International Journal For Technological Research In Engineering
Volume 5, Issue 3, November-2017 *ISSN (Online): 2347 - 4718*

IMPROVING PRIVACY PRESERVING COLLABORATIVE FILTERING BASED RECOMMENDATION SYSTEMS

Vidhya Raval[1], Mrs. Risha Tiwari[2]
Post Graduate Student, Professor, Dept. of Computer Engg., Hasmukh Goswami Collage of Engineering, Ahmedabad, Gujarat, India.

Abstract: Collaborative filtering (CF) systems are being widely used in E-commerce applications to provide recommendations to users regarding products that might be of interest to them. The prediction accuracy of these systems is dependent on the size and accuracy of the data provided by users. However, the lack of sufficient guidelines governing the use and distribution of user data raises concerns over individual privacy. Users often provide the minimal information that is required for accessing these E-commerce services. In this project , we propose a framework for obfuscating sensitive information in such a way that it protects individual privacy and also preserves the information content required for collaborative filtering. An experimental evaluation of the performance of different collaborative filtering systems on the obfuscated data proves that the proposed technique for privacy preservation does not impact the accuracy of the predictions
Keywords: Accuracy, privacy.

I. INTRODUCTION

Recommender Systems are software tools and techniques providing suggestions for items to be of use to a user [21]. The suggestions relate to various decision-making processes, such as what items to buy, what music to listen to, or what online news to read. “Item” is the general term used to denote what the system recommends to users. A RS normally focuses on a specific type of item (e.g., CDs, or news) and accordingly its design, its graphical user interface, and the core recommendation technique used to generate the recommendations are all customized to provide useful and effective suggestions for that specific type of item. In their simplest form, personalized recommendations are offered as ranked lists of items. In performing this ranking, RSs try to predict what the most suitable products or services are, based on the user’s preferences and constraints. In order to complete such a computational task, RSs collect from users their preferences, which are either explicitly expressed, e.g., as ratings for products, or are inferred by interpreting user actions. For instance, a RS may consider the navigation to a particular product page as an implicit sign of preference for the items shown on that page.

II. LITERATURE SURVEY & ANALISIS

Collaborative Filtering

The term 'Collaborative Filtering' (CF) was first introduced in the Tapestry system [21], for filtering electronic documents through e-mail and Usenet postings. In this system, auser explicitly requests recommendations based on reviews of a specific set of known individuals. The drawback of this system is that it requires a close-knit group of people who are aware of each other's interests. The lack of scalability of this system for larger networks ledto the development of more Automated Collaborative Filtering systems (ACF) [54]. TheGroupLens CF system [52] pioneered the research on ACF by using pseudonymous usersto provide ratings for movies and Usenet news articles. Some of the other recommendationsystems such as the e-mail based music recommendation system [66], Ringo, and the web-based movie recommendation [30], Video Recommender, also developed ACF algorithms forrecommendations. All three systems use neighborhood-based prediction algorithms such asPearson's correlation and vector similarity. These algorithms are referred to as memory-based algorithms because they use the raw data in the database to make recommendations.Model-based approaches such as Bayesian network models and cluster-based models wereproposed in [16]. These algorithms first develop cluster-based models or Bayesian network models on the database. The models are then used for making predictions for userson items that have not yet been rated by them. This makes model-based CF algorithmsfaster and less memory-intensive. Hybrid memory-model based approaches have also beendeveloped to improve accuracy of predictions [49].

Data Obfuscation Techniques

The abundance of information available online has resulted in the loss of individual privacy [18]. Several methods have been proposed and implemented for privacy preservationof sensitive data sets [32]. The term data obfuscation [7] is used as a generalization of all approaches that involve distorting the data for privacy preservation and other purposes. Oneof the more common techniques is cryptography, where sensitive data is encrypted with akey and is accessible only to an authenticated user. In several applications, it is necessary to provide different levels of precision of data, based on the type of user requesting access. Theencryption of data does not provide this capability. The usability of the data is therefore restricted only to a narrow set of users. Secure multi-party encryption techniques proposeto perform computations on data in the encrypted form [55].

