Addressing Sparsity Data and Cold Start Problem on Collaborative Filtering Recommender System for E-Commerce: A Review

by Mahasiswa Mahasiswa

Journal of Engineering and Applied Sciences
MEDWELL PUBLICATIONS

Addressing Sparsity Data and Cold Start Problem on Collaborative Filtering Recommender System for E-Commerce: A Review

[1,2]Hanafi, [2]Nanna Suryana and [2]Abdul Sammad Bin Hasan Bashari
[1]*Department of Information Technology, University of Amikom Yogyakarta, Yogyakarta, Indonesia*
[2]*Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia*

Key words: Recommender system, collaborative filtering, social recommender, cold start, sparsity data, E-commerce

Corresponding Author:
Hanafi
Department of Information Technology, University of Amikom Yogyakarta, Yogyakarta, Indonesia

Page No.: 2025-2037
Volume: 15, Issue 9, 2020
ISSN: 1816-949x
Journal of Engineering and Applied Sciences

Abstract: Recommender systems are an important technique for creating effective communication between users and retailers in E-commerce services. Good communication and easy to find the product will increase marketing target. On the other hand, will give significant effect to achieving the target value of transactions between users and retailers in online shopping industry. Recommender systems have begun to implement in the mid-90's and many researchers have given the effort to enhance some weaknesses of existing system stronger also because there are many changes of social paradigm and E-commerce industry. One of the models is quite successful recommender system is collaborative filtering, but there is a major drawback of this model is in dealing with the cold start and sparsity of data. The problem rise when new user and new item is coming. There are many solution strategies to handle the problem. In these paper, we show many possible solutions include in there exploring algorithm model and exploiting information from implicit and explicit information that comes from social media, a feature of product/item and user profiles.

INTRODUCTION

Began in the beginning 20th century, the growth of internet users has increased significantly; this is influent E-commerce provider serve millions of items for consumers. Choosing among in too big options of things is very difficult for users. It becomes primary reason how in E-commerce services, recommender system absolute have an important role to serve benefits about consumers and retailers. Also, recommender systems are application tools involve a technique that is providing pieces of advice for items to users or customers what product fit. The advice relates to various decision-making process like what items to buy what music to listen what movie to watch or what online news to read. The first automatically recommender system was developed and implemented by Grouplens[1]. It was used this method to detect Usenet news which are same to be interesting in particular user, according to literature from business and marketing journal has conducted analyzed when recommender system has ability to increase sales improve customer satisfaction[2].

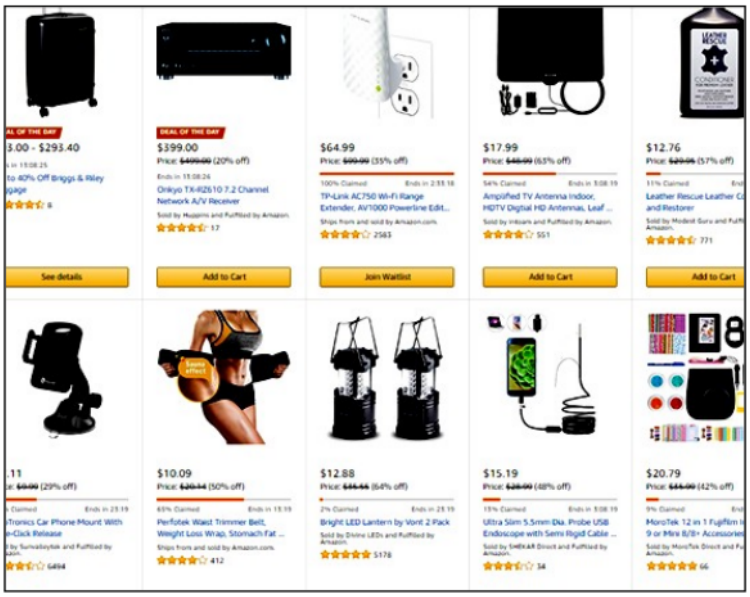

Fig. 1: Implementation recommender system in E-commerce

In the real business world, there are many motivations as to why E-commerce business provider need to exploit this technology[3]:

- Increase quantity of products/items sell
- Increase users/customers Fidelity
- Increase the user service satisfaction
- Sell more diverse items/product

Better understanding of what the users/customers want. Selecting product in E-commerce to need a lot of time and also required fully knowledge about product, so because of this reason very essential needed recommender system, on the other hand, leaving manually technique. Targeted consumers are much better solution to communicate more users in which buyer candidate will show a recommendation about only those products items which they may be interested in.

There is some E-commerce company that very intent and involve large resources to developt system doing research and implemented in the early born of the recommender system, we mention Amazon as a retailer for many variant products, instance most favorite book store[4]. Look at Fig. 1 and 2, the displaying of some items that served by a recommender system.

The large company of movie retailer such as Netflix also companies that focus to develop and implemented Recommender system to serve better recommendation for his film product to customers.

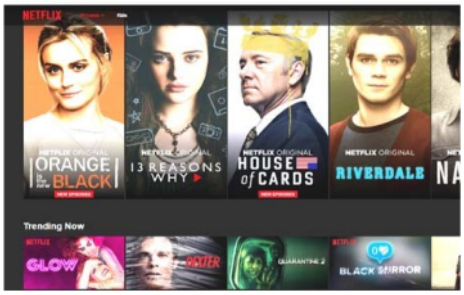

Fig. 2: Sample recommender system for movie

Nowadays specific recommendations that called personalized have a bigger opportunity because those have big potential to increase in availability of information quality from the internet. Reference[5] many E-commerce websites have being implemented recommender system machine for example Movie Lens, Last.fm, Netflix, Amazon, YouTube, Facebook, Google, etc. They create the strategy from all possible information that available from the internet to create optimum user preferences, so that, they can serve the best recommendations to their buyer candidate (Table 1).

According to reference[6, 7], there are four primary technical goals of recommender system to implement in the system retailer.

Relevant: The frequent target of a recommender system is to serve information about items that suitable to users'

Table 1: Sample recommender system in large company

Systems	Product goal
Amazon.com	Books and others product
Netflix	Films streaming, DVD, video
Jester	Jokes
Grouplens	News
Movielens	Movie
Last.fm	Music
Google news	News
Google search	Advertisement
Facebook	Friends, groups, advertisement
Pandora	Music
Youtube	Video streaming
Tripadvisor	Travel products
IMDb	Movies

interest. So, the product fit can found easily. It needed by users or customer to help them to get especially product when they conduct for shopping.

