ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

Addressing Interpretability and Cold-Start in Matrix Factorization for Recommender Systems

Kavita Mogal, Prof. Kurhade N.V

P.G. Student, Department of Comp Engineering Sharadchandra Pawar College of Engineering, Otur, Pune, India

Professor, Department of Comp Engineering Sharadchandra Pawar College of Engineering, Otur, Pune, India

ABSTRACT: Online buying is the most favorable in 21st century, around one trillion items should be buy and sold through using online shopping. Recommender systems suggest to users items that they might like (e.g., news articles, songs, movies) and, in doing so, they help users deal with information overload and enjoy a personalized experience. One of the main problems of these systems is the item cold-start, i.e., when a new item is introduced in the system and no past information is available, then no effective recommendations can be produced. The item cold-start is a very common problem in practice: modern online platforms have hundreds of new items published every day. To address this problem, we propose to learn Local Collective Embeddings–a matrix factorization that exploits items' properties and past user preferences while enforcing the manifold structure exhibited by the collective embeddings. We present a learning algorithm based on multiplicative update rules that are efficient and easy to implement. Experiments on two item cold-start use cases: news recommendation and email recipient recommendation, demonstrate the effectiveness of this approach and show that it significantly outperforms six state-of-the-art methods for item cold-start. In This scenario B2B as well as B2B problem occurred frequently. With this research we proposed an approach which will eliminate the matrix factorization problem using cold start recommendation for online users. The experimental analysis shows how proposed system provide the better results than classical algorithms.

KEYWORDS: co-clustering, overlapping co-clustering, cold-start, product recommendations, interpretability.

I. INTRODUCTION

In B2B recommender systems, only positive ratings are present: the products that the clients have already purchased. Recommender systems are aimed to help users of online platforms to deal with the large volumes of information and to provide them a personalized experience. This is achieved by suggesting items of interest to the users based on their explicit and implicit preferences. Recommender systems use a number of different technologies, but may be broadly classified into two groups: content-based and collaborative filtering systems. Content-based systems examine the properties of the items and recommend items which are similar to the ones the user preferred in the past. They model the taste of a user by building a user profile based on the properties of the items the user liked, and use the profile to compute the similarity with new items. Items which are most similar to the user's profile are recommended. Collaborative filtering systems, on the other hand, ignore the properties. of the items and base their recommendations on community preferences. They recommend items that users with similar tastes and preferences liked in the past. Two users are considered similar if they have many items in common. One of the main problems for recommender systems is the cold-start problem, i.e., when a new item or user is introduced in the system. In this study we focus on the problem of producing effective recommendations for new items: the item cold-start. Collaborative filtering systems suffer from this problem as they rely on the previous ratings of the users. Content based approaches, on the other hand, may still produce recommendations using the description of the items and are the default solution to the item cold-start. However, they tend to achieve lower accuracy and, in practice, they are seldom the only choice. We introduce a new

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

method for recommendation, LCE, that combines the content and collaborative information in a unified matrix factorization framework while exploiting the local geometrical structure of the data; We propose a simple and efficient learning algorithm, based on multiplicative update rules and prove its convergence; • We conduct an extensive experimental study and we show that the proposed methods outperform six stateof-the-art methods for item-cold start recommendation.

II. LITERATURE SURVEY

Exploiting the local geometric structure of the data to discover better low-dimensional representations has been exploited by Cai et al. [1]. Inspired by the success of using the nearest neighbour graph for label propagation in semisupervised learning, they propose a clustering technique. The algorithm favours factorizations for which similar instances have similar low-dimensional representations. The authors show that, by imposing this constraint, they outperform classical clustering algorithms and classical factorization techniques. In this work, we impose such geometrical constraints but for collective factorization for which we can handle multiple data sources, i.e., the content and collaborative data matrices

The Regression Based Latent Factor model [2] for applications where ”bag-of-words” representation of items is natural. The main idea is regularizing the user and item factors simultaneously through the user features and the words associated with the items. The user ratings are modelled as user’s affinity to the item’s topics, where the user’s affinity to topics and topic assignments to items are learned jointly in supervised fashion. In our experiments we only use the item feature as user features are not available. As recommended by the authors, we run 20 EM iterations with 100 samples, drawn after 10 burn-in samples. In the email recommendation experiment, we test for $k \in \{10, 25, 50\}$ factors, while for the news recommendation experiment we test only $k = 10$ due to the long running times.

