ENS 491-492 – Graduation Project

Final Report

Project Title: Differential Privacy Framework for Movie Recommendation System

Group Members: Mert Kosan, M. Mucahid Benlioglu

Supervisor(s): Yücel Saygın

Date: 21.05.2017

Sabancı Üniversitesi

Contents

Executive Summary 3

Problem Statement and Objectives 4

Project Result and Achievements 6

Benchmarking and Testing 6

Differentially Private Queries 7

Is Project Successfully Completed? 9

Realistic Constraints of Project 10

Impact 11

Project Management 12

Conclusion and Future Work 13

References 14

Executive Summary

Establishing privacy without losing utility of data is very hard problem in computer science. You should utilize the most effective algorithms and technics that is suitable your own systems. In this project, we tried to establish differentially private movie recommendation system. Project had two main sides. First one is building a tool that recommends a movie to the user by using item-item collaborative filtering algorithm. Second one is making this recommendation system differentially private. Both was using k-nearest neighbor algorithm to choose movie to recommend. The most important realization of project was making a good differentially private recommendation system is very hard problem to solve. We bound the sensitivity of the queries based on specific database to make system differentially private. We made the queries differentially private step by step to see the better results. Also, our results of benchmarking and testing of our findings were very impressive which proves reliable tool is created in this project.

Problem Statement and Objectives

With today’s advancing technology we have started to gather enormous amount of data in all areas from important and sensitive data such as medical reports of patients to entertainment such as games and movies. As the data grew, need for statistical analysis and providing meaningful results from the data is also occurred. But most of the time, because of the sensitivity of the data these statistical results can give away the specific and sensitive data of a specific person, such as medical test results of person, which is against the privacy constraints. Our motivation is that given a database that contains sensitive information and given an adversary that has considerable amount of information regarding to the database, the data of a person remains private and cannot be accessed by the queries provided by the database.

Current approaches to privacy either prevents researchers to use the data and provide useful applications to users or does not guarantee the privacy. For instance, some data is given out anonymously, but if an adversary knows some information about the database and the specific person that he wanted to learn about, he/she can easily bypass the privacy that anonymity provides. To be clearer we will provide the following scenario about a medical database:

> Suppose that a medical institute recently started to apply differential privacy in their databases, an adversary knows that and has the anonymous data before they applied the differential privacy. An important CEO, John, goes to this institute to get tested for a disease, along with 2 others this week. Our adversary also knows about this and pays off the other 2 people to learn their results. As a result, our adversary plans to learn John’s result by getting the count of total positive results from previous anonymous database, combine it with the result that he took from 2 other guys that are tested, and ask our new differentially private database a simple query: *“How many people are diagnosed with this disease so far?”.*

Now the question in this story seems very innocent at first, but given the knowledge our adversary possesses, he/she can retrieve the information whether John has the given disease or not, which is a clear violation of the privacy and a proof that anonym data does not guarantee this privacy.

But, our differentially private database aims to give such an answer to these seemingly innocent queries, so that adversaries cannot retrieve the sensitive information even if they have enormous amount of knowledge about the database as described in the scenario, and while doing so it still does not disturb the result very much, which allows researchers with non-violent intentions to continue their research and reach to meaningful results.

In the light of this motivation, we have decided to apply differentially private queries to movie databases, to give recommendations to users, while keeping their ratings private. To reach this goal, we have divided our tasks into smaller tasks such as a query that counts how many user gave ratings to a given movie.

Following section will give a more detailed insight about our achievements and progress towards our ultimate goal.

Project Result and Achievements

As we mentioned in the progress report 2, we had two main tasks to do until final to build a tool. First one was applying tests to our algorithm to check their correctness, second was applying differential privacy to queries of our model, recommendation system. Plan was not changed and we established our remaining tasks to complete our project. Below tasks which are done and how are done will be introduced.

Benchmarking and Testing

To remind reader of this report, we found a similarity between movies by using adjusted cosine similarity given knowledge of rating of user to movie. Please refer to progress report 2 to recall how adjusted cosine similarity work on item-based collaborative filtering. However, to continue our project more effectively, we had made test on adjusted cosine similarity algorithm.

At the beginning of the process, we realized that our data is very large to take effective testing, so clean our data to control the test result and itself effectively. However, cleaning operation didn’t happen randomly, we choose 100 most rated movies to create mini database (it is not mini, testing complexity is bound number movie that database has). With this minimization, we increased our performance of time complexity with sufficient number of data.

After a new database is created, we have divided our data into two. One is for training, one is for testing data. In our test model, we used %80 training data, %20 test data. With testing data, we tried to predict rating of user to movie. For each movie, we found a k-nearest movie and prediction value is weighted average of this k movie. As we also know real rating, we calculate difference of prediction and real ratings. This difference corresponds error adjusted cosine similarity algorithm. For each test case, we applied same procedure and at the end, we average whole error to calculate mean absolute error (MAE) of algorithm (MAE is mentioned in progress report 2). In our database, user can rate movie 1 to 5 (worst rating is 1, best rating is 5). We found mean error 0.66 which indicates adjusted cosine similarity is algorithm we can use in movie recommendation system.