Novelty: Recommender systems are very powerful to help a user find good products or items. The other important things the goals of recommender systems are to serve information that has not seen in the past before. Repeated serving of popular information probably reduction in marketing target.

Serendipity: It means products recommended are unpredictable. The systems can serve better information so users feel lucky to get the information. There no clear definition and no consensus to explain serendipity but serendipity must have include three component are novel, relevant and unexpected.

Diversity: Or increasing recommendation diversity. It means diversity makes sure to users that the information shown is not repeated, so users are not bored.

The basic idea of recommender system is recommending items by learning user's profile, user's previous activities and the kind of items available in the system. Many methods to approaches used by the E-commerce platform to recommend products to user's candidate. Most essential strategies for recommender systems sure content-based, collaborative filtering and hybrid methods. The most powerful and popular approach that called Collaborative filtering is the best approaches build the most efficient recommender systems[8]. Implemented recommender system will increase marketing target[9-11].

MATERIALS AND METHODS

Basic method recommender system: The recommender system is a result from system which involves many processing and many information from users and retailer used a unique method to get better recommendation for users about item products, especially in E-commerce. As basic method[3, 5], there are three algorithm models whom development and implementation in real business world and research area include.

Content based: The system tries to learn to recommend items or product that are similar to the ones that the user liked in the past. The similarity of items will be calculated using the features contents related to the compared items. For example, if a user has rated a film that belongs to the comedy genre, then the system can learn to recommend other movies from this genre. Older content based recommendation procedures go for coordinating the properties of the client profile against the properties of the items. Much of the time, the properties of the item are essentially watchwords that are removed from the items portrayals. Semantic ordering methods speak to the item and users profiles utilizing ideas rather than keyword.

Collaborative filtering: The first and most straightforward execution of this approach makes suggestions to the dynamic user based on item that different user with comparable tastes like before. The comparability in taste of two users is compute in view of the likeness in the rating history of the users. This is the motivation behind why alludes collaborative filtering as "individuals to-individuals connection". Collaborative filtering is thought to be the most famous and generally executed method in recommender systems.

Hybrid recommender system: Hybrid recommender systems these recommender systems depend on the blend of the previously mentioned strategies. The main idea of hybrid system consolidating procedures A and B try to utilize the upsides of A to settle the burdens of B. For example, collaborative filtering strategies experience have big problem in new items issues or that they cannot prescribe items that have no rating. This does not confine content-based methodologies, since, the prediction for new items based on their depiction (includes) that are regularly effortlessly accessible. Given (at least two) essential recommender system methods, several ways have been proposed for mixing them to create new hybrid model.

Variant of rating type: One of important thing to generate recommendation are users activity for example user purchasing, and user clicks to product, items that were seen by user, sometimes, many E-commerce portal support available opinion comment for his/her product. Product interest by users include rating. Rating is representation of interest feel by users about products. It is indicator of degree of interesting user for product.

According to[3, 5, 12] rating play role very important to generate recommendation. There are many kinds of rating type that was implemented in recommendation technique, for examples:

Numerical rating such as the 1-5 start. Many large E-commerce companies using this model instance Amazon, Lazada group, mobile application provider iTunes, play store.

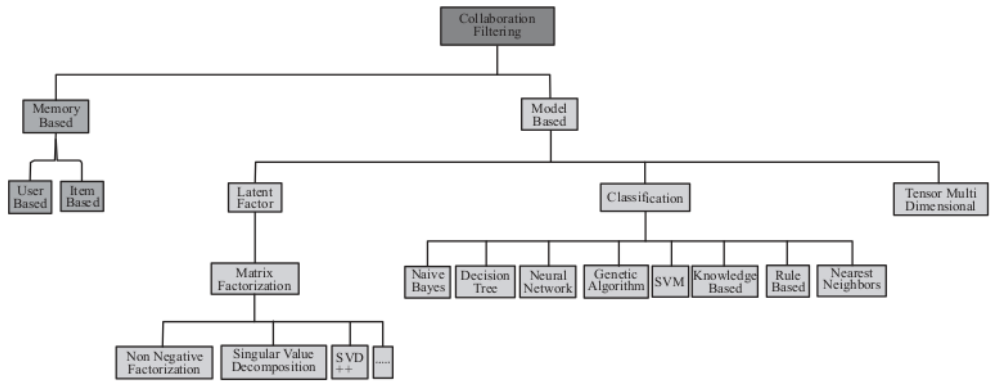

Fig. 3: Variants collaborative filtering technique

Continues rating. The rating is special on continuous scale. An example the system that implemented in Jetster joke recommendation engine take a value rating between -10 and 10.

Ordinal rating, such as "strongly agree, agree, neutral, disagree, strongly disagree. Binary rating that model choice in which the user is simply asked to decide if a certain item is good or bad. Unary rating can include that a user has observed or purchased an item or otherwise rated the item positively.

Variant of collaborative filtering: Collaborative filtering technique makes mixture of the rating that served by many users to make product prediction of users needed. The major challenge for design collaborative filtering is how to solve rating matrices are sparse. Rating is the important thing as representation of user interest for a product or item (Fig. 3).

The main idea of collaborative filtering technique is rating as representation of the feeling of users to items can be imputed because rating s analytic is having highly related between various user and item. Most of the technique of collaborative filtering are focus on influence inter user correlation and inter item correlation for predict recommendation process.

According to Fig. 4, we show the variants of collaborative filtering. The development of recommendation very strong was influenced of disadvantage of older model that was assembled. For example, born the model based collaborative filtering consider to addressing of disadvantaged as mayor problem of memory based was famous as scalability problem. Born of model based to enhance disadvantages of memory based.

Memory based collaborative filtering: In the earliest collaborative filtering technique about in the middle 90's, the most popular method is memory based also famous as

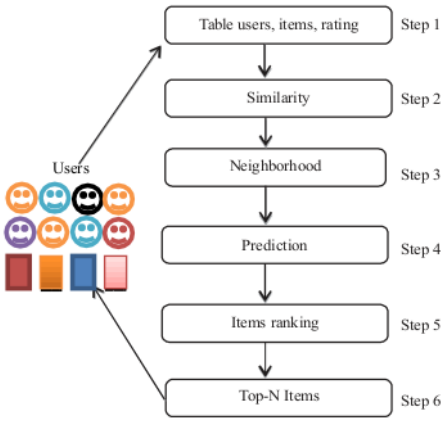

Fig. 4: Step of collaborative filtering technique

neighborhood based, this technique very powerful to predict rating. Memory based is the method use prediction based on statistical method. These is several statistical methods that used Cosine, Spearman, Pearson.

