Review on Techniques of Collaborative Tagging

Ms. Benazeer S. Inamdar[1], Mrs. Gyankamal J. Chhajed[2]

[1]Student, M. E. Computer Engineering, VPCOE Baramati, Savitribai Phule Pune University, India

benazeer.inamdar@gmail.com

[2]Assistant Professor, Computer Engineering Department, VPCOE Baramati, Savitribai Phule Pune University, India

gjchhajed@gmail.com

Abstract- Collaborative tagging is one of the most diffused and prominent services available on the web. It has emerged as one of the best ways of associating metadata (tag) with web resources like image, bookmark, post etc. With the increase in the kinds of web objects becoming available, collaborative tagging of such resources is also developing along new dimensions. In this survey paper, we have reviewed and analyzed different privacy enhanced technologies (PET) of collaborative tagging like tag suppression, tag perturbation, tag recommendation and tag prediction.

Keywords— Social bookmarking, Collaborative tagging, Tag prediction, Tag recommendation, Data perturbation, Granularity, Filtering.

INTRODUCTION

Collaborative tagging became popular with the launch of sites like Delicious and Flickr. Since then, different social systems have been built that support tagging of a variety of resources. Given a particular web object or resource, tagging is a process where a user assigns a tag to an object. On Delicious, a user can assign tags to a particular bookmarked URL. On Flickr, users can tag photos uploaded by them or by others. Whereas Delicious allows each user to have her personal set of tags per URL, Flickr has a single set of tags for any photo. On blogging sites like Blogger, Wordpress, Livejournal, blog authors can add tags to their posts.

The main purpose of collaborative tagging is to classify resources based on user feedback, expressed in the form of tags. It is used to annotate any kind of online and offline resources, such as Web pages, images, videos, movies, music, and even blog posts. Nowadays collaborative tagging is mainly used to support tag-based resource discovery and browsing. Consequently, collaborative tagging would require the enforcement of mechanisms that enable users to protect their privacy by allowing them to hide certain user generated contents, without making them useless for the purposes they have been provided in a given online service. This means that privacy preserving mechanisms must not negatively affect the accuracy and effectiveness of the service, e.g., tag-based browsing, filtering, or personalization. Tag suppression is the privacy-enhancing technology (PET) is used to protect end user privacy. Tag suppression is a technique that has the purpose of preventing privacy attackers from profiling user's interests on the basis of the tags they specify. It can affect the effectiveness of policy based collaborative tagging systems.

TECHNIQUES OF PRIVACY PRESERVATION

1. Collaborative Filtering Using Data Perturbation

Collaborative filtering techniques are becoming increasingly popular in E-commerce recommender systems as data filtration is most demanding way to reduce cost of searching in E-commerce application. Such techniques suggest items to users employing similar users' preference data. People uses recommender systems to deal with information overload. Although collaborative filtering systems are widely used by E-commerce sites, they fail to preserve users' privacy as data is exposed to filter engine in unencrypted form. Since many users might decide to give wrong information because of privacy concerns, collecting high quality data from users is very tough task. Collaborative filtering systems using these data might produce inaccurate recommendations.

1.1 Randomized Perturbation Techniques

In this paper, H. Polat and W. Du propose a randomized perturbation technique to protect individual privacy while still producing accurate recommendations results. Although the randomized perturbation techniques attach randomness to the original data to prevent the data collector from learning the private user data, the method can still provide recommendations with decent accuracy.

These approaches basically suggest perturbing the information provided by users. In this, users add random values to their ratings and then submit these perturbed ratings to the recommender system. After receiving these ratings, the system perform an algorithm and sends the users some information that allows them to compute the prediction [1].

Advantage:

This approach makes it possible for servers to collect private data from users for collaborative filtering purposes without compromising users' privacy requirements. This solution can achieve nearly accurate prediction compared to the prediction based on the original data.

Limitations:

The accuracy of this scheme can be provide most accurate result if more aggregate information is disclosed along with the concealed data, especially those aggregate information whose disclosure does not compromise much of users' privacy. This kind of information includes distribution, mean, standard deviation, true data in a permuted manner, etc.

