International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.1, 2015

RESEARCH ARTICLE

ISSN: 2321-7758

PRIVACY PROTECTION USING TAG SUPPRESSION IN COLLABORATIVE TAGGING

VAISAGAN[1], N.MADHAN[2] , L.JOSHUA[2,]

[1]Asst. Professor, [2] M.Tech Student

Department of Information Technology, Jeppiaar Engineering College

Article Received: 18/02/2015 | Article Revised on:24/02/2015 | Article Accepted on:28/02/2015

ABSTRACT

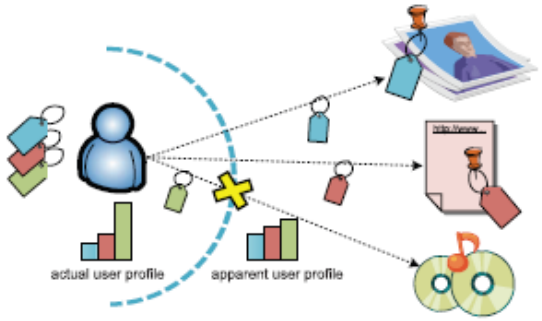
Collaborative tagging is one of the most popular services available online, and it allows end user to loosely classify either online or offline resources based on their feedback, expressed in the form of free-text labels (i.e., tags). Although tags may not be perse sensitive information, the wide use of collaborative tagging services increases the risk of cross referencing, thereby seriouslycompromising user privacy. In this paper, we make a first contribution toward the development of a privacy-preserving collaborative tagging service, by showing how a specific privacy-enhancing technology, namely tag suppression, can be used to protect end-user privacy. Moreover, we analyze how our approach can affect the effectiveness of a policy-based collaborative tagging system that supports enhanced web access functionalities, like content filtering and discovery, based on preferences specified by end users.

INTRODUCTION

Collaborative tagging is one of the most diffused and popular accommodations available online. First provided byconvivial bookmarking sites only-for example, Ambrosial (http://ambrosial.com), Digg (http://digg.com), Stumble-Upon (http://stumbleupon.com) —it is currently fortifiedby proximately any type of convivial web application, and it is utilizedto annotate any kind of online and offline resources(e.g., webpages, images, videos, movies, music, and even blog posts).The main purport of collaborative tagging is to loosely relegate resources predicated on end-user's feedback, expressed in the form of free-text labels (i.e., tags). The novelty of such an approach to content/resource categorization has been visually perceived, in recent years, as a challenging research topic.

In fact, collaborative tagging may be the substratum for a semantic network connecting online resources predicated on their characteristics, and not only their URIs. At the same time, the undefined semantics of tags, which are per sequivocal and expressed in multiple languages, makes it arduous to enforce semantic interoperability and to granta plausible level of precision when determining the "meaning" of a tagredicated on such considerations, most research work has investigated how to efficaciously reuse tag amassments(referred to as folksonomies) in the semantic Web framework(visually perceive, e.g., [1], [2], [3]), and analyzed collaborative tagging practices to enforce

strategies addressing the semantic ambiguity issue (e.g., as in [4]), by statistically analyzing tag amassments to infer, whenever possible, a semantic alignment of at least a subset of tags.



EXISTING SYSTEM

The aim of this layer will be to enforce utilize predilections, intensionally denoting resources on the substructure of the set of tags associated with them, and, possibly, other parameters concerning their trustworthiness (the percentage of users who have integrated a given tag, the gregarious relationships and characteristics of those users, etc.). This is an incipient research topic, and, to the best of our cognizance, the only work addressing this issue is reported in [5], where a multilayer policy-predicated collaborative tagging system is described.

Consequently, collaborative tagging requires the enforcement of mechanisms that enable users to bulwark their privacy by sanctioning them to obnubilate certain utilizer-engendered contents (unless they optate otherwise), without making them useless for the purposes they have been provided in a given online accommodation. This denotes that privacy-preserving mechanisms must not negatively affect the accommodation precision and efficacy (e.g., tag-predicated browsing, filtering, or personalization).

OVERVIEW OF THE PROPOSED APPROACH

As we discussed in Section 1, convivial bookmarking accommodations are among the most used convivial accommodations, and, thanks to their fortification to collaborative tagging, they can be currently considered as the most valuable cognizance acquisition implements, as far as online resources are concerned.

We have additionally pointed out that collaborative tagging is not exploited to its full potential, since it is typically used just to fortify tag-predicated resource browsing and search, despite the fact that collaborative tagging systems can be facilely enhanced without modifying their core architecture, because they provide access to the amassed information via APIs, which can be facilely exploited by external applications. One of the reasons is that the size of the amassed data sets is too immensely colossal to sanction the enforcement of even simple mechanisms, concerning, for example, personalization, content filtering, and quality assessment.