Collaborative filtering models use the collective power of user's interest given by rating items to produce recommendation. The fundamental test in planning collaborative filtering techniques is that the underline rating matrices are inadequate. There are two variants of techniques are normally utilized powered in collaborative filtering which are mentions to as memory based methods and model based methods.

Memory-based methods: Memory based methods are equaled as neighborhood based collaborative filtering algorithm. These were among oldest collaborative filtering algorithms; the basic ideas were the rating of user and item mixing are predicted on the basis of their neighborhoods. These will be explaining in two ways:

User-based collaborative filtering: User based method are an effort to get recommendation that defined in order to identify similar users to the target users for whom the value rating prediction will be calculated. In order to determine the neighborhood of the target user i, her similarity to all the other users is calculated. A similarly function need to be defined between the rating special by users. Similarity computation is very difficult because different users may have different scale.

One of sample method to measure the similarity Sim(u,v) between the rating vector of two user that mention on above is called Person correlation coefficient:

$$\mu\mu = \frac{\sum_{kal_u} r_{uk}}{|L_u|} \forall_u \uparrow \in \{1, ..., m\}$$

Next, Person correlation coefficient between the rows (users) u and v are explaining bellow:

$$Slm(u, v) = Pearson(u, v) = \frac{\sum_{ka_u u_{\pi_u}} (r_{uR} - \mu m), (r_{uR} - \mu v)}{\sqrt{\sum k_{a_{\pi_u}} \cap I_u (r_{uR} - \mu_u)^2}, \sqrt{\sum k_{a_{\pi_u}} \cap I_v (r_{vR} - \mu_v)^2}}$$

For this situation, the evaluations given by similarly invested user of an objective user A are utilized as a part of request to make the propose for A. Along these lines, the essential thought is to decide user, who are like the objective user A and prescribe evaluations for the surreptitiously appraisals of A by registering weighted midpoints of the appraisals of this recommender system.

Item-based collaborative: So as to make the rating prediction for target item B by user A, the initial step is to decide a set S of item that are most like target item B. The appraisals in item set S which are determined by an are utilized to anticipate whether the user A will like item B. Basic compute to implement neighborhood methods are used similarity formulation explained on bellow:

- If a is the active user for whom we seek recommendations, u another user and i and j two items we will denote
- I(a), I(u) and I(a&u) = I(a)∩I(u) the sets of items consumed by a, u, both a and u, respectively
- U(i), U(j) and U(i&j) = U(i)∩U(j) the set of users who consumed i, j and both i and j, respectively
- $\bar{I}(u)$ the line of matrix R for user u and $\bar{c}(i)$ its column for item i
- $\bar{I}(u)$ the average of $\bar{I}(u)$ (average rating given by u or everage number of items consumed by u) and $\bar{c}(i)$ (i's average rating or number of users who consumed I)

$$\bar{I}(u) = \frac{1}{c}\sum_{i=1}^{c} r_{ui} \quad \bar{c}(i) = \frac{1}{L}\sum_{u=1}^{L} r_{ui}$$

The similarity between users a and u can be defined through many similarity measures, for example cosine, Pearson Correlation Coefficient (PCC) (Eq. 1) or asymmetric cosine (Eq. 2) similarities (equation) (Eq. 3-5) bellow, respectively:

$$sim(a, u) = \cos\left[\vec{I}(a), \vec{I}(u)\right] \frac{\sum_{i=1}^{C} r_{ai} \times r_{ui}}{\sqrt{\sum_{i=1}^{C} (r_{ai})^2 \sum_{i=1}^{C} (r_{ai})^2}} \quad (1)$$

$$sim(a, u) = PCC\left[\vec{I}(a), \vec{i}(u)\right] \frac{\sum_{i\in i\cap I(u)} \left[r_{ai} - \bar{I}(a)\right]\left[r_{ui} - \bar{I}(u)\right]}{\sqrt{\sum_{i\in I(a)\cap I(u)} \left[r_{ai} - \bar{I}(a)\right]^2} \sqrt{\sum_{i\in I(a)\cap I(u)} \left[r_{ai} - \bar{I}(a)\right]^2}} \quad (2)$$

$$Sim(a, u) = asym - \cos_\alpha\left[\bar{I}(a), \bar{I}(u)\right] \frac{\sum_{i=1}^{C} r_{ai} \times r_{ui}}{\left[\sqrt{\sum_{i-1}^{C} r_{ai}^2}\right]^\alpha \times \left[\sqrt{\sum_{i=1}^{C} r_{ai}^2}\right]^{1-\alpha}} \quad (3)$$

$$Sim(i, j) = \cos\left[\vec{c}(i), \vec{c}(j)\right] = \frac{\sum_{u=i}^{L} r_{ui} \times r_{uj}}{\sqrt{\sum_{u=1}^{L} (r_{ui})^2} \sqrt{\sum_{u=1}^{L} (r_{ui})^2}} \quad (4)$$

$$Sim(i, j) = PCC\left[\vec{c}(i), \vec{c}(j)\right] = \frac{\sum_{u\in U(i)\cap U(i)} \left[r_{ui} - \bar{c}(i)\right]\left[r_{uj} - \bar{c}(j)\right]}{\sqrt{\sum_{u\in U(i)\cap U(j)} \left[r_{ui} - \bar{c}(i)\right]^2} \sqrt{\sum_{u\in U(i)\cap U(j)} \left[r_{uj} - \bar{c}(j)\right]^2}} \quad (5)$$

$$Sim(i, j) = asym - \cos_\alpha\left[\vec{c}(i), \vec{c}(j)\right] \frac{\sum_{u=i}^{L} r_{ui} \times r_{uj}}{\left[\sqrt{\sum_{u=1}^{L} r_{ui}^2}\right]\left[\sqrt{\sum_{u=1}^{L} r_{uj}^2}\right]^{1-\alpha}} \quad (6)$$

Benefit memory based: Memory based or very popular as neighborhood model is the eldest method in based collaborative filtering family. In fact, in real industry and researcher changes to model based for interesting. Although, there are many benefit, beside have crucial problem in this method (Table 2). The major benefits of Neighborhood method according reference[12, 3, 5] are:

Table 2: Item based and user based suitable use

CF model	Point similarity	Schema to compute
User-based	User-user similarity	When number or items larger than data of users Users don't change frequently
Item-based	Item-item similarity	When number or users larger than data of items Items don't change frequently

Stability: This method is robust to additional users and have no effect because the reason from above. In the other hand just little effect by the additional users, items and rating.