This method [3] handles the cold-start in two steps: (1) factorizing the collaborative matrix to learn latent factor representation of the users and items, (2) learning a mapping between the user/item attributes and the corresponding latent factors. When a new user/item arrives in the system, the mapping from (2) is used to infer the factors from the attributes, and then the factors are used to make predictions. As proposed by the authors, we used Bayesian Personalized Ranking (BPR) to factorize the collaborative matrix. To learn the mappings we used the K-Nearest-Neighbour and BPR optimization; however, the kNN mapping was superior in all cases and thus we report the results for only for kNN mapping. This is consistent with their experimental results. We test with different number of latent factors $k \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$.

Shmueli et al. [4] consider a similar scenario of news recommendation (Section 5.3), i.e., predicting the articles a user is most likely to comment on. They combine content based and collaborative filtering approach using a latent factor model. The odds that a user will comment an article are estimated as the inner product of the user and article factors, where the article factors are represented as the sum of the latent factors associated with the textual content (tags–named entities) and the commenter’s. A modification of the model for real-time scenarios is presented in [3]. The authors show that the recommendation accuracy grows as the number of commenter’s grows. However, in an item cold-start scenario articles are not yet associated with commenter’s, thus an can only be represented with the latent factors of the textual tags

Singh and Gordon [5] propose the idea of collective matrix factorization, a general framework for multi-relational factorization models. They subsume models on any number of relations as long as their loss function is a twice differentiable decomposable loss. In their work, they address both rating prediction and item recommendation. The matrix factorization approach proposed in this work is based on a similar idea of collective factorization. However, we enforce non-negativity constraints on the factorization to obtain sparse and interpretable factors, and we consider the specific scenario of cold-start recommendations

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

Soboroff [6] proposed a technique based on Latent Semantic Indexing (LSI) for combining the collaborative filtering input and the document content for recommendation of textual items. The method builds a content profile for each user as a linear combination of the preferred documents. LSI is then applied to the user profiles to discover topics in the collection and implicitly learn commonalities among the user profiles. Incoming documents are projected into the LSI space and compared to user profiles. The documents are recommended to the users with the most similar profiles. The author argues that applying LSI on the user profiles instead of the documents allows one to take into account the collaborative input and consequently improves the recommendation performance. However, the system is not evaluated in the cold-start scenario

According to [7] the problem of item cold-start is of great practical importance because of two main reasons. First, modern online platforms have hundreds of new items everyday and effectively recommending them is essential for keeping the users continuously engaged. Second, collaborative filtering methods are at the core of most recommendation engines, as they tend to achieve the state-of-the-art accuracy. However, to produce recommendations at the expected accuracy they require that items are rated by a sufficient number of users. Therefore, it is crucial for every collaborative recommender to reach this state as soon as possible. Having methods that produce accurate recommendations for new items will allow enough feedback to be collected in a short amount of time, making effective collaborative recommendations possible.

As the data is intrinsically influenced by time we sort the mailboxes chronologically. We divide the messages in 80% training and 20% testing, resulting in 10 independent train/test subsets. Only the recipients that appear in the training period are considered as potential receivers. We tune the hyper-parameters of the methods on independent validation set, 10% of the training set. Finally, we evaluate the statistical significance of the differences in performance by using a Wilcoxon signed rank test [8].

The output of the algorithms is a ranking of the past recipients of how likely they are to be recipients of the new email. The feedback from the users is explicit, i.e., we have ground truth of who are the recipients of the mail as specified by the user. To evaluate the ranking produced by each algorithm we use the state-of-the-art metrics from Information Retrieval: Micro and Macro F1, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (NDCG) [9]

Recently, matrix factorization techniques have been extensively used in the recommendation systems and topic modelling literature. Many collaborative filtering systems approximate the collaborative matrix by applying techniques such as Singular Value Decomposition (SVD) or UV decomposition [10]. Similar matrix factorization techniques have been used to discover topics in a document collections by decomposing the content, i.e., document-term matrix. Non-negative Matrix Factorization (NMF) is one such approach that factorizes the document-term matrix in two nonnegative, low-rank matrices, where one matrix corresponds to the topics in the collection and the other represents the extent to which documents belong to these topics. Due to the non-negativity constraints, NMF produces a so-called “additive parts-based” representation of the data that increases the sparsity and interpretability of the hidden factors.