Differentially Private Queries

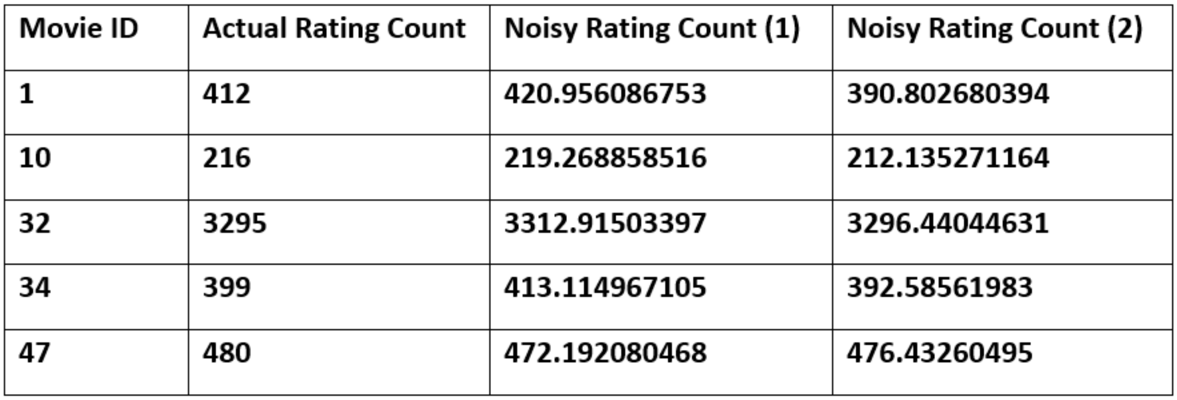
In differential privacy concept, there are two main factors which are sensitivity of query and epsilon (ε) which indicates how private the query is. So, these two factor should be picked carefully while creating a query which is differentially private. In our tool, we choose them by making tests upon them. Sensitivity of query is more obvious than ε, because it represents how response of query will change in the worst case when one user data does not exist in our case. However, picking ε depends on some tests upon data. At first, we picked bigger ε and checked whether query is differentially private or not, if not we decrease ε, and test again. After we found sensitivity and ε values, we utilize Laplace mechanism to randomize noise which will add to original data. Laplace mechanism can generate positive noise as well as negative noise. So, noisy data can be lower or higher than original data. The important feature of differential privacy is every run of queries, results of query is changing, you cannot see same result. Below, we introduced one of the example differentially private query that we build. You can find details on the example.

- *Query of number of rating given movie*

  In this query, we firstly calculate the sensitivity. Sensitivity of this query is 1, it is because in the model we are considering if one user won’t give us her/his ratings, how this issue affects our query. If we delete one user from database, given movie is affected at most 1 count as users can only rate once per movie. Then this interpretation tells us sensitivity of this query is 1.

  Secondly, we tried to answer second important question. How can we choose ε? We had two constraints, which are ε should satisfy differential privacy and it should not ruin utility of data. We started from 0.9 ε value and decreased by 0.1 for each case. We tried to find best option for ε value. At the end, we decided to choose 0.1 for ε. It ensures that two constraints are satisfied by 0.1 ε value.

  Below figure represents examples of our differentially private query. First column is movie id which is identical for each movie. Second row is real rating count of movie. Third is noisy (differentially private) rating count of a movie for query is sent firstly. And column fourth is noisy rating count of a movie for queries sent secondly for the same movie. As we are expecting, it is good model. Also, first and second result of queries are also different which is accurate and must for differential privacy.



Movie ID	Actual Rating Count	Noisy Rating Count (1)	Noisy Rating Count (2)
1	412	420.956086753	390.802680394
10	216	219.268858516	212.135271164
32	3295	3312.91503397	3296.44044631
34	399	413.114967105	392.58561983
47	480	472.192080468	476.43260495

Is Project Successfully Completed?

In this project, there were many tasks to do, we tried to finish every task professionally and until differential privacy part, everything is done appropriately. However, in differential privacy part, as we have many difficulties to establish very operative system, we changed our plan for differential privacy. This change gives us and our model better and more robust solution. With innovative approach to make database privacy which is making differentially private step by step, this made our approach to differential privacy more generic. As applicability of differential privacy is very hard, we made some queries of database differentially private.

Realistic Constraints of Project

Primary focus of the project is that, providing differentially private query responses to databases, while making sure that privacy constraints does not disrupt the utility a usability of the statistical data. Therefore, our greatest constraints were to ensure the privacy and while doing so keep the utilization high enough so that the data still has a meaning in queries.

Our project and works towards satisfying these constraints were mainly in the research level, we have calculated the sensitivity cofactors to keep the data private and try different approaches to queries to keep the sensitivity factor low enough so that the random noise we add to the data does not disturb the utilization of the data so much that it becomes no longer meaningful.