Efficiency: One of strongest memory based collaborative filtering is they have efficiency. There are no need training step those very expensive costly.

Simplicity/Easy implemented: These approaches relatively simple to implement just need one parameter.

Disadvantages memory based: Although, memories based have absolute powerful to predict recommendation, there are many problem seriously. It will be enough significant effect when the number of the users or number of products/items growth larger. According example calculation by reference[5] for example cases, when the number of users m is of the order of a few hundred million, how many time need to compute, look the formula $O(m^2 n')$ running time of a user based method will become impractical even for occasional offline computation.

For example, cases where $m = 10^3$ and $n' = 100$. Then $O(m^2 n') = O(10^{18})$ calculation will be required. If we use assumption a 10 GHz computer will require 10^3 sec, it need approximately 115.74 days. Exactly, the approach such an example cases is very not practical from scalability.

This problem become a reason many researcher try to reduce the time to compute[13, 14]. They almost break the large scale data and conduct to compute part by part data. In fact the reducing time are not enough better when comparing over model based.

Computing approach of model based collaborative filtering: Memory based having crucial problem in scalability. It is a reason why many researcher changes fully concern to model based to improve time to compute in collaborative filtering to much more efficient for the time and computer resources. However, this method raise many benefit, there is also have some disadvantages rise too. For example, major problem in this method are cold start and sparsity data. Both of problem will happen when new user and news item actually new come in table matrix of collaborative filtering. While sparsity data will happen when data rating collected by user to an item is not enough. This condition will have impact the result of recommendation will not accurate.

In model based techniques, machine learning and data mining strategies are used as a part of the setting of model based. In situations where the model is parameterized, the parameters of this model are found out inside the setting of an improvement structure. A few cases of such model based techniques incorporate decision trees, rule based models, bayesian method and neural network.

The benefit of memory based model is that they are easy to implement and the subsequent suggestions are frequently simple to clarify. Then again, memory based prediction don't work extremely well with meager collaborative filtering. Although, memory based having some advantage as E-commerce recommendation engine, they have critical problem. It is the reason why model based was born and becoming concerns many researcher interests to solve the challenge. There are several artificial intelligent, machine learning and data mining method which popular used that show in below, include:

- Decision tree collaborative filtering[15]
- Rule Based Collaborative Filtering[5]
- Naive Bayes collaborative filtering[16]
- Neural network collaborative filtering[17]
- Deep learning machine[18, 19]
- Latent factor models collaborative filtering[20, 21]
- Support Vector Machine (SVM)[3]

Main challenge of collaborative filtering: Notice that while memory based techniques produce ranked lists of items, model-based techniques predict ratings, through score which can be used also to rank recommendations. According reference[6,5]. In practice, all of collaborative filtering systems suffer from several drawbacks.

New user/item: Collaborative systems impossible giving better recommendation to new users, since, they have not rated a sufficient number of items to determine their preferences. The same problem arises for new items, which have not obtained enough ratings from users. This problem is known as the cold start recommendation problem.

Scalability: Especially for memory based systems generally have a scalability issue, because they need to calculate the similarity between all pairs of users (resp. items) to make recommendations.

Sparsity: the number of available ratings is usually extremely small compared to the total number of pairs user-item; as a result the computed similarities between users and items are not stable (adding a few new ratings can dramatically change similarities) and so, predicted ratings are not stable either.

Table 3: Example case of cold start problem

Parameters	Hanafi	Paijo	Amin	Budi	Siti
Batman	4	?	4	?	5
superman	5	?	?	5	3
Iron man	3	2	5	3	?
Hulk	?	3	?	?	1
Antman	2	?	3	?	?
Troy	?	2	4	1	4

Information: Of course memory-based and model-based techniques use very limited information, namely ratings/purchases only. They could not use content on users or items, nor social relationships if these were available.

Cold start and sparsity data causes: Cold start and sparsity data are main problem in collaborative filtering based. These causes of problem is added new users or new items also both of them[22]. Recommender system based on collaborative filtering will obtain recommendation with better accuracy when available good information about rating. Rating is feeling interested in items/products for users.

Accuracy degree of recommender system was leveraged not only availability of information but quality of information is needed absolutely. There is not enough information to build recommendation in case recommender system called sparsity data. Cold start is family problem in recommender system in these case extreme sparsity data, in the other hand when there is no rating that gave the users to an item. Impact of cold start problem is no recommendation will obtain in the system and impact of sparsity data is the result of suggestion is not accurate. As we show in Table 3, table is filled in number of rating and sign of "?", That mean user has no giving a rating from a movie. It needs an effort to predict a score that unpredicted before by users.

RESULTS AND DISCUSSION

General work frame recommender: As illustrated in Figure 5, the framework of a model Collaborative Filtering recommender system followed:

- Collect data step
- Prepare before process step
- Collaborative filtering step

For the first time process, user or customer data are collected through website ecommerce activity and placing in the server database. The next step ensures prepare before processing implemented are very crucial to make sure the data integrity and reliability. According these data, collaborative filtering algorithms are implemented to predict user interests, in the other hands actually rating and make recommendation related items in order to efficiency in the time and effort.

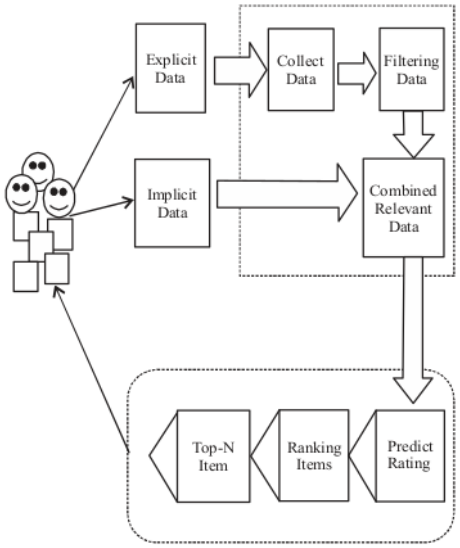

Fig. 5: General framework recommender system

Collect data: Collection data is very crucial step of the entire recommender system. The grouped data mostly breakdown into four type: demographic data, production data, user behavior and user rating.

