ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Privacy-Preserving Enhanced Collaborative Tagging

Thammalla Alekhya
M.Tech Student,
Department of CSE,
KLR College of Engineering and Technology.

S.S.Madhavi
Associate Professor & HOD,
Department of CSE,
KLR College of Engineering and Technology.

Abstract:

Collaborative tagging is one of the most popular and diffused services available online. The main purpose of collaborative tagging is to loosely classify resources based on end-users feedback, expressed in the form of tags. Content/resource categorization has been seen a challenging research topic in recent year. Tag suppression is a privacy enhancing technique for the semantic Web. In this paper, users are assigned a tag to resources on the Web revealing their personal preferences. However, in order to avoid privacy attackers from profiling users based on their interests, they may wish to refrain from tagging certain resources. Consequently, tag suppression protects user privacy to a certain manner, but at the cost of semantic loss incurred by suppressing tags. In a nutshell, this technique poses a trade-off between privacy and suppression. In this paper, this trade off is investigated in a systematic fashion and provides an extensive theoretical analysis. User privacy is measure as the entropy of the users tag distribution after the suppression of some tags.

Keywords:

social bookmarking, tag suppression, privacy-enhancing technology, Shannon’s entropy, privacy-utility tradeoff.

1. Introduction:

Collaborative tagging became popular with the launch of sites like Flickr and Delicious. Since then, different social systems have been built that support tagging of a variety of resources. For a particular web object or resource, tagging is a process where a user assigns a tag to an object. A user can assign tags to a particular bookmarked URL on Delicious and 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, Livejournal, Wordpress, blog authors can add tags to their posts.The main purpose of collaborative tagging is to classify resources based on user feedback 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 browsing and discovery.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 filtering, browsing, or personalization.Tag suppression is the privacy-enhancing technology (PET) is used to protect privacy of end user. Tag suppression is a technique that has the purpose of preventing privacy attackers from profiling users interests on the basis of the tags they assign. It can affect the effectiveness of policy based collaborative tagging systems.

2. Literature Survey:

There are numerous approaches for collaborative tagging like data perturbation, tag prediction and tag recommendation.

2.1 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 use recommender systems to deal with information overload.

2.1.1 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 performs an algorithm and sends the users some information that allows them to compute the prediction [8].

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.

Limitation:

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.

2.1.2 SVD (Singular Value Decomposition):

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].

Limitation:

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

2.2 Tag Prediction:

Tag prediction concerns about 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.

2.2.1 Social Tag Prediction:

In this paper, D. Ramage, P. Heymann, 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 and giving the deeper understanding into the relationships between tags [3].

Limitation:

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 and higher entropy tags are harder to predict. When considering tags in their natural (skewed) distributions, data scarcity issues lead to dominate, so each tag improves classifier performance. This method performs poor in case of popular tags and distribution becomes poor with overall performance.

2.2.2 Granularity of User Modeling:

In this paper, Frias-Martinez, M. Cebrian, 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 levelswhich are used to improve accuracy. Prediction accuracy improves if the level of granularity matches the level of participation of the user in the community [4].

Limitation:

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).
2)How and when user models evolve from one granularity to the next.

2.3 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].

2.3.1.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 is searched by using TF-IDF (Term Frequency-Inverse Document Frequency). 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).

2.3.2 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. Implementation Details:

The architecture consists of privacy and policy layer. The aim of privacy layer is to preserve privacy of end user by applying tag suppression techniques and the aim of policy layer will be to enforce user preferences.

3.1 Tag Categorization:

Delicious dataset is used for processing. Dataset contain records in the form of triples (username, bookmark, tag). It contain 420 millions of these triples, but only subset of 12,41,029 triples is considered for processing. In this, Tags in dataset is categories into a few high-level tag categories using coarser categorization. Hierarchical cluster is formed by using Lloyds algorithm.

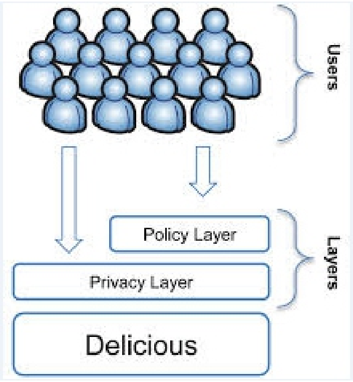


Figure 1: Architecture of enhanced social tagging service

Lloyds algorithm is used to group tags into 20 categories and again this main category is divided into 10 subcategories, result of this clustering into total 200 subcategories. The tags in subcategory are sorted in decreasing order of proximity to the centroid.

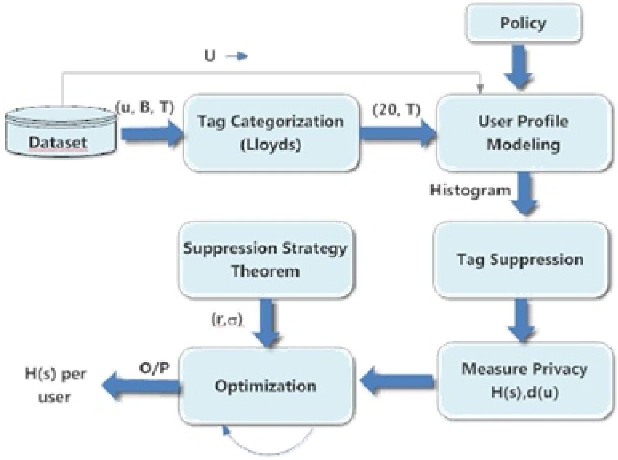


Figure 2: Architecture of collaborative tagging system with tag suppression.