III. PROBLEM STATEMENT

The proposed work first investigate data mining approaches like Apriori, FP Tree, FIUT and find the issues of existing system, System also focus on database security like SQL injection with parallel data mining top k retrieval approach on synthetic and real time dataset in distributed framework and provide recommendation for B2B as well as B2C scenario.

Objectives of System

- To design an algorithm for mining Multiple-level association rules from large databases.
- To improve the efficiency of algorithm by reducing the repeated database scan operations.
- To compare the performance of proposed algorithm with the existing algorithms.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

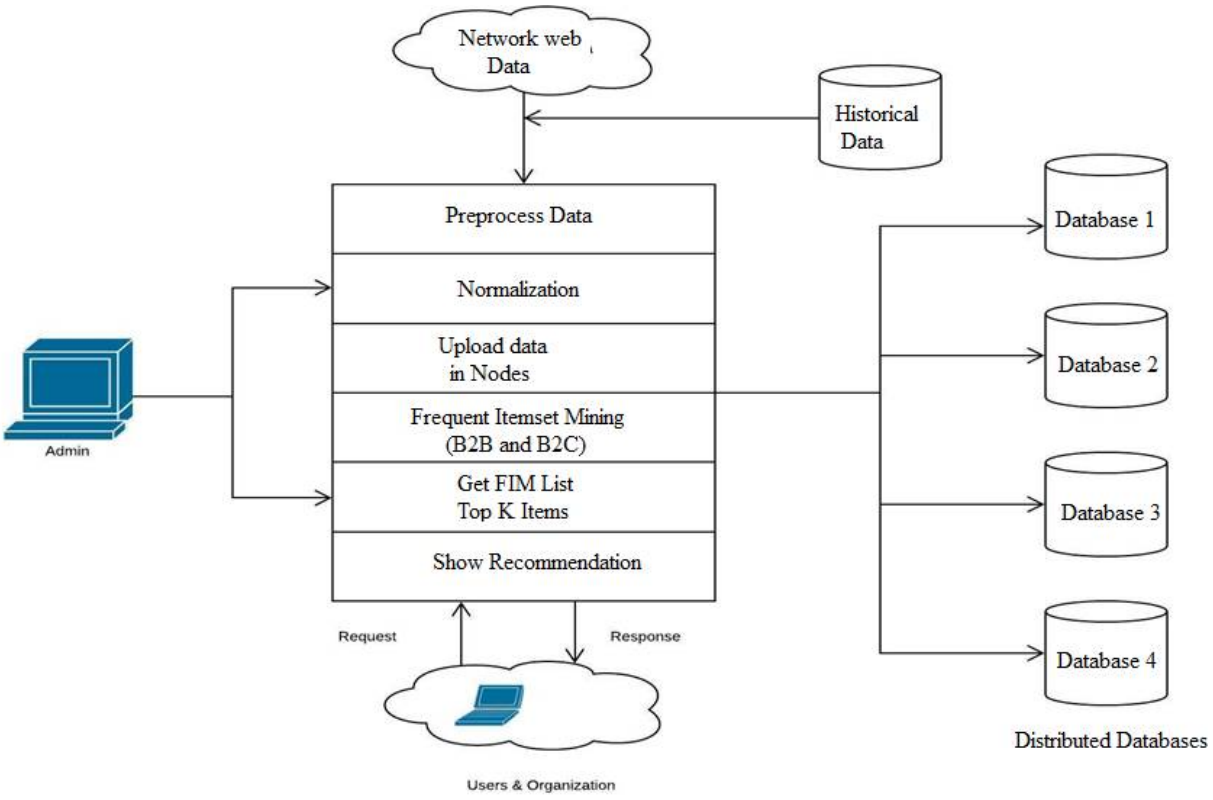
- To implement the algorithm and test it for real time datasets to reduce costly, repeated database scans.
- To embed system on distributed framework and check its functionality.

Scope of System

- With exponential increase in number of everyday internet users there has arise the need to understand their internet usage activity for further improvement in services provided to them.
- Need to improve the e-Business using attractive recommendations.
- The data involved in these cases is a huge chunk of data which needs to be studied thoroughly to understand user needs.
- To provide great user experience to users in their day to day activity this Big Data needs to be analyzed.