Demographic data from users: Mostly E-commerce provider needs users to register on their systems and fill in the form for personal data before using the services. The personal data often includes name, address, occupancy, telephone number, etc. The engine on server analysis of demographic data, E-commerce server, can build the user profiles and push promotional messages to web portal when they come in a more special product.

Production data: Retailers always make a classification their product based on their functionality, location, pricing, etc. For instance, a fashions portals usually add a tag to their fashions products, to make easy the users to find what they need more enjoy. Hence, the production data are easy to access by the server.

User behavior: If users conduct browsing to a website, or they are watching to a series of movie, users are likely to be listened by the server which stores a large amount of behavior data. Such as the clicking the items which they interest, the purchasing date of a product or even the number of clicks on a website. These data are often of large volumes and need to be analyzed by especially in data mining methods.

User rating: Several sites serve rating systems and give suggestion to consumers to rate items that they have

experienced, such as movies, songs, products, news and web services. These ratings represent the preferences of a customer and receive increasing concentration from the business player. Sometimes, items have various attributes which need to be rated respectively. Accordingly, some rating systems provide users the opportunity to rate items based on multiple criteria which can greatly enrich the rating information. All of the data that mentioned in above usually play an important role in the recommender system. It will be effectively used. Although, as we explain in section I, collaborative filtering has no need information from the users (user features) and items (item feature), it concerns on feedback from the user include the implicit feedback (user behavior) and the explicit feedback (user rating).

Pre process: Data preprocess have become an important part of recommender systems. Those have a responsibility to make sure the input data of collaborative filtering completely. Preprocess is usually divided into the following 3 steps.

Data cleaning: Some of the consumers may rate the items arbitrarily, for instance, customers giving most items the highest rating. It has objective to save the time, which is likely to reduce the reliability of the rating data on the whole. Specific outlier detection algorithms can tackle these problems to some extent. For example, after choosing part of the ratings as training data and establishing a classifier model based on machine learning algorithms, the outliers can be removed with satisfying accuracy.

Generate of implicit ratings: Mainly collaborative filtering based recommender systems treat explicit user ratings as valuable data. However, lots of them do not give rate the items they have already paid, which leads to the problem of sparsity data and an extreme condition is cold start. There are many benefits in this decade because of growth of social media, mobile device and application technology. Many specific user behaviors are collected and stored in the server with significant potential information which may become the key to addressing this serious problem. For example, recommender systems receive the tremendous size of user ratings and user behaviors as training set and then applying specialized machine learning technique on it, for example, deep learning, rule based, neural network or decision tree, etc.

Data mixing: Both of the explicit and implicit rating data are combined into a matrix table, the rating matrix table, as shown in Table 3. There are still a plenty of missing values in this matrix which need to be filled in through collaborative filtering.

Metrics of collaborative filtering: General procedures of collaborative filtering include in their predicting losing values, ranking items and selecting Top-N items. Also, the rating matrix is not complete, the priority task of collaborative filtering is how to estimate these missing components using the availability data. After these process finished, a list of items are ranked following to predicted ratings and Top-N of them are chosen as the recommendation system. Once a recommender system is built, After this step finished, the next problem is how to evaluate the result of recommendation. The technique to measure metrics of recommender systems are divided into three criteria.

Measure accuracy recommendation: Accuracy metrics are utilized to assess either the prediction precision of evaluating the rating of specific user item mixed the precision of the top-k ranking predicted by a recommender system. Commonly, the ratings of a set R of sections in the rating matrix are covered up, and the precision is assessed over these covered up entries. Distinctive classes of strategies are utilized for the two cases.

Reference[12] Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) which is defined by Eq. 7 and 8:

$$MAE = \frac{\sum_{(u,t)-R_{test}} \left| R_{u,t} - R'_{u,t} \right|}{\left| R_{test} \right|} \quad (7)$$

The square root of the aforementioned quantity is followed to as the root mean squared error or RMSE:

$$RMSE = \sqrt{\frac{\sum_{(u,t)-R_{test}} \left| R_{u,t} - R'_{u,t} \right|}{\left| R_{test} \right|}} \quad (8)$$

where, $|R_{test}|$ represents the number of ratings in test set. $R_{u,t}$ is the predicted rating for user u on item I and $R_{u,t}$ is the actual actual rating in test set. A lower MAE or RMSE represents a higher predictive accuracy rating in test set. A lower MAE or RMSE represents a higher predictive accuracy.

Recall and precision: These are as the most popular metric for evaluating information retrieval system:

$$recall = \frac{\sum_u L(N,u)}{\sum_u L(u)}$$

Where:

$$precission = \frac{\sum_u L(N,u)}{UN}$$

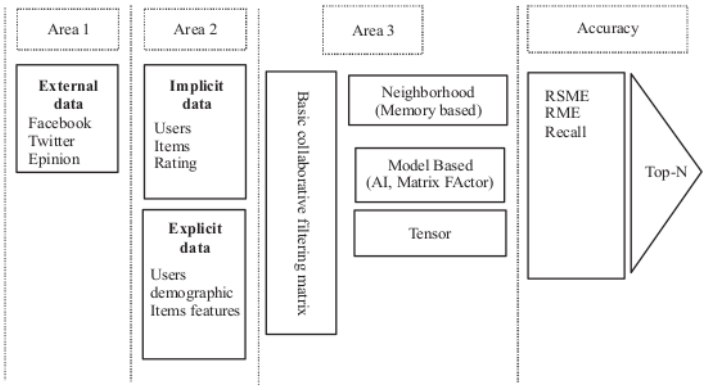

Fig. 6: Possibility strategy to improve problem

where, U is the number of users. The upper limit of precision is 1 which means all of the items in recommendation list are relevant. Sometime both of them have conflict, the formulation approach to solve them between recall and precision are:

$$F_1 = \frac{2 \times recall \times precission}{recall + precission}$$

Strategy to addressing cold start and sparsity data: In classical method of collaborative filtering in neighborhoods recommender system also popular called memory based recommender system have big problem in cold start problem. In the a decade many researcher from industrial and academic people have been trying to make new algorithm approach to obtain better results. We show figure from literature that were release began from 10 years ago. There is too much research area to get better result recommendation; in this study, we collecting literature then make summary, finally, the conclusion is there are three area methods have possible to create new method to good deal in sparsity data and cold start problem.