3.1.1 Steps of categorization process:

a) Co-occurrence matrix.:

Computation and recording of simultaneous occurrence of two tags are done under common resource in the form of co-occurrence matrix. Tags may then be modeled as numeric vectors of co-occurrences, obtain as columns or rows within this matrix.

b) Cosine Distance:

A quantitive measure of semantic dissimilarity, namely the cosine distance between tag vectors, under the principle that similar tag should induce similar co-occurrence profiles

c)Clustering:

Clustering of tag is done using the Lloyd's algorithm. Replacing all tags within each cluster by a common representative tag and minimizing average semantic distance.

3.2 Tag Suppression:

Tag suppression is privacy enhancing technology (PET). It is used to protect end user privacy. It is a technique that has the purpose of preventing privacy attackers from profiling users interests on the basis of the tags they specify. In collaborative tagging, users tag resources on the web for e.g. music, images, and bookmark according to their personal preferences. In this way users interest is get reveal and any attacker able to collect such information. To avoid this, user may adopt privacy enhancing technology based on data perturbation. Tag suppression is data perturbative technique that allows a user to refrain tag of certain resources in such a manner attacker is not able to capture their interest precisely.

3.2.1 Privacy Enhancing Techniques (PET):

a) Refrain Tag:

In this, tags are refrain by applying-S = q – r / 1 – d
Where,
d = suppression rate (total no of tag remove), r = suppression strategy.

b) False Tag

Distort profile of user so that attacker is unable to make prediction of user interest.

c) Replace Tag:

Replace specific tags that show interest of user by general tag.

3.3 Measure Privacy:

Information theoretic criteria are used to quantify the privacy of user profile. Two fundamental quantities of information theory, namely Shannons entropy and Kullback-Leibler (KL) divergence is used to measure privacy. H(s) is Shannons entropy and d (u) is Kullback-Leibler (KL) divergence.

Mathematical Model:

Input Set: I (u, T, B)
Where,
u = User,
T = Tag,
B = Bookmark

Output Set:

i.Shannon Entropy: H (s1, s2, …sn) Where,
s1......n = Users in dataset
ii.KL Divergence: d (u1, u2, …un)
Where,
u1……un = Users in dataset

Processing Set:

1.Shannon Entropy:

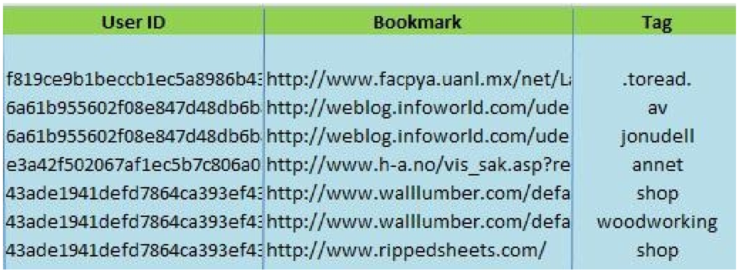
It is used to measure privacy.
H(s) = - ∑ si log2 si
Where,
si = PMF for all categories.
H(u) = log2 n
It indicates uniform distribution.

2.KL Divergence:

D(s || u) = log2 n - H(s) Where,
p = uniform distribution.

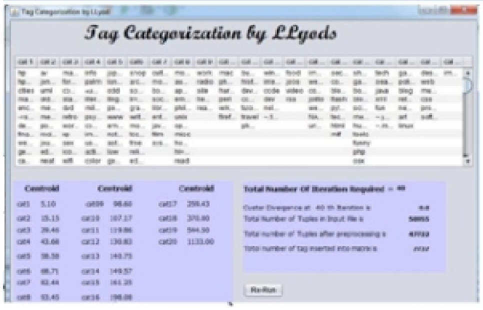
4. Dataset and Results:

Data is collected from Delicious to evaluate this approach. Delicious dataset is used for processing. Dataset contain records in the form of triples (username, bookmark, tag). Dataset contain 420 millions of these triples, but only subset of 12,41,029 triples is considered for processing. In this, Tags in dataset is categories into a few high-level tag categories using coarser categorization.

