. These are called the four V’s because first letter of all characteristics is V [Xiaomeng Su.] [8].

1- Volume (data size): For flawless determination, Volume of Big Data is one of the most important as it is very large data quantity from the datasets having the dimensions of terabytes to petabytes or to zettabyte.

2- Velocity (speed of change): From dealings, a very huge volume of data having extraordinary revive frequency causing in data streams upcoming at countless speed and the time will be very minute often by these data streams. Since batch processing to real time streaming, there is a change.

3- Variety (data structure): Different data sources are used to fetch the data. Firstly, data can be fetch by both internal and external data sources. Using numerous applications, semi-structured data, structured data and different data formats are more significantly fetched like transaction as well as log data. There is a change by sole structured data to more unstructured data increasingly or hybrid of these two.
4- Veracity: It is a term used to show trustworthiness and authenticity and states doubt of availability of data. Sometimes available data is challenging to be trusted and disordered. Quality and accuracy with several procedures of big data are tough to control. Big data and analytics technology grants the permission to work with such types of data. Some business owners do not trust the information used for making the decisions that is doubtful.

3.3. Apache Spark
Apache Spark is basically a platform or an open source handling background of big data constructed nearby haste, easy to use and trendy analytics. Apache Spark came into existence in UC Berkeley’s AMPLab in 2009 and as an Apache project it was located in 2010. It is advantageous as it replaced Map Reduce and Big data technologies such as Storm and Hadoop. It is better because it supports continuous data streaming, SQL requests, matrix data mining as well as Deep learning. These proficiencies would be used individually or mutually by the developers for running a solo data channel use case. Spark enables us to write the applications in a quick manner in Scala, Python or Java and having set of above 80 high-level built-in operators. And used for the data query purpose in shell interactively.

3.3.1. Spark Collaborative Filtering
CF (Collaborative filtering) [9] is an algorithm that depends on ALS (alternating least squares) method which is used to train, validation, and test datasets [10] and is recycled to produce recommendations. Automatic predictions is generated by it according to the concern of the consumer with the help of collaborating or gathering preferences commencing various costumers. An essential supposition of the collaborative filtering methodology if two person (e.g. person A & person B) having equivalent judgement on any kind of issue then person A’s judgement will be more probable to have person B’s judgement rather than a customized person. Collaborative filtering (CF) algorithm is very light in computation and very easy to execute in a very limited period of time. Collaborative filtering is created to execute the workflow according to clients demand in the training set for each and every client.

3.3.2. Spark MLlib
Spark MLlib is basically a Machine Learning library of Spark. Its main focus is on making of useful machine learning accessible and stress-free. It includes common learning algorithms and utilities such as regression, classification, clustering, collaborative filtering, dimensionality reduction (size decreasing) and even lower level optimization primitives and higher level channel APIs.[ Nitesh Ghimire ][11].Spark MLlib Machine Learning (ML) algorithms will be applied in this paper for better results.
4. Experiments and Evaluation Metrics

4.1. Datasets
Large (20 million) MovieLens dataset is used in my research. Here all the dataset is comprising numerous files that belong to Movies and Movie ratings. File rating.csv and movies.csv will be used here. In file rating.csv dataset consist of 20 million movie ratings from circa 130000 users and prearranged as movieID, userID, timestamp and rating. Rating to each movie is placed or stored in every row that is acknowledged by movieID, by one of the users acknowledged by userID. Further in the file movies.csv the dataset comprise circa 27000 movies that is organized as Category, MovieID and title as well.

4.2. Performance Evaluation Criteria
I use statistical precision metrics for the assessment of the precision of the recommendation system. RMAE (Root Mean Absolute Error) is metric in the recommendation system that is broadly practiced using collaborative filtering (CF) for the measurement of deviation of the recommendations by their factual customer definite ratings. Here is the formula given below.

\[ RMAE = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (\text{Predicted}_j - \text{Actual}_j)^2} \]

This is the equation that indicates the average of the actual variance in the center of n pairs; forecast tallies of consumer ratings and definite customer ratings in the test dataset for user-item pairs. The decreased importance of RMAE would contribute to the precision of estimation of consumer ratings by improved recommendation method.

5. Result
Basic purpose to write this paper is to build a recommendation system that would be able to predict what kind of movies are to recommend to the customers. To attain the better results we add rating and
movies to test and train datasets then our model is trained by using our ratings afterwards the RMSE is calculated for this innovative model on the assessment set and in this way we predict what ratings would be given to the movies by you, which ratings are not delivered previously by you.

### Table 2. 80-20 Dataset

| Iteration# | λ = 0.01 | RMSE | λ = 0.1 | RMSE | λ = 1 | RMSE | λ = 2 | RMSE | λ = 3 | RMSE |
|------------|-----------|-------|----------|-------|-------|-------|-------|-------|-------|-------|
| 1          | 0.01      | 0.894266 | 0.1 | 2.897784 | 1 | 3.66501 | 2 | 3.681256 | 3 | 3.68338 |
| 3          | 0.01      | 0.909961 | 0.1 | 0.855669 | 1 | 1.370541 | 2 | 3.478916 | 3 | 3.681486 |
| 5          | 0.01      | 0.94266 | 0.1 | 0.837761 | 1 | 1.327533 | 2 | 2.209487 | 3 | 3.672479 |
| 10         | 0.01      | 0.891064 | 0.1 | 0.823867 | 1 | 1.323342 | 2 | 2.176589 | 3 | 3.419291 |
| 15         | 0.01      | 0.882942 | 0.1 | 0.821514 | 1 | 1.323342 | 2 | 2.176588 | 3 | 3.123759 |
| 20         | 0.01      | 0.88883 | 0.1 | 0.819942 | 1 | 1.323343 | 2 | 2.176588 | 3 | 3.111463 |