According to Fig. 6 from the first one area is many researchers exploit information and create algorithm from external data, for example, a social network like Facebook, Twitter, Google, etc. This method aims to exploit user data from side area.

According to strategy solution [Area 1]: The growing application of social networks used by the community at large, it is affecting the development of existing application features in the social network is increasingly diverse. What is the role of social network for recommender system? With the social network then the behavior of society can be detected, e.g., passion, life style, hobby, tendency of interest in a group, favorite places, interest in a product and so on. In this study[23] researcher proposed a method to improve sparsity data involve social network behavior concern graph based method use random walk to detect relationship between users, items together with item content, user profile and social network information. The experiment show performs well over existing method. In this study[24] researcher develops method how to measure degree of connection between one node with other as how to measure connection between node use fuzzy linguistic. This approach has objective to generate trust via similar friendship[25]. In here[26] different with some other researcher, they involve social media impact to generate tend users on social media to for product. Researcher use sentiment analysis to detect user tendency. The result experiment could improve cold start and sparsity problem. Also[27] use similarly method in[26] to detected positively feedback. In Poirier *et al*.[28] also involve social media to generate trust thought opinion classification. Other methods have proposed in[29] use user tags. In this study show, user tags have the ability to detect relationship cross domain between user and items. Results show a significantly better-comparing tag based recommender than for the people based recommender.

According to reference[30] use method through generates trust aware in social network use deep learning machine. They develop the novel deep learning matrix factorization approach to handle the trust aware recommendation problem in social networks also propose a novel method have objective to improve the quality of the initial vectors for matrix factorization by using the deep auto encoder technique. The result significantly better recommendation accuracy in particular for sparse data and cold-start users in comparison with other methods.

According to reference[6] social network have potential benefit to make recommender system working much better. There are many social contexts can be explained in social network application:

- Social network can be used as side information to improve more effective and efficient recommender system
- Social context in network centric and social trust perspective
- User interaction perspective. Interaction of user in social network creates some feedback form, for example, comment or tags. The tags can be assumed with collaboratively and classify content

Strategy solution in [Area 2]: In this study part many researchers have conducted the research to exploit between users and items information. As we know when users or items would be registered as members is E-commerce portal, we should be fill the form for detail product or users profile. This information very important to generate a part of prediction machine, there are many machine learning method who involve developing classification, probabilistic, similarity and decision. Many popular machine learning are Naive Bayes, for example, in this study[31], researcher used mapping users and items feature. How can they deal with cold start problem? In this study, researchers created a scheme to predict the rating on new users and new items by referring to users and items mapping that have been set up based on features owned by the user and the item.

Reference in this study[16], researchers could improve the scalability also performance of a previous approach to handling cold-start situations that use filterbots or surrogate users that rate to product items based only on user or item attributes. They have done create new schema in a very small number of simple filterbots to make collaborative filtering algorithms more powerful. According to reference[15] there is difference approach to deal with cold start, they use interview with the user. In this study, they introduced functional Matrix Factorization (fMF), a novel cold start suggestion technique that takes care of the issue introductory interview development inside the context of learning user and items profiles. In particular, fMF develops a decision tree for the underlying interview with every hub being an inquiry question, empowering the recommender to question a user adaptively as per her earlier reactions. All the more imperatively, we relate latent profiles for every hub of the tree as a result confining the latent profiles to be an element of conceivable responses to the inquiries questions which enables the profiles to be step by step refined through the interview procedure in view of user reactions. We build up an iterative improvement algorithm that interchanges between decision tree construct and latent profiles extraction and a regularization scenario that assesses the tree structure.

The major problem for latent factor and machine learning method are how to the technique could predict missing value the rating or on the other hand the method have ability to predict rating value even data rating are sparse or there is extremely condition no rating that called cold start problem. Many researcher use side data to generate information. User data demographics to use classification user based[32] also to optimum classification improving decision three[15], enhance content feature classification[34]. Some researcher use Bayesian to design classification, instance[34].

One of popular machine learning that have ability to compute probabilistic is neural network. Some researcher empower this method to generate probabilistic community rating[17]. In recent year some researcher enhance neural network variant[35] new technology with new generation neural network technology as we know deep learning, this method attempt to improving correlation between users feature information and content feature also involve side information[19, 36, 18, 37] the result of new method according this research show could improve over classical neural network method. Next, one of family of neural network who has evidence enough powerful to improve the problem is restricted boltzmann machine, It is like neural network method but only use single input[38].

Strategy solution in [Area 3]; Optimization latent factor: One of method to addressing cold start and sparsity data are latent factor model. This model have being recommender system because have ability to reduce missing value. This model influence well known dimensionally reduction technique to fill in the missing value. The main idea of dimensionally reduction methods is that the reduced, rotated and make complete especially representation can be powerful estimated from an incomplete data matrix. Some of most successful realization of latent factor method is based matrix factorization. In original basic matrix, matrix factorization characterizes both items and users by vectors of factors inferred from item rating pattern. High quality relationships between item and user lead recommendation.

There some of family matrix factorization implement in recommender system[39] also a variant of matrix factorization is Singular Value Decomposition (SVD). SVD is categorical unconstraint matrix factorization. In this study[40], researcher have main idea to optimization lower rank approximation could remove data noise brought by unstable user behaviors thus lead to better recommendation quality. The result of experiment show that the SVD based collaborative filtering approach not only improves the prediction accuracy but also has better performance. According[41] they was trying to improve matric factorization to deal cold start problem. This method adopted include matrix factorization by data fusion. In this study, we will show matrix factorization variants model which very popular to improve cold start, reference use the one of variant matric factorization that

named non negative matrix factorization with stochastic gradient descent. Some researcher used other variant named non negative matrix factorization with alternating least squares that was designed by[42] and semi non negative matrix factorization with missing data[20]. Especially to facing cold start problem case, there is particular of matrix factorization those very powerful when comparing with other method, in the early year[43] proposed three algorithm method to solving cold start, those are describe for imputing missing data: imputation use SVD, nearest-neighborhood and regression. After this, According reference[44], researcher proposed same method but combined with add row average as a method for the estimation of missing values in gene microarray data, the result of experiment show results show that KNN performs better than SVD and row average method can provides very fast, more accurate and more powerful approach to estimating missing data reference[45]. Next, researcher proposed three strategy addressing missing value; remove sample missing value, estimate missing value use machine learning and use weighting to combining with machine learning.