IV. PROPOSED SYSTEM

In some cases the decision on whether or not to make a purchase is based largely on price but in many instances the purchasing decision is more complex, with many more considerations affecting the decision-making process before the final commitment is made. Retailers understand this well and attempt to make use of it in an effort to gain an edge in a highly competitive market. Specifically, in an effort to make purchasing more likely, in addition to balancing the salability and profit in setting the selling price of a product.



- System proposed a R-OCuLaR factorization hash or matrix has used for runtime recommendation to end user.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

- In FIM, on Real time transactional dataset apply the proposed algorithms and extract the frequent item sets from historical data
- For the database security apply SQL injection and prevention approach that can be provide the highest security against internal as well as external attackers.
- FP- Tree itemset mining is another example for parallel data mining and frequent item set mining scheme.

V. RESULTS AND DISCUSSIONS

The proposed system has evaluated in distributed environment, for the system performance evaluation, calculate the matrices for accuracy. The system is executed on java 3-tier architecture framework with INTEL 2.7 GHz i3 processor and 4 GB RAM with deep learning approach. The below figure 2 and figure 3 shows the precision, recall and f-score of B2B as well as B2C recommendation respectively.

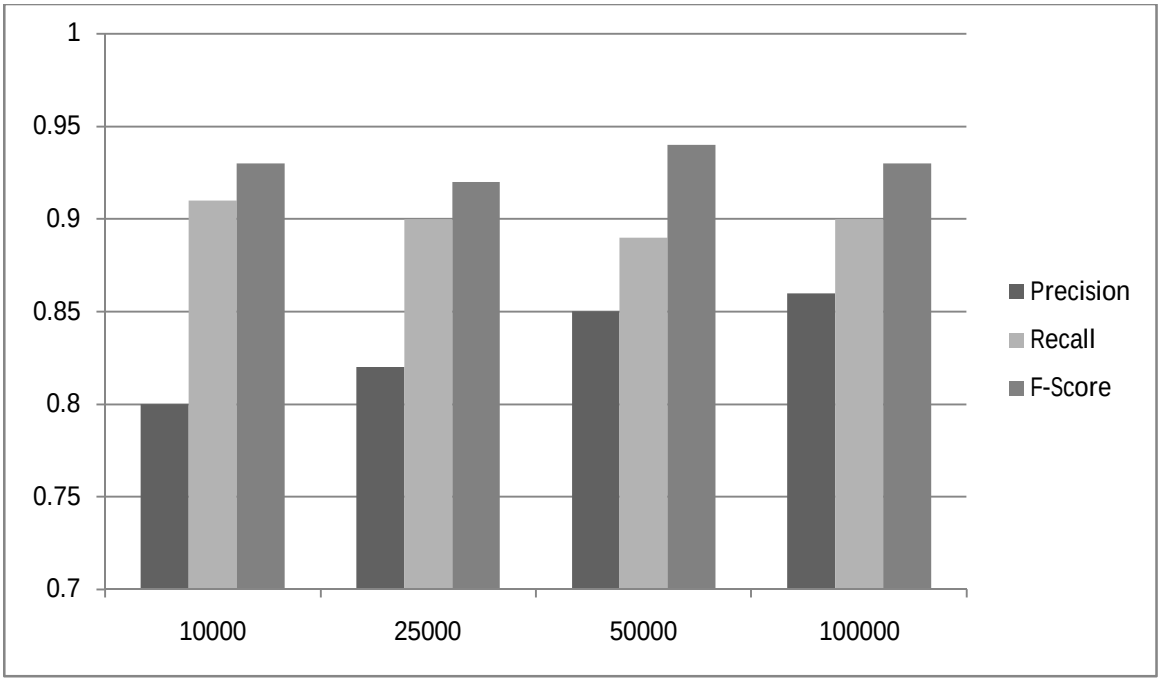


Figure 2 : Performance evaluation of B2B recommendation with various instances

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

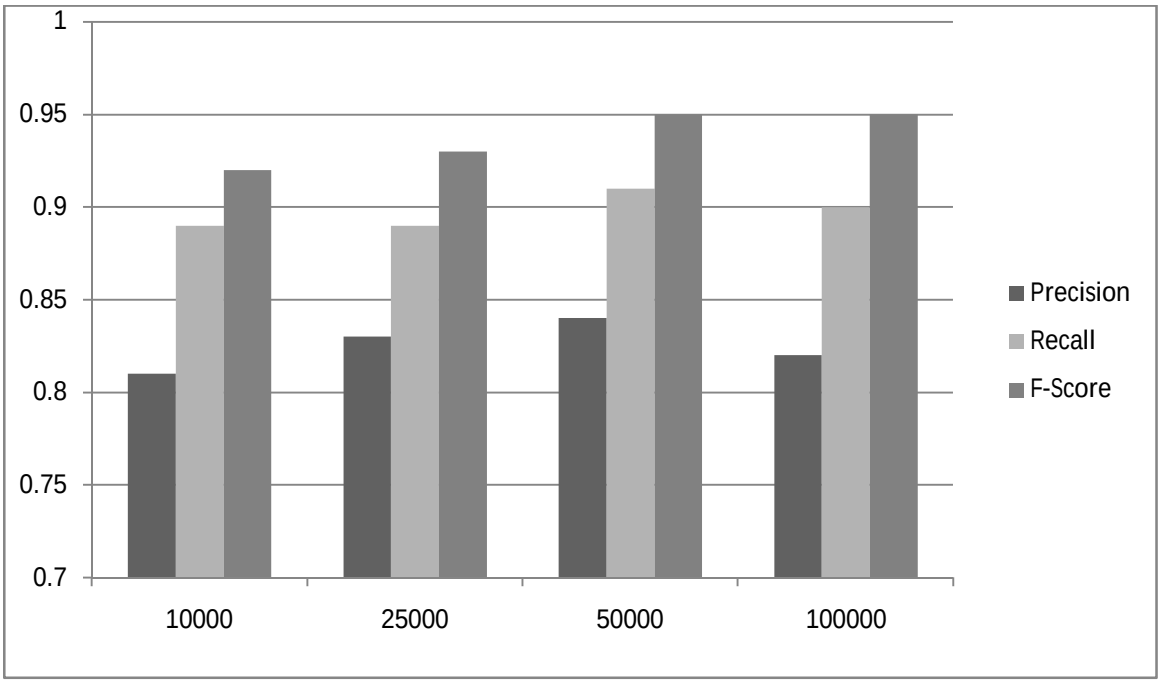


Figure 3: Performance evaluation of B2C recommendation with various size of instances

VI. CONCLUSION