Data anonymization [34] attempts to classify data into fixed or variable intervals. Theusefulness of the obfuscated data and the privacy factor are dependent on the choice of theinterval. A large interval makes the data less useful, while an interval that is too small does not provide sufficient privacy protection of the data. In [62], the author proposes a generalization and suppression approach to obtaining the required anonymity level: generalizationreplaces a value with a less specific value, while suppression does not release a value at all.This guarantees that each data item will relate to at least k other entries, even ifthe records are directly linked to external information. K-anonymization has been provento be an NP-hard problem [39]. Various algorithms, such as the k-optimal anonymization algorithm [62], the simulated annealing technique [68], and the condensation-based k-anonymization [2], have been proposed to optimize and solve the generalization/suppressionproblem, but even the most optimum algorithm that uses an approximation technique nowhas a polynomial complexity. The other drawback of the anonymization technique is theloss of information. The generalization approach categorizes quantitative information intointervals, thus reducing the granularity of the information. Furthermore, data entries thatare not possible to generalize are suppressed. This leads to a complete loss of information regarding certain fields.

III. PROPOSED WORK

Our proposed method for privacy preservation deals with the limitations of Nearest Neighbour Data Substitute ion (NeNDS) algorithm. The framework for privacy preserving collaborative filtering uses a centralized CF server that eliminates cold start problem.

Privacy Preserving Framework for proposed method

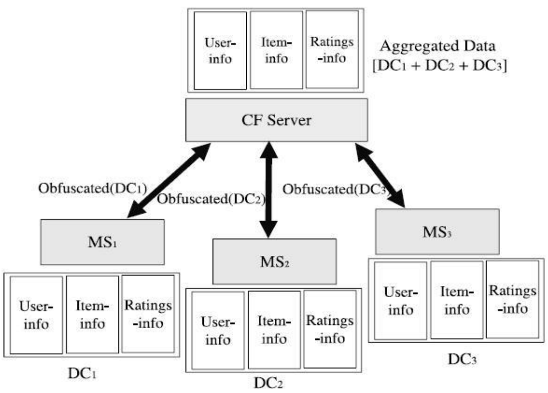

Fig. Privacy Preserving framework for Collaborative filtering [17]

The privacy framework serves as a wrapper that obfuscates the relevant fields of data before they are fed to the CF engine. A diagrammatic view of the model is shown in Figure 6.1 using an example having three meta-store fronts [MS1, MS 2, MS3] such as Amazon, C-net, Yahoo that wish to share information in a privacy preserving way. Each MSi's has three databases, a User-info database that stores demographic informationregarding its users, an Item-info database that stores information regarding the items in its inventory, and a Rating s-info database that stores information regarding the ratings provided by the users on the items purchased. The databases are obfuscated and sent to the central CF server. The CF engine combines the information from all three meta-store fronts and creates three aggregated databases as shown.

Recommendations are made for all the unrated items for each record in the ratings database. The aggregate database is then divided back into the three individual databases, which are now populated with recommendations for unrated items. The databases are then sent back to the meta-store fronts. The stores provide recommendations to their users based on the results obtained from the CF engine. Since the databases are dynamic in nature, the MSi obfuscate the updated databases periodically and send them to the CF server so that the recommendations are made on the most recent ratings of individuals. This type of framework allows different e-commerce vendors to share proprietary information about their customers without violating their privacy.