CONCLUSION

Nowadays, ecommerce company growth significantly includes quantity of users or quantity of retailer. It is a good condition for ecommerce business and will emerge many benefit for customer also ecommerce provider. Although, it have been emerge many benefits and opportunity, many technical problem have happen that called cold start problem and sparsity data. The reason for these problems is new user and new item have come in the recommender system.

Because the problem on above, there are many possible method done by many researchers involve many information resource, some algorithm method and machine learning also data mining. In some research, many researchers empower external information instance social media, tags, location and time to detect behavior users. This effort has result improving recommender system.

The second strategy to deal with cold start is create enhance algorithm to optimizing with many strategy for instance latent factor including matrix factorization, SVD, non-negative matrix factorization, semi non negative matrix factorization, alternating least square.

There are many possibilities to improving recommender system in accuracy, scalability, efficiency point of view. In latest research, the mostly popular research have emerge because trend of social media as external information and as explicit data also emerging new method for example deep learning machine. The trend will be conduct the research how to combine one method to with other. The method famous named hybrid recommendation system.

ACKNOWLEDGEMENT

We would like to acknowledge for University Teknikal Malaysia Melaka (UTeM) and Amikom University for fully support for conduct the research and to thank all my friends and colleagues for their helpful, suggestion, discussion and encouragements.

REFERENCES

01. Resnick, P., N. Lakovou, M. Sushak, P. Bergstrom and J. Riedl, 1994. Group lens: An open architecture for collaborative filtering of Netnews. Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, Oct. 22-26, Chapel Hill, North Carolina, United States, ACM Press, pp: 175-186.
02. Ansari, A., S. Essegaier and R. Kohli, 2000. Internet recommendation systems. J. Mark. Res., 37: 363-375.
03. Ricci, F., L. Rokach and B. Shapira, 2015. Recommender Systems Handbook. 2nd Edn., Springer, Berlin, Germany, ISBN:978-1-14899-7636-9, Pages: 997.
04. Linden, G., B. Smith and J. York, 2003. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Comput., 7: 76-80.
05. Handbook, R.S., 2015. Recommender Systems Handbook. 2nd Edn., Springer, Berlin, Germany,.
06. Resnick, P. and H.R. Varian, 1997. Recommender systems. Commun. ACM., 40: 56-58.
07. Kotkov, D., S. Wang and J. Veijalainen, 2016. A survey of serendipity in recommender systems. Knowl. Based Syst., 111: 180-192.
08. Chen, S., S. Owusu and L. Zhou, 2013. Social network based recommendation systems: A short survey. Proceedings of the International Conference on Social Computing (SocialCom), September 8-14, 2013, IEEE, Alexandria, Virginia, ISBN:978-0-7695-5137-1, pp: 882-885.
09. Nadine, U., H. Cao and J. Deng, 2016. Competitive recommendation algorithm for E-commerce. Proceedings of the 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), August 13-15, 2016, IEEE, Changsha, China, ISBN:978-1-5090-4094-0, pp: 1539-1542.
10. Reshma, R., G. Ambikesh and P.S. Thilagam, 2016. Alleviating data sparsity and cold start in recommender systems using social behaviour. Proceedings of the International Conference on Recent Trends in Information Technology (ICRTIT), April 8-9, 2016, IEEE, Chennai, India, ISBN:978-1-5090-0433-1, pp: 1-8.
11. Bobadilla, J., F. Ortega, A. Hernando and A. Gutierrez, 2013. Recommender systems survey. Knowl. Based Syst., 46: 109-132.

12. Yang, Z., B. Wu, K. Zheng, X. Wang and L. Lei, 2016. A survey of collaborative filtering-based recommender systems for mobile internet applications. IEEE. Access., 4: 3273-3287.
13. Ahn, H.J., 2008. A new similarity measure for collaborative filtering to alleviate the new user Cold-starting problem. Inform. Sci., 178: 37-51.
14. Toscher, A., M. Jahrer and R. Legenstein, 2008. Improved neighborhood-based algorithms for large-scale recommender systems. Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition, August 24-27, 2008, ACM, Las Vegas, Nevada, ISBN:978-1-60558-265-8, pp: 1-4.
15. Zhou, K., S.H. Yang and H. Zha, 2011. Functional matrix factorizations for cold-start recommendation. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2011, ACM, Beijing, China, ISBN:978-1-4503-0757-4, pp: 315-324.
16. Park, S.T., D. Pennock, O. Madani, N. Good and D.D. Coste, 2006. Naive filterbots for robust cold-start recommendations. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2006, ACM, New York, USA., ISBN:1-59593-339-5, pp: 699-705.
17. Devi, M.K., R.T. Samy, S.V. Kumar and P. Venkatesh, 2010. Probabilistic neural network approach to alleviate sparsity and cold start problems in collaborative recommender systems. Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), December 28-29, 2010, IEEE, Coimbatore, India, ISBN:978-1-4244-5965-0, pp: 1-4.
18. Wang, H., N. Wang and D.Y. Yeung, 2015. Collaborative deep learning for recommender systems. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 10-13, 2015, ACM, Sydney, Australia, ISBN: 978-1-4503-3664-2, pp: 1235-1244.
19. Wei, J., J. He, K. Chen, Y. Zhou and Z. Tang, 2016. Collaborative filtering and deep learning based hybrid recommendation for cold start problem. Proceedings of the IEEE 2nd and 14th International Conference on Dependable, Autonomic and Secure Computing and Pervasive Intelligence and Computing and Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/ DataCom/CyberSciTech), August 8-12, 2016, IEEE, Auckland, New Zealand, ISBN:978-1-5090-4066-7, pp: 874-877.
20. Ding, C., T. Li and M. Jordan, 2006. Convex and semi-nonnegative matrix factorizations for clustering and low-dimension representation. Lawrence Berkeley Nat. Lab. Tech. Rep., 1: 1-19.
21. Agarwal, D. and B.C. Chen, 2009. Regression-based latent factor models. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 28-July 01, 2009, ACM, Paris, France, ISBN:978-1-60558-495-9, pp: 19-28.
22. Park, D.H., H.K. Kim, I.Y. Choi and J.K. Kim, 2012. A literature review and classification of recommender systems research. Exp. Syst. Appl., 39: 10059-10072.
23. Shang, S., S.R. Kulkarni, P.W. Cuff and P. Hui, 2012. A random walk based model incorporating social information for recommendations. Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP), September 23-26, 2012, IEEE, Santander, Spain, ISBN:978-1-4673-1024-6, pp: 1-6.
24. Perez, L.G., F. Chiclana and S. Ahmadi, 2011. A social network representation for collaborative filtering recommender systems. Proceedings of the 11th International Conference on Intelligent Systems Design and Applications (ISDA), November 22-24, 2011, IEEE, Cordoba, Spain, ISBN:978-1-4577-1676-8, pp: 438-443.
25. Guo, G., J. Zhang and D. Thalmann, 2014. Merging trust in collaborative filtering to alleviate data sparsity and cold start. Knowl. Based Syst., 57: 57-68.
26. Krauss, C. and S. Arbanowski, 2014. Social preference ontologies for enriching user and item data in recommendation systems. Proceedings of the IEEE International Conference on Data Mining Workshop (ICDMW), December 14, 2014, IEEE, Shenzhen, China, ISBN: 978-1-4799-4273-2, pp: 365-372.
27. Tomeo, P., I.F. Tobias, T.D. Noia and I. Cantador, 2016. Exploiting linked open data in cold-start recommendations with positive-only feedback. Proceedings of the 4th Spanish Conference on Information Retrieval, June 14-16, 2016, ACM, Granada, Spain, ISBN:978-1-4503-4141-7, pp: 1-11.
28. Poirier, D., F. Fessant and I. Tellier, 2010. Reducing the cold-start problem in content recommendation through opinion classification. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) Vol. 1, August 31-September 3, 2010, IEEE, Toronto, Ontario, Canada, ISBN:978-1-4244-8482-9, pp: 204-207.