TAG SUPPRESSION

In our scenario of collaborative tagging, users tag resources on the web, for example, music, pictures, videos or bookmarks, according to their personal predilections. Users therefore contribute to describe and relegate those resources, but this is ineluctably at the expense of revealing their profile. To evade being accurately profiled by tagging systems, or in general by any assailer able to accumlate such information, users may adopt a privacy-enhancing technology predicated on data perturbation.

The data-perturbative technology considered in this work is tag suppression, a technique that sanctions a utilizer to forbear tagging certain resources in such a manner that the profile resulting from this perturbation does not capture their intrigues so precisely. Our conceptually simple technique bulwarks utilizer privacy to a certain degree, but at the cost of the semantic loss incurred by suppressing tags.

Utilizer Profile Model

In the scenario of gregarious bookmarking, a utilizer browses the web bookmarks pages and assigns tags to them according to his/her profile of fascinates. As in our anterior work on tag suppression [8], we consider n tag categories, indexed by 1; . . . ; n, and model the profile of a utilizer as a probability mass function (PMF), that is, a histogram of relative frequencies of tags across these categories. Our model of utilizer profile is identically tantamount to the tag clouds that numerous collaborative tagging accommodations use to visualize the tags posted by users; a tag cloud is a visual representation in which tags are weighted according to their frequency of use.

Quantifying the Privacy of a Utilizer Profile

To make the presentation of our privacy criterion suited to a wider audience, next we shall review two fundamental quantities of information theory, namely Shannon's entropy and Kullback-Leibler (KL) divergence. Recall [25] that the Shannon entropy is defined as a quantification of the skepticality of the outcome of a desultory variable distributed according to such PMF, and that it is maximized, among all distributions on f1; . . . ; ng, by the uniform distribution. The KL divergence is often referred to as relative entropy, as it may be considered as a generalization of the entropy of a distribution, relative to another.

Optimization of the Privacy-Suppression

Tradeoff Equipped with a quantitative measure of privacy, now we are intrigued with culling a suppression strategy r so that s maximizes for a given . Formally verbalizing, we would relish to solve the multiobjective optimization quandary given by the privacy-suppression function. Albeit this optimization will be carried out for suppression rate as a quantification of utility, which makes the quandary tractable, the remnant of our work adheres to assess the loss in data utility and precision due to tag suppression in terms of certain percentages regarding missing tags on bookmarks, on the one hand, and on the other, in terms of erroneous positives and negatives.

$$P(\sigma) = \max_{\substack{0 \le r_i \le q_i \\ \sum_i r_i = \sigma}} H\left(\frac{q-r}{1-\sigma}\right)$$

Another paramount aspect that follows directly from our formulation is the intuitive fact that there must subsist a tag suppression rate beyond which the privacy-suppression function achieves its maximum value or critical privacy Pcrit ¼ HðuÞ ¼ log2 n. We refer to this suppression rate as the critical suppression rate and define it formally as crit ¼ minf: PðÞ ¼ Pcritg: ð3Þ Interestingly, it can be shown that crit ¼ 1 n mini qi, which implicatively insinuates that critical privacy is never procured for < 1, provided that q has at least one zero component. To visually perceive this, next we adumbrate a proof.

EXPERIMENTAL ANALYSIS

In this section, we delve into the impact that tag suppression may have on an enhanced collaborative tagging system predicated on Dainty. With this aim, Section 6.1 first examines the data set that we used to conduct the experimental evaluation. To make utilizer profiles tractable, Section 6.2 summarizes the methodology that we followed for mapping tags into a diminutive set of paramount categories of interest.

Data Set

In our experiments, we utilized the Delectable data set retrieved by the Distributed Artificial Perspicacity Laboratory (DAILabor), at Technische Universita¨t Berlin [37]. This data set includes those bookmarks and tags marked as public by approximately 950,000 users. The information is organized in the form of triples (username, bookmark, tag), each one modeling the action of a utilizer associating a bookmark with a tag. The data set contains 420 millions of these triples. It is worth mentioning that no preprocessing has been done, though usernames have been anonymized by applying a hash function. The data set that we considered in our analysis is a subset of the entire data set described above.

Tag Categorization

The representation of a utilizer profile as a normalized histogram across these 59,505 tags would be certainly unfeasible from sundry practical perspectives, mainly concerning the unavailability of data to reliably, accurately measure fascinates across such fine-grained categorization, and, should the data be available, its inundating computational intractability. Further, in our experiments but additionally in data mining procedures, a coarser categorization makes it more facile to have an expeditious overview of the utilize fascinates. For example, for users posting the tags "welfare," "Dubya" and "Katrina" it would be preferable to have a higher caliber of abstraction that enables us to conclude, directly from the inspection of the utilizer profile, that these users are intrigued with politics

Privacy

In our architecture, a utilizer designates a suppression rate designating the fraction of tags he/she is disposed to eliminate. The numerical method culled is the interiorpoint algorithm [39], [40], [41] implemented by the Matlab R2011a function fmincon. The algorithm in question makes utilization of the soi-disant barrier functions and has a polynomialtime intricacy with veneration to the number of subcategories.