1.2 SVD(Singular Value Decomposition) Based Collaborative Filtering

In this paper, H. Polat and W. Du proposed SVD-based collaborative filtering technique to preserve privacy. The method used is a randomized perturbation-based system to protect users' privacy while still providing recommendations with decent accuracy. In this, the same perturbative technique is applied to collaborative filtering algorithms based on singular-value decomposition [2].

Limitations:

Even though a user disguises all his/her ratings, it is evident that the items themselves may uncover sensitive information. The simple fact of showing interest in a certain item may be more revealing than the ratings assigned to that item.

1.3 Random Data Perturbation Techniques

In this paper, H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar proposed a technique which is used to preserve data privacy by adding random noise, while making sure that the distort noise still preserves the “signal” from the data so that the patterns can still be accurately estimated. Randomization-based Techniques are used to generate random matrices [3] .

The following information could lead to reveal of private information from the perturbed data.

a) Attribute Correlation: Real time data has strong correlated attributes, and this correlation can be used to filter off additive white noise.
b) Known Sample: The attacker sometimes has specific background knowledge about the data or a collection of independent samples which may overlap with the original data. This may not happen every time in collection of independent samples from background knowledge.
c) Known Inputs/Outputs: There is large probability that the attacker knows a small set of private data and their perturbed counterparts. This equivalency can help the attacker to estimate other private data.
d) Data Mining Results: The particular pattern discovered by data mining also provides a certain level of knowledge which can be used to guess the private data to a higher level of accuracy.
e) Sample Dependency: Most of the attacks assume the data as independent samples from some unknown distribution. This consideration may not hold true for all real applications.

Limitations:

Some of the challenges that these techniques face in preserving the data privacy. It showed that under certain conditions it is relatively easy to breach the privacy protection offered by the random perturbation based techniques.

1.4 Deriving Private Information from Randomized Data

In this paper, Z. Huang, W. Du, and B. Chen are proposed a method to correlation which affects the privacy of a data set disguised using the random perturbation scheme. There are two methods to reconstruct original data from a disguised data set. One scheme is based on PCA (Principal Component Analysis), and the other scheme is based on the Bayes estimate. Results have shown that both the PCA-based schemes and the Bayes estimate (BE) based scheme can reconstruct more accurate data when the correlation of data increases [4].

The BE based scheme is always better than the PCA-based schemes. To defeat the data reconstruction methods that exploit the data correlation, authors proposed a modified random perturbation, in which the random noises are correlated. The experiments show that the more the correlation of noises resembles that of the original data, the superior privacy preservation can be achieved.

2. Tag Prediction

Tag prediction concerns the possibility of identifying the most probable tags to be associated with a non tagged resource. Tags are predicted based on resources content and its similarity with already tagged resources.

2.1 Social Tag Prediction

In this paper, P. Heymann, D. Ramage, and H. Garcia-Molina proposed a tag prediction technique. Tag is predicted based on anchor text, page text, surrounding hosts, and other tags applied to the URL. An entropy-based metric which captures the generality of a particular tag and informs an analysis of wellness of the tag which can be predicted. Tag-based association rules can produce very high-precision predictions as well as giving deeper understanding into the relationships between tags [5].

Limitations:

The predictability of a tag when the classifiers are given balanced training data is negatively correlated with its occurrence rate and with its entropy. More popular tags are harder to predict and higher entropy tags are harder to predict. When considering tags in their natural (skewed) distributions, data sparsity issues lead to dominate, so each tag improves classifier performance. This method perform poor in case of popular tags and distribution becomes poor with overall performance

2.2 Granularity of User Modeling

In this paper, Frias-Martinez, M. Cebria´n, and A. Jaimes proposed a tag prediction technique based on granularity. One of the characteristics of tag prediction mechanisms is that, all user models are constructed with the same granularity. In order to increase tag prediction accuracy, the granularity of each user model has to be adapted to the level of usage of each particular user. In this, canonical, stereotypical and individual are the three granularity levels which are used to improve accuracy. Prediction accuracy improves if the level of granularity matches the level of participation of the user in the community [6].

Limitations:

This approach doesn't investigate the following two areas: (1) how to identify the scope of information used in the construction of the models (i.e., size and shape of clusters in the stereotypical case), and (2) how and when user models evolve from one granularity to the next.