Since our project is at research level we have no other “realistic constraints” such as economic, environmental, health-safety or manufacturability etc.

Impact

When you analyze current applications, differential privacy is not ingrained and widely-used system for applications. It isn’t because it is very ineffective system, it is because it is hard system to build. Apple has announced (as we mentioned in Proposal) that they are using differential privacy in their data preserving models with IOS 10. However, in the future we expect that differential privacy will be used widely in applications especially in recommendation system. We didn’t build a complete system to commercial yet as we are still researching about how can we deal with sensitivity and epsilon values and how we can build better system that we built. However, because differential privacy establishing secure system for user, besides security, better systems with more data; in future applications would use differential privacy in their systems. We hope that differential privacy will be a topic causing célèbre.

Project Management

At the beginning of project, dominant topic was differential privacy on text analytics. However, as days are passing, we realized that differential privacy of recommendation system should be dominant topic. Text analytic remains second plan of this project. It is because, there is no accepted solution on text analytics on internet and constituting of applied tool is troublesome.

Second important change of project is on differential privacy of movie recommendation system. When we make database differentially private, we are making private users' ratings. However, this makes sensitivity is tremendously high. So, we decided to bound sensitivity of queries for each or specific database so that it can give us better results.

However, these changes don't affect our progress and we don't lose any time in terms of project management as we build our plan effective and adaptable for other plans. Also, we changed programming languages in the implementation process. However, this affects our progress positively. New programming languages (Python) was more suitable for our plans. Actually, we lost a little bit time with Java, but in that process, we learned many things for example, differential privacy on histogram publishing systems.

With our supervisor, we tried to establish every task at the right time. Also as we are doing project 2 people, we shared tasks equally and effectively with each other to make better project, to get efficient and effective results.

Conclusion and Future Work

We started our research by creating a query to calculate the similarity between movies and by using this similarity result giving the user meaningful recommendations. As we succeeded this task we have directly tried to reach our real goal, which is creating recommendations with differentially private queries. But as we progressed, we have observed that the obvious sensitivity factor disturbs the utilization of the data more than it should, which prevents us from making meaningful recommendations. Therefore, to find a way around, while still ensuring the privacy, we have decided to divide the goal into small tasks as explained in detail in project results/achievements section.

What is left to finish is putting together these small tasks and finding the optimal epsilon (ε) and the optimal sensitivity factor. In addition to that, since the correctness of the data is corrupted with random noise of the differential privacy, in future the recommendation algorithm might be reinforced with probabilistic reasoning and static similarity to improve the accuracy of the recommendation even under some uncertain conditions.

References

Dwork, C., & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science: Vol. 9: No. 3–4*, 211-407.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In S. Halevi, & T. Rabin, *Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, vol 3876.* Springer, Berlin, Heidelberg.

Elasticsearch. (n.d.). Retrieved from http://db-engines.com/en/ranking/search+engine

Ghosh, A., Roughgarden, T., & Sundararajan, M. (2012). Universally Utility-maximizing Privacy Mechanisms. *SIAM Journal on Computing, 41*(6), 1673-1693.

Greenberg, A. (2017, March 23). *Apple's 'Differential Privacy' Is About Collecting Your Data*. Retrieved from https://www.wired.com/2016/06/apples-differential-privacy-collecting-data

Gursoy, M. E., Inan, A., Nergiz, M. E., & Saygin, Y. (2016, September 9). Privacy-Preserving Learning Analytics: Challenges and Techniques. *IEEE Transactions on Learning Technologies*, pp. 68-81.

Inan, A., Gursoy, M. E., Esmerdag, E., & Saygin, Y. (2016). Graph-based modelling of query sets for differential privacy. *Proceedings of the 28th International Conference on Scientific and Statistical Database Management* (p. 3). ACM.

Item-item collaborative filtering. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Item-item_collaborative_filtering

Kosan, M., & Benlioglu, M. M. (2016). *Differential Privacy Framework for Text Analytics (Proposal).*

Kosan, M., & Benlioglu, M. M. (2017). *Differential Privacy Framework for Movie Recommendation System (Progress Report II).*

Kosan, M., & Benlioglu, M. M. (2017). *Differential Privacy Framework for Text Analytics (Progress Report-I).*

MovieLens. (n.d.). Retrieved from https://grouplens.org/datasets/movielens/

Norton, R. M. (1984). The double exponential distribution: Using calculus to find a maximum likelihood estimator. *The American Satisfaction, 38*(2), 135-136.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th international conference on World Wide Web* (pp. 285-296). ACM.

Simon, P., & Garfunkel, A. (1964). *The Sound of Silence*. Retrieved from https://youtu.be/4zLfCnGVeL4

Xu, J., Zhang, Z., Xiao, X., Yang, Y., & Yu, G. (2013). Differentially private histogram publication. *The VLDB Journal*, 797-822.