Flowchart of the Proposed System

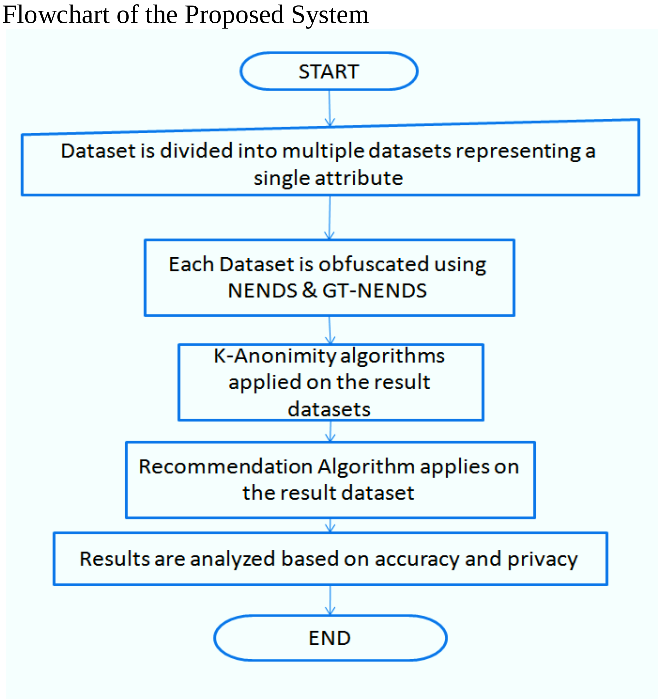

Fig.4.2. Flow chart of praposed system

Algorithm of NENDS :
NeNDS(c)
For each i ϵ [1,m] do
NHsize = [N/c+1]
(b) $\Sigma^i_{in} =$ (NH1, NH2…NH $_{NHSize}$)
For each NHj ϵ Σ^i_{in} do
Tree = Create Tree(NHj,0, NHsize)
dj = depth(Tree)
For each path$_k$ in Tree$_j$ of length d$_j$ - 1 Candidate Set = Candidate Set + (path$_k$)
NH'$_j$ = min(Candidate Set)
2. Σ^i_{out} = (NH'$_1$, NH'$_{2…}$ NH'$_{NHSize}$)
Create Tree (NH, Tree, Size)
If Tree = 0 then Tree = NH[0]
If NH = 0 then Return Tree
Children Tree = NH - Ancestors(Tree) - Identical(Parent, NH)
Child(Tree) = Sort(Children Tree)
Tree = Child(Tree)

IV. IMPLEMENTATION

Data set Details
Dataset: MovieLens(100K, 1M)

- The dataset contains 100,000 ratings from 1000 users on 1700 movies.
- The dataset contains 1 million ratings from 6000 users on 4000 movies.

Accuracy of the algorithm will be checked against both the datasets.

Description of Dataset:
The dataset with 100k ratings (4.8 MB):
This data set consists of:

- 100,000 ratings (1-5) from 943 users on 1682 movies.
- Each user has rated at least 20 movies.
- Simple demographic info for the users (age, gender, occupation, zip)

The dataset having 1m ratings (5.8 MB):
This dataset consists of:

- There are 1000054 ratings, 95580 tags applied to 10681 movies.
- There are 71567 users which have rated the movies.
- All users have rated at least 20 movies.

Work Done So Far

- Collected and Analyzed Movielens Datasets to be exercised
- Converted the dataset files into .csv files to be used in Mahout.
- Installed Eclipse Kepler (Service Release 2).
- Configured Apache Mahout Distribution 5.0 with Eclipse.
- Practiced some sample run in the environment to understand the behavior inside the environment.

Comparison for Various Datasets

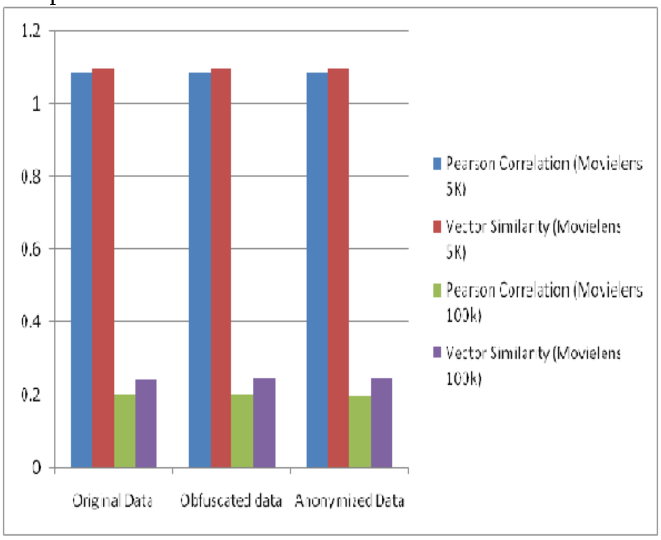

Fig 5.4.Comparision for various dataset

MAE for Anonymized Dataset

	Prediction Accuracy			
Collaborative Filtering Algorithm	**Original Data**	**Obfuscated data**	**Anonymized Data**	**Error %**
Pearson Correlation (Movielens 5K)	**1.083**	**1.084**	**1.085**	**0.2**
Vector Similarity (Movielens 5K)	**1.095**	**1.096**	**1.097**	**0.2**
Pearson Correlation (Movielens 100k)	**0.198**	**0.198**	**0.197**	**0.1**
Vector Similarity (Movielens 100k)	**0.241**	**0.242**	**0.243**	**0.1**