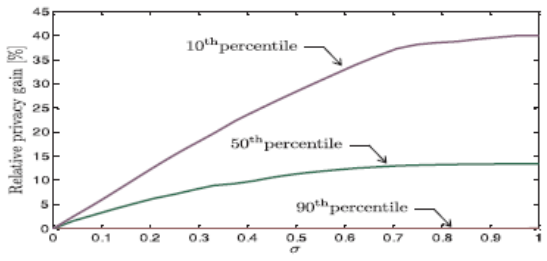
According to our hierarchical clustering, each category is composed of 10 subcategories. The two examples of subcategories shown here additionally illustrate a key result of the categorization process—tags in each subcategory are sorted in decrementing order of proximity to the centroid, which in practice betokens that those tags at the top of the list are the most representative tags of the subcategory they belong to.

1. This particular utilizer is identified by the string 674f779ba3b445937fd9876054a6e in [37]. Superimpose the optimal suppression strategy on the genuine utilizer profile q, to reflect the proportion of tags that the utilizer should eliminate from each subcategory of q to become the uniform distribution. Eminent is the fact that ri ¼ qi minj qj for any active subcategory i.
2. Lastly, the privacy aegis that users achieve as aresult of the suppression of tags. More accurately, we consider the case when all users in our data set have adhered to tag suppression and utilize the same suppression rate.

Data Utility

As we have just optically discerned, our approach avails users bulwark their privacy. Nevertheless, as in any perturbative mechanism, this bulwark comes at the expense of a loss in data utility.

In this section, we assess quantitatively the degradation in data utility caused by our privacy-bulwarking mechanism. In our precedent work on tag suppression [8], we utilized a preliminary, simplified measure of loss in data utility, namely the tag suppression rate. In this work, we do evaluate the impact that suppression has on utility, by 188 IEEE TRANSACTIONS ON ERUDITION AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014



We represent the ostensible profile of a particular utilizer, that is, the perturbed profile resulting from the suppression of tags and observed from the outside. We only show the active subcategories of this profile, i.e., those subcategories tagged by the utilizer. In this particular case, the utilizer posted 190 tags belonging to 49 subcategories.

Precision in Content Filtering

We quantitatively evaluate the degradation in the relegation of web content due to the suppression of tags. The subcategories of our example are "entertainment for children" and "entertainment for adults," identified, after the categorization process, as the subcategories 62 and 68, respectively.3 The threshold values for these subcategories are t62 ¼ 60% and t68 ¼ 10%. That verbally expressed, suppose w is the profile of a webpage and that w62 and w68 are the components of this profile, corresponding to the aforementioned subcategories To quantify the loss in the precision of this filter, we contemplate the following measures of utilitythe number of mendacious negatives and mendacious positives, precision, and recall. In our scenario, an erroneous negative is defined as a resource that changes from the initial state gainsaid to the final state granted, as a consequence of tag suppression.

COLLABORATIVE TAGGING CONCLUSIONS AND FUTURE WORK

Collaborative tagging is currently an astronomically popular online accommodation. Albeit nowadays it is fundamentally used to support resource search and browsing, its potential is still to be exploited. One of these potential applications is the provision of web access functionalities such as content filtering and revelation. The latter implements tag suppression, a privacy-preserving technology predicated on data perturbation.

The cumulation of these two accommodations sanctions us then to broaden the functionality of collaborative tagging systems and, at the same time, provide users with a mechanism to preserve their privacy while tagging.

REFERENCES

[1]. P. Mika, "Ontologies Are Us: A Unified Model of Social Networks and Semantics," Proc. Int'l Semantic Web Conf. (ISWC '05), Y. Gil, E. Motta, V. Benjamins, and M. Musen, eds., pp. 522-536, 2005.

[2]. X. Wu, L. Zhang, and Y. Yu, "Exploring Social Annotations for the Semantic Web," Proc. 15th Int'l World Wide Web Conf. (WWW), pp. 417-426, 2006.

[3]. B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and S. Gerd, "Evaluating Similarity Measures for Emergent Semantics of Social Tagging," Proc. 18th Int'l Conf. World Wide Web (WWW), pp. 641-650, 2009.

[4]. C. Marlow, M. Naaman, D. Boyd, and M. Davis, "HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, to Read," Proc. 17th Conf. Hypertext and Hypermedia (HYPERTEXT), pp. 31-40, 2006.

[5]. B. Carminati, E. Ferrari, and A. Perego, "Combining Social Networks and Semantic Web Technologies for Personalizing Web Access," Proc. Fourth Int'l Conf. Collaborative Computing: Networking, Applications and Worksharing, pp. 126-144, 2008.