29. Guy, I., N. Zwerdling, I. Ronen, D. Carmel and E. Uziel, 2010. Social media recommendation based on people and tags. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, July 19-23, 2010, ACM, Geneva, Switzerland, ISBN:978-1-4503-0153-4, pp: 194-201.
30. Deng, S., L. Huang, G. Xu, X. Wu and Z. Wu, 2017. On deep learning for trust-aware recommendations in social networks. IEEE. Trans. Neural Netw. Learn. Syst., 28: 1164-1177.
31. Gantner, Z., L. Drumond, C. Freudenthaler, S. Rendle and L.S. Thieme, 2010. Learning attribute-to-feature mappings for cold-start recommendations. Proceedings of the IEEE 10th International Conference on Data Mining (ICDM), December 13-17, 2010, IEEE, Sydney, New South Wales, Australia, ISBN:978-1-4244-9131-5, pp: 176-185.
32. Safoury, L. and A. Salah, 2013. Exploiting user demographic attributes for solving cold-start problem in recommender system. Lect. Notes Software Eng., 1: 303-307.
33. Sun, D., C. Li and Z. Luo, 2011. A content-enhanced approach for cold-start problem in collaborative filtering. Proceedings of the 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), August 8-10, 2011, IEEE, Dengleng, China, ISBN:978-1-4577-0535-9, pp: 4501-4504.
34. Salakhutdinov, R. and A. Mnih, 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. Proceedings of the 25th International Conference on Machine Learning, July 05-09, 2008, ACM, Helsinki, Finland, ISBN:978-1-60558-205-4, pp: 880-887.
35. Salakhutdinov, R., A. Mnih and G. Hinton, 2007. Restricted boltzmann machines for collaborative filtering. Proceedings of the 24th International Conference on Machine Learning, June 20-24, 2007, ACM, New York, USA., ISBN:978-1-59593-793-3, pp: 791-798.
36. Li, S., J. Kawale and Y. Fu, 2015. Deep collaborative filtering via marginalized denoising auto-encoder. Proceedings of the 24th ACM International Conference on Information and Knowledge Management, October 18-23, 2015, ACM, Melbourne, Australia, ISBN:978-1-4503-3794-6, pp: 811-820.
37. Elkahky, A.M., Y. Song and X. He, 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. Proceedings of the 24th International Conference on World Wide Web (WWW'15), May 18-22, 2015, International World Wide Web Conference Committee, Geneva, Switzerland, ISBN:978-1-4503-3469-3, pp: 278-288.
38. Gunawardana, A. and C. Meek, 2008. Tied boltzmann machines for cold start recommendations. Proceedings of the ACM Conference on Recommender Systems, October 23-25, 2008, ACM, Lausanne, Switzerland, ISBN:978-1-60558-093-7, pp: 19-26.
39. Paterek, A., 2007. Improving regularized singular value decomposition for collaborative filtering. Proceedings of the International Conference on Knowledge discovery in databases (KDD-Cup07), August 12, 2007, ACM, San Jose, California, USA., pp: 39-42.
40. Ge, S. and X. Ge, 2012. An SVD-based collaborative filtering approach to alleviate cold-start problems. Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), May 29-31, 2012, IEEE, Sichuan, China, ISBN:978-1-4673-0025-4, pp: 1474-1477.
41. Ocepek, U., J. Rugelj and Z. Bosnic, 2015. Improving matrix factorization recommendations for examples in cold start. Exp. Syst. Appl., 42: 6784-6794.
42. Takacs, G., I. Pilaszy, B. Nemeth and D. Tikk, 2009. Scalable collaborative filtering approaches for large recommender systems. J. Mach. Learn. Res., 10: 623-656.
43. Hastie, T., R. Tibshirani, G. Sherlock, M. Eisen, P. Brown and D. Botstein, 1999. Imputing missing data for gene expression arrays. Technical Report, Division of Biostatistics, Stanford University.
44. Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown and T. Hastie *et al*., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics, 17: 520-525.
45. Li, Y., 2014. Sparse mach in E-learning models in bioinformatics. Ph.D Thesis, University of Windsor, Windsor, Ontario.

Addressing Sparsity Data and Cold Start Problem on Collaborative Filtering Recommender System for E-Commerce: A Review

ORIGINALITY REPORT

18%	7%	13%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

2%

★ arxiv.org

Internet Source

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Addressing Sparsity Data and Cold Start Problem on Collaborative Filtering Recommender System for E-Commerce: A Review

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13