In this research work, the main focus is on the problem of cold start and work on recommender systems. Proposed machine learning has focused primarily on the accuracy of prediction. This work explicitly addresses the aspect of interpretability. By formulating co-clustering as a constrained matrix factorization approach. System also investigate data mining approaches like Apriori, FP Tree, FIUT and find the issues of existing system, System also focus on database security like SQL injection with parallel data mining top k retrieval approach on synthetic and real time dataset in distributed framework and provide recommendation for B2B as well as B2C scenario.

VII. FUTURE WORK

To implement this system with distributed Hadoop Distribution File System (HDFS) environment on large scale data will be the future of system.

REFERENCES

[1] D. Cai, X. He, X. Wu, and J. Han. Non-negative matrix factorization on manifold. In International Conference on Data Mining, 2008
[2] Z. Gantner, L. Drumond, C. Freudenthaler, S. Rendle, and L. Schmidt-Thieme. Learning attribute-to-feature mappings for cold-start recommendations. In International Conference on Data Mining, 2010.
[3] E. Shmueli, A. Kagian, Y. Koren, and R. Lempel. Care to comment?: recommendations for commenting on news stories. In World Wide Web Conference, 2012.
[4] E. Shmueli, A. Kagian, Y. Koren, and R. Lempel. Care to comment?: recommendations for commenting on news stories. In World Wide Web Conference, 2012.
[5] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In ACM Conference on Knowledge Discovery and Data Mining, 2008.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

[6] I. Soboroff. Combining content and collaboration in text filtering. In IJCAI Workshop on Machine Learning for Information Filtering, 1999.
[7] A. T. Gediminas Adomavicius. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 2005
[8] J. Dem˘sar. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research, 2006
[9] R. Baeza-Yates and Ribeiro-Neto. Modern information retrieval. ACM press New York, 1999
[10] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. Computer, 2009.