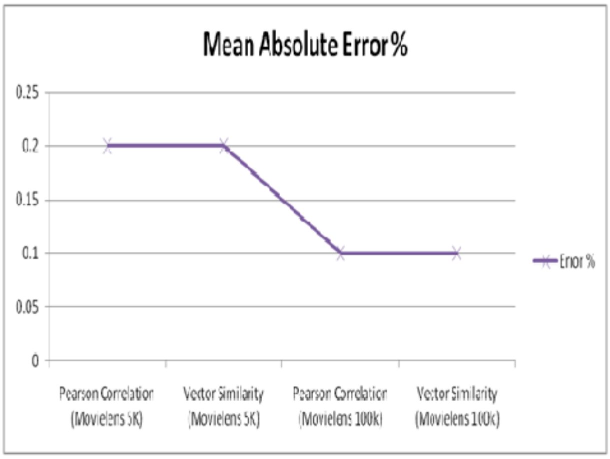

Fig 5.5. Mean absolute error

Comparison of Execution Time

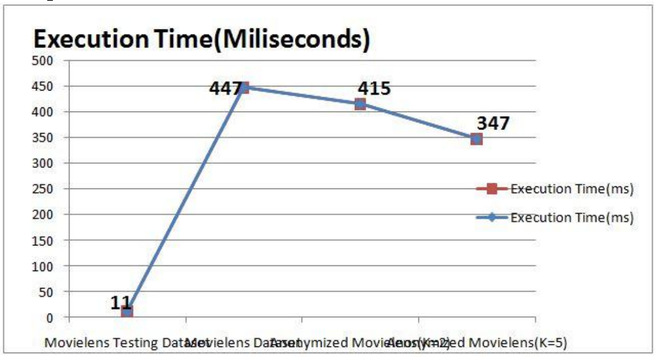

Fig .5.6.Comparision of execution time

Comparison of Memory Usage

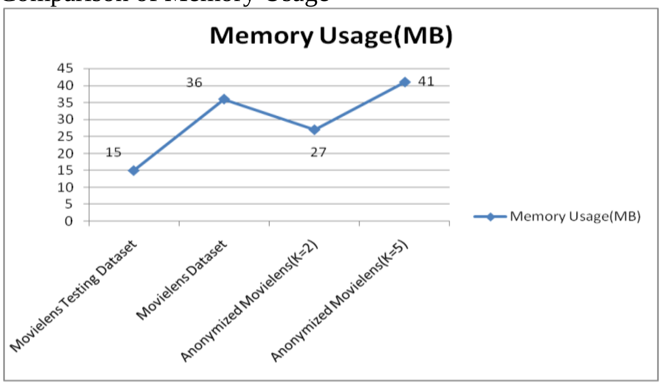

Fig .5.7 .Comparision of memory usage

V. CONCLUSION

This thesis proposes privacy preserving framework for collaborative filtering applications. The problem of privacy is still not a well-understood one. While there is a definite need for privacy, there is no clear-cut answer to the question of what information is considered private and when a database is considered to be breached. In general, if any information about an individual revealed from a database can be obtained in any other way without access to the database, the information is no considered as private. Gaining access to such information is not considered as a privacy breach.

NeNDS-based transformations obfuscate individual records by permuting each dataset individually. Anyquery made to the database is guaranteed to reveal an answer that is close to the truth but different from the exact truth.