### Table 3. 60-40 Dataset

| Iteration# | λ = 0.01 | RMSE | λ = 0.1 | RMSE | λ = 1 | RMSE | λ = 2 | RMSE | λ = 3 | RMSE |
|------------|-----------|-------|----------|-------|-------|-------|-------|-------|-------|-------|
| 1          | 0.01      | 3.493471 | 0.1 | 3.057215 | 1 | 3.663589 | 2 | 3.680112 | 3 | 3.682051 |
| 3          | 0.01      | 0.949115 | 0.1 | 0.863132 | 1 | 1.366073 | 2 | 3.480148 | 3 | 3.68017 |
| 5          | 0.01      | 0.933184 | 0.1 | 0.845621 | 1 | 1.325257 | 2 | 2.199994 | 3 | 3.671156 |
| 10         | 0.01      | 0.933073 | 0.1 | 0.835627 | 1 | 1.323686 | 2 | 2.175678 | 3 | 3.415876 |
| 15         | 0.01      | 0.93275 | 0.1 | 0.832075 | 1 | 1.323686 | 2 | 2.175677 | 3 | 3.12107 |
| 20         | 0.01      | 0.93258 | 0.1 | 0.830576 | 1 | 1.323686 | 2 | 2.175678 | 3 | 3.109063 |

Graph 1. RMSE Values with 80-20 Dataset

Graph 2. RMSE Values with 60-40 Dataset
6. Conclusion

By the use of ironic set of approaches Spark’s MLlib library offers accessible data analytics. For Collaborative Filtering (CF) the employment of ALS (Alternating Least Squares) is one that is a best fit in the recommendation engine. Owing to its precise environment, collaborative filtering (CF) technique is a very expensive technique as it subsequently have need of modernizing its model as soon as innovative customer predilections thrash out. Consequently, to have a scattered calculation engine like Spark for the performance of model computation in any real-world recommendation engine is significant such as we have recently fabricated here in this research paper.

We have given explanation on how a model via Spark is built, how particular parameter selection is accomplished by using a reduced dataset, and how the model is updated every time that advanced customer predilections arise. Furthermore, the usage of recommender is also clarified for poles apart circumstances and the joining of the results with product metadata (e.g. movie titles) are also explained for the presentation of results in an appropriate manner.

In forthcoming research we will ensure detailed progressive study and move to enhanced upshots and tend to use our projected model in a web environment for the production of online movies recommendations or predictions.

7. Acknowledgments

I am delighted to convey my appreciation to my advisor Prof. Xia XiaoLing for her excellent supervision and motivation in the completion of this project.

I want to thank my husband, Kapil Dev Choudhry, for his never-ending support and encouragement during my master’s studies at Donghua University Shanghai, China.

8. References

[1] Khorasani, E. S., Zhenge, Z., & Champaign, J. (2016, December). A Markov chain collaborative filtering model for course enrollment recommendations. In Big Data (Big Data), 2016 IEEE International Conference on (pp. 3484-3490). IEEE.

[2] Yang, Z., Wu, B., Zheng, K., Wang, X., & Lei, L. (2016). A Survey of Collaborative Filtering-Based Recommender Systems for Mobile Internet Applications. IEEE Access, 4, 3273-3287.

[3] Ruining He, & Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In proceedings of the 25th international conference on World Wide Web, pages 507–517. International World Wide Web Conferences Steering Committee,2016

[4] Li Zhuang, Feng Jing, & Xiao-Yan Zhu. Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management, pages 43–50. ACM, 2006.

[5] Lina L Dhande and Girish K Patnaik. Analyzing sentiment of movie review data using naive bayes neuralclassifier. 2014.

[6] Callen Rain. Sentiment analysis in amazon reviews using probabilistic machine learning.2013.
[7] Neelu Rani, Nishant Singh, Sujay Pawar, et al. Sentiment analysis by data mining of past movie reviews/ratings. Imperial Journal of Interdisciplinary Research, 3(6), 2017.

[8] Xiaomeng Su. Introduction to bigdata. NTNU.

[9] “Collaborative Filtering”, Wikipedia https://en.wikipedia.org/wiki/Collaborative_filtering

[10] Y. Koren, R. Bell, C. Volinsky, “Matrix Factorization Techniques for Recommender Systems“, in Computer Journal, Volume 42 Issue 8, August 2009, Pages 30-37 https://dl.acm.org/citation.cfm?id=1608614

[11] Nitesh Ghimire. Movie Rating Prediction System By Using-: "Apache Classification Techniques for Large Scale Datasets With "Apache Spark"

[12] “Collaborative Filtering. RDD based API” The Spark MLlib implementation http://spark.apache.org/docs/latest/mllib-collaborative-filtering.html. 2014.

[13] J. Sangeetha and Dr. V. Sinthu Janita Prakash. International Journal of Mechanical Engineering and Technology (IJMET) Volume 10, Issue 02, February 2019.

[14] G. Bathla,” A graph-based model to improve social trust and influence for social recommendation,” The Journal of Supercomputing, 2017, 1-19