3. Recommendation Approach

In this paper, G. Adomavicius and A. Tuzhilin proposed a tag recommendation approach. It suggests to users the tags to be used to describe resources they are bookmarking. It is enforced by computing tag based user profiles and by suggesting tags specified on a given resource by users having similar characteristics/interest [7].

3.1 Content-based Recommendation Approach

Content-based recommendation systems try to recommend items similar to those a given user has preferred in the past. The basic process performed by a content-based recommender consists in matching up the attributes of a user profile in which preferences and interests are stored, with the attributes of a content object (item), in order to recommend to the user new interesting items.

a)Heuristic-based

In this item profile (keyword format) is searched by using TF-IDF. User profile (weights of keywords for each user) and cosine similarity are calculated.

b) Model-based

In this Bayesian classifiers and Probability measures are used in content-based approach. Some of the model-based approaches provide rigorous rating estimation methods utilizing various statistical and machine learning techniques.

Limitations:
1. Limited Content Analysis (insufficient set of features).
2. Overspecialization (recommend too similar items).
3. New User Problem (not enough information to build user profile).

3.2 Collaborative based

In this, the user is recommended items that people with similar tastes and preferences liked in the past. Collaborative recommender systems (or collaborative filtering systems) try to predict the utility of items for a particular user based on the items previously rated by other users. The utility *u*(*c*, *s*) of item *s* for user *c* is calculated based on the utilities u (cj, s) assigned to item *s* by those users *cj*∈*C* who are “similar” to user *c*.

a) Heuristic-based

In this, correlation coefficient and cosine-based Similarity measurements are used. Heuristic based methods are also known as memory based methods. Memory-based algorithms essentially are heuristics that make rating predictions based on the entire collection of previously rated items by the users.

b) Model-based

In this, Cluster models and Bayesian networks are used. Some of the model-based approaches provide various rating estimation methods utilizing various statistical and machine learning techniques.

Limitations:

1. New User Problem (not enough information to build user profile).
2. New Item Problem (too few have rated on new items).
3. Sparsity (too few pairs of users have sufficient both-rated items to form a similar group among them).

3.3 Hybrid based

Hybrid based methods combine collaborative and content-based methods. It predicts the absolute values of ratings that individual users would give to the yet unseen items.

a) Heuristic-based

i) Adding Content-based Characteristics to Collaborative Models. In this, content-based profile is used to calculate similarity between users and a user can be recommended an item not only when this item is rated highly by users with similar profiles.
ii) Adding Collaborative Characteristics to Content-based Models and developing a Single Unifying Recommendation Model.

b) Model-based

In this, combine content-based and collaborative components by:
i) Incorporating one component as a part of the model for the other.
ii) Building one unifying model.

CONCLUSION

In this paper we have reviewed and analyzed different methods to preserve privacy of collaborative tagging. We have reviewed different techniques like Tag perturbation, tag prediction and tag recommendation. We can conclude that each approach has its own significance and importance in preserving privacy of end user. Each tag based approach uses different algorithm and evaluation technique for preserving privacy.

REFERENCES:

1) H. Polat and W. Du, "Privacy-Preserving Collaborative Filtering Using Randomized Perturbation Techniques," *Proc. SIAM Int'l Conf. Data Mining (SDM),* 2003.

2) H. Polat and W. Du, "SVD-Based Collaborative Filtering with Privacy," *Proc. ACM Int'l Symp. Applied Computing (SASC),* pp. 791-795, 2005.

3) H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," *Proc. IEEE Int'l Conf. Data Mining (ICDM),* pp. 99- 106, 2003.

4) Z. Huang, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," *Proc. ACM SIGMOD Int'l Conf. Management Data,* pp. 37-48, 2005.

5) P. Heymann, D. Ramage, and H. Garcia-Molina, "Social Tag Prediction," *Proc. 31st Ann. Int'l ACM SIGIR Conf. Research Development Information Retrieval,* pp. 531-538, 2008.

6) E. Frias-Martinez, M. Cebria´n, and A. Jaimes, "A Study on the Granularity of User Modeling for Tag Prediction," *Proc. IEEE/ WIC/ACM Int'l Conf. Web Intelligence Intelligent Agent Technology (WIIAT),* pp. 828-831, 2008.

7) G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Trans. Knowledge Data Eng.,* vol. 17, no. 6, pp. 734-749, June 2005

IJERGS