AnonymizedNeNDS takes privacy a step further by generalization and suppression of the data to a state where the values in the database are clearly different from the originalvalues. The inter-relationship among the data items are preserved, which makes this approach an excellent candidate for data mining applications. The privacy preserving framework can be used to share information amongmultiple meta-store fronts for information for mutual gain. New sellers suffer an initialsetback, referred to as cold-start, because of the lack of a data pool to provide recommendations to its users. The cold start problem can be averted by the presence of a shared CFengine. The experimental results indicate that the accuracy of CF engines remains nearlythe same in spite of the preliminary data obfuscation process. Although the rank scoringmetric indicated that the utility of the ranking order is decreased by data obfuscation, theerror is only about 5% on average, which is an acceptable trade-off, given the benefits of arobust privacy-preservation mechanism.

VI. FUTURE WORK

Some interesting problems for future work are listed below.
Anonymized NeNDS-based Data Obfuscation can be performed only on static databases. For dynamic databases, or databases that undergo constant changes, the algorithm can be applied to the database periodically. However, this could be time-intensive for some applications because the entire database has to be obfuscated and anonymized each time. An interesting problem is to study ways in which data obfuscation and anonymization can be applied only to the parts of the database that have been modified without losing clustering information of the data. The collaborative filtering framework proposed here assumes that the users of the E-commerce sites are valid users and are not malicious. The framework does not include mechanisms to avoid shilling or targeted attacks on the CF system. Methods such as building a web of trust, and trust-aware CF have been proposed to counter such targeted attacks. An evaluation of the performance of the NeNDS-based CF framework that incorporates trust-based techniques is an interesting future work.

REFERENCE

Data Mining: Concepts and Techniques, Second Edition Jiawei Han and MichelineKamberUniversity of Illinois at Urbana-Champaign

[1] Recommender Systems Handbook Francesco Ricci LiorRokach · BrachaShapira · Paul B. Kantor Springer

[2] Privacy-Preserving Data Mining Models and Algorithms Charu C. Aggarwal, IBM T.J. Watson Research Center, USAPhilip S. Yu, University of Illinois at Chicago, USA

WEB

[3] www.wikipedia.org

PAPERS

[4] ShlomoBerkovsky, YanivEytani, TsviKuflik, Francesco Ricci. Privacy-Enhanced Collaborative Filtering"

[5] Sheng Zhang, James Ford, FilliaMakedon. Department of Computer Science, Dartmouth College. "A Privacy-preserving Collaborative Filtering Scheme with Two-way Communication".

[6] HuseyinPolat and Wenliang Du. "Privacy-Preserving Collaborative Filtering". International Journal of Electronic Commerce / Summer 2005, Vol. 9, No. 4, pp. 9–35.

[7] Husain Polat,Wenliang Du. "SVD based Collaborative Filtering with Privacy".SAC'05, March 1317, 2005, Santa Fe, New Mexico, USA.

[8] ShlomoBerkvosky,YanivEytani,TsviKuflik,Francesco Ricci. "Enhancing Privacy and Preserving Accuracy of a Distributed Collaborative Filtering".RecSys'07, October 19–20, 2007, Minneapolis, Minnesota, USA.

[9] Joseph A. Calandrino, Ann Kilzer, Arvind Narayanan, Edward W. Felten and VitalyShmatikov. "You Might Also Like: Privacy Risks of Collaborative Filtering".

[10] Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan. "Collaborative Filtering Recommender Systems", Foundations and Trendsin Human–Computer Interaction. Vol. 4, No. 2 (2010)

[11] AnirbanBasu, Hiroaki Kikuchi, and JaideepVaidya, "Privacy-preserving weighted Slope One predictor for Item-based Collaborative Filtering".

[12] HuseyinPolat, Wenliang Du "Privacy-preserving top-N recommendation on horizontally partitioned data". 2005

[13] AnirbanBasu, JaideepVaidya, Hiroaki Kikuchi. "Efficient Privacy-Preserving Collaborative Filtering Based on the Weighted Slope One Predictor" ,JISIS, volume: 1, number: 4, pp. 26-46 2011

[14] Richard Cissée, SahinAlbayrak. "An Agent-Based Approach for Privacy-Preserving Recommender Systems". AAMAS'07 May 14–18 2007, Hono lulu, Hawai'i, USA.