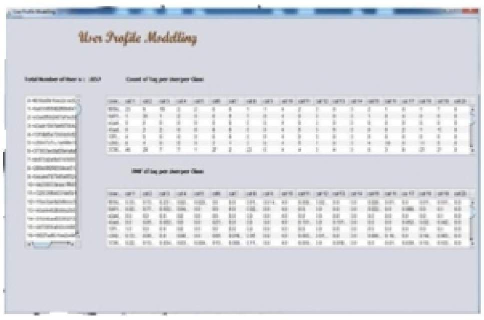


User ID	Bookmark	Tag
f819ce9b1beccb1ec5a8986b43	http://www.facpya.uanl.mx/net/La	.toread.
6a61b955602f08e847d48db6b	http://weblog.infoworld.com/ude	av
6a61b955602f08e847d48db6b	http://weblog.infoworld.com/ude	jonudell
e3a42f502067af1ec5b7c806a0	http://www.h-a.no/vis_sak.asp?re	annet
43ade1941defd7864ca393ef43	http://www.walllumber.com/defa	shop
43ade1941defd7864ca393ef43	http://www.walllumber.com/defa	woodworking
43ade1941defd7864ca393ef43	http://www.rippedsheets.com/	shop

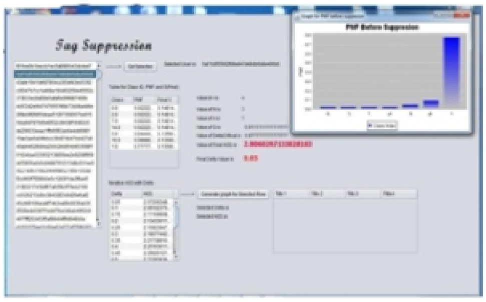
This shows the tag categorization process. A tag is categorized into 20 categorize by using Lloyds algorithm. Lloyds algorithm is a clustering algorithm.



This shows the user profile modeling. A PMF (Probability Mass Function) is calculated of each user per category.

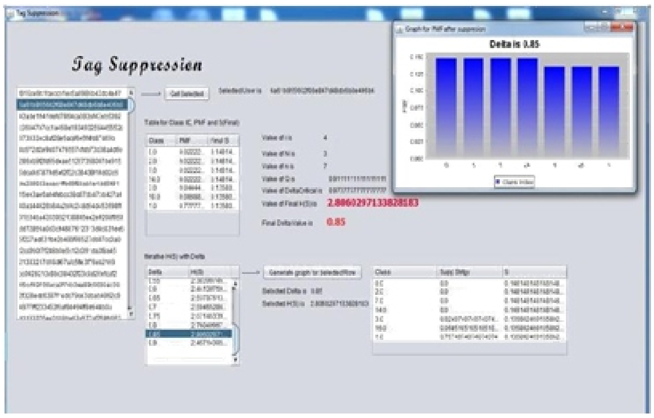


This shows PMF before suppression .



This shows the PMF after suppression . A tag is suppressed here so that user's privacy is get protected.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal


Tag Suppression
Get Selected
Graph for PMF after suppression
Delta is 0.85
2.8060297133828183
Final Delta Value is 0.85
Generate graph for selectedRow

5. Conclusion:

In this paper, the privacy of end user is preserved using tag suppression. The enhanced collaborative tagging architecture is proposed that consists of a bookmarking service and two additional services built on it. The former service enables users to set policies both to block undesired web content and to denote resources of interest. The Tag suppression is a privacy preserving technology based on data perturbation. The combination of these two services allows broadening the functionality of collaborative tagging systems and, at the same time, providing users with a mechanism to preserve their privacy while tagging. Future scope is an extensive performance evaluation of collaborative tagging system architecture, showing its effectiveness in terms of privacy guarantees, data utility, and filtering capabilities for two key scenarios, for example, parental control and resource recommendation.

References:

[1]Javier Parra-Arnau, Andrea Perego, Elena Ferrari, Jordi Forne´, and David Rebollo-Monedero,” Privacy-Preserving Enhanced Collaborative Tagging”, IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO. 1, January 2014.

[2]H. Polat and W. Du, SVD-Based Collaborative Filtering with Privacy, Proc. ACM Intl Symp. Applied Computing (SASC), pp. 791-795, 2005.

[3]P. Heymann, D. Ramage, and H. Garcia-Molina, Social Tag Prediction, Proc. 31st Ann. Intl ACM SIGIR Conf. Research Development Information Retrieval, pp. 531-538, 2008.

[4]E. Frias-Martinez, M. Cebrian, and A. Jaimes, A Study on the Granularity of User Modeling for Tag Prediction, Proc. IEEE/ WIC/ACM Intl Conf. Web Intelligence Intelligent Agent Technology (WIIAT), pp. 828-831, 2008.

[5]H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, On the Privacy Preserving Properties of Random Data Perturbation Techniques, Proc. IEEE Intl Conf. Data Mining (ICDM), pp. 99- 106, 2003.

[6]Z. Huang, W. Du, and B. Chen, Deriving Private Information from Randomized Data, Proc. ACM SIGMOD Intl Conf. Management Data, pp. 37-48, 2005.

[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.

[8]H. Polat and W. Du, Privacy-Preserving Collaborative Filtering Using Randomized Perturbation Techniques, Proc. SIAM Intl Conf. Data Mining (SDM), 2003.

Authors:

Thammalla Alekhya, M.Tech Student,Department Of CSE, KLR College Of Engineering And Technology

Mrs. S.S.Madhavi, working as an Associate professor in the Department of Computer Science and Engineering, KLR College Of Engineering and Technology, pursuing her Ph.D. in Computer Sci-ence and Engineering from JNTUH, Hyderabad.Her research area is BigData.