ISSN:0975-9646

A Survey on Collaborative Tagging

Pradnya Deshmane, Prof.N.R.Wankhade

[1]*M.E. Student, Kalyani Charitable Trust's Late.G.N.Sapakal College of Engineering, Nashik*
[2]*Professor, Kalyani Charitable Trust's Late.G.N.Sapakal College of Engineering, Nashik*
Department of Computer Engineering, Savitribai Phule Pune University

Abstract **- Online social sites / systems has lots of facility to express feeling , deliverable content management , sharing of content , private information management etc. Tagging is also one of the most import feature while sharing some deliverable information with other user. Its extended facility is Collaborative tagging is which is also 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 per se sensitive information, the wide use of collaborative tagging services increases the risk of cross referencing, thereby seriously compromising user privacy. There some systems work on such user's privacy. All these techniques and research are studied to analyze such collaborative tagging system so that user preference is considered at time of tagging and his privacy is preserved.**

Keyword- **Tagging, Collaborative tagging, privacy preservation, free-text label**

I. INTRODUCTION

Collaborative tagging mainly focus on to task to loosely classify resources based on end user's feedback expressed in form of free-text labels. Recently research on content /resource categorization is in progress. Collaborative tagging may be the basis for a semantic network connecting online resources based on their characteristics, and not only their URIs. But meaningless tags are difficult to semantic analysis hence research also focuses on effective use of tags collections[1] [2] [3]. To analyze tags most effectively and properly statistical analysis of such collections of tags is focused [4]. While doing research on such collaborative tagging user preference is also considered to which we can call as policy layer. Hence research is also taking a path in that way also [5]. At the same time while working on policy layer and user preference analysis of tags tends to critical crisis of breaching user's privacy [6][7]. This privacy threat generates new facet to research. To test the privacy issue experiment is carried out on huge data set with tag suppression technique [8]. Tag suppression technique extends and tag suggestion / tag recommendation by the system and prediction is takes place and triggers the research accordingly and using tags proper web search possibilities are also considered[9][10][11][12][13][14][15][16][17]-[19].After that to trigger semantic research user to user relations using set of tags is considered also weight for particular relation is also in focus as a new technique[20] . Hence privacy protection in social tagging is another issue arises that suggest perturbing tags technique which manipulate exact tags and convert it to general tags[21][22][23][24][25]. Another aspect is also introduce which may crack the privacy. In this aspect user profile is considers and analyzed that unique profile tends towards revealing of identity. To test whether this way is proper or not measuring privacy of user profile techniques is developed [25]. Hence this statistical analysis works on user tagging policy and profile management policy that helps to preserve privacy of user in social networking.

II. RELATED WORK

P. Mika proposed a paper "Ontologies Are Us A Unified Model of Social Networks and Semantics" [1] This paper demonstrates the application of this representation by showing how community-based semantics emerges from this model through a process of graph transformation. We illustrate ontology emergence by two case studies, an analysis of a large scale folksonomy system and a novel method for the extraction of community-based ontologies from Web pages. Though it considers only two case studies it demonstrates that community based semantics can be analyzed from data

X. Wu, L. Zhang, and Y. Yu proposed a paper "Exploring Social Annotations for the Semantic Web" [2]. In this paper they explore a complement approach that focuses on the "social annotations of the web" which are annotations manually made by normal web users without a pre-defined formal ontology. Compared to the formal annotations, although social annotations are coarse-grained, informal and vague, they are also more accessible to more people and better reflect the web resources' meaning from the users' point of views during their actual usage of the web resources. But this system focuses only on bookmarking web based system. Also this system not considers the tagging approach.

B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and S. Gerd proposed a paper "Evaluating Similarity Measures for Emergent Semantics of Social Tagging" [3].Here they build an evaluation framework to compare various general folksonomy-based similarity measures, which are derived from several established information-theoretic, statistical, and practical measures. Their framework deals generally and symmetrically with users, tags, and resources. For evaluation purposes we focus on similarity between tags and between resources and consider different methods to aggregate annotations across users. This approach shows that we can define relations of users from tags. This is important from privacy preserving point of view.

C. Marlow, M. Naaman, D. Boyd, and M. Davis proposed a paper "HT06 Tagging Paper, Taxonomy, Flickr, Academic Article, to Read"[4]. In this paper, they provide a

short description of the academic related work to date. They offer a model of tagging systems, specifically in the context of web-based systems, to help us illustrate the possible benefits of these tools. Since many such systems already exist, we provide a taxonomy of tagging systems to help inform their analysis and design, and thus enable researchers to frame and compare evidence for the sustainability of such systems. They also provide a simple taxonomy of incentives and contribution models to inform potential evaluative frameworks. They present a preliminary study of the photo-sharing and tagging system Flickr to demonstrate our model and explore some of the issues in one sample system. Hence this paper is just giving us the basic idea about how tag functionality works in web based systems. This is important to clear basic ideas about tagging.

B. Carminati, E. Ferrari, and A. Perego proposed a paper "Combining Social Networks and Semantic Web Technologies for Personalizing Web Access"[5].This paper discussed about the issue that how to assess the trustworthiness of Web metadata ? They discuss how such issue can be addressed through the use of collaborative and Semantic Web technologies. The system we propose is based on a Web-based Social Network, where members are able not only to specify labels, but also to rate existing labels. Both labels and ratings are then used to assess the trustworthiness of resources' descriptions and to enforce Web access personalization. This system helps us to trigger thoughts on building a module that predict the unwanted tags.

R. Gross and A. Acquisti propose a paper "Information Revelation and Privacy in Online Social Networks"[6]. In this paper they study patterns of information revelation in online social networks and their privacy implications. We analyze the online behavior of more than 4,000 Carnegie Mellon University students who have joined a popular social networking site catered to colleges. We evaluate the amount of information they disclose and study their usage of the site's privacy settings. We highlight potential attacks on various aspects of their privacy, and we show that only a minimal percentage of users change the highly permeable privacy preferences. This paper forces us to give a thought on privacy on social web based system. This study of 4000 students proves that social data can breach privacy of particular user.

S.B. Barnes proposed a paper "A Privacy Paradox: Social Networking in the United States"[7]. This article will discuss the uproar over privacy issues in social networks by describing a privacy paradox; private versus public space; and, social networking privacy issues. This paper express exactly what privacy means and it guides us to build a one of the important layer of our system.

J. Parra-Arnau, D. Rebollo-Monedero, and J. Forne proposed a paper "A Privacy- Preserving Architecture for the Semantic Web Based on Tag Suppression"[8]. They propose an architecture that preserves user privacy in the semantic Web via tag suppression. In tag suppression, users may wish to tag some resources and refrain from tagging some others in order to hinder privacy attackers in their efforts to profile users' interests. Following this strategy, their architecture helps users decide which tags should be suppressed. This is one of the important strategy that helps us to find way for privacy preserving technique that may breach through tags.

J. Voß proposed a paper "Tagging, Folksonomy & Co - Renaissance of Manual Indexing?"[9]. This paper gives an overview of current trends in manual indexing on the Web. Along with a general rise of user generated content , there are more and more tagging systems that allow users to annotate digital resources with tags (keywords) and share their annotations with other users. This paper also shows that tagging should better be seen as a popular form of manual indexing on the Web.

G. Adomavicius and A. Tuzhilin proposed a paper "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions"[10]. This paper also describes various limitations of current recommendation methods and discusses possible extensions that can improve recommendation capabilities and make recommender systems applicable to an even broader range of applications. In this way we can get an idea of recommendation system.

P. Heymann, D. Ramage, and H. Garcia-Molina propose a paper "Social Tag Prediction"[11]. In this paper, they look at the "social tag prediction" problem. Given a set of objects, and a set of tags applied to those objects by users, can anyone predict whether a given tag could/should be applied to a particular object? They investigated this question using one of the largest crawls of the social bookmarking system del.icio.us gathered to date. Their results have implications for both the study of tagging systems as potential information retrieval tools, and for the design of such systems. This helps us to understand how tags are playing role to predict exact object that user shared with each other.

E. Frias-Martinez, M. Cebrian, and A. Jaimes propose a paper "A Study on the Granularity of User Modeling for Tag Prediction"[12]. One of the characteristics of tag prediction mechanisms is that, typically, all user models are constructed with the same granularity. In this paper they hypothesize and empirically demonstrate that 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 case user modeling is considered for predictions from tags.

Z. Yun and F. Boqin proposed a paper "Tag-Based User Modeling Using Formal Concept Analysis" [13]. All the tags used and resources collected by a user constitute the userpsilas personal tag space, which contains valuable information that can be used for building and enhancing the user model. In this paper, they propose an approach to mine user profile from onepsilas personal tag space. This system helps us to learn tag space mining.

A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke proposed a paper "Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering"[14].In this paper they consider the collaborative tagging. Data mining techniques, such as clustering, provide a means to remedy these problems by identifying trends and reducing noise. Tag clusters can also be used as the basis for effective

personalized recommendation assisting users in navigation. They present a personalization algorithm for recommendation in folksonomies which relies on hierarchical tag clusters. Their basic recommendation framework is independent of the clustering method, but they use a context-dependent variant of hierarchical agglomerative clustering which takes into account the user's current navigation context in cluster selection.

M. Bundschus, S. Yu, V. Tresp, A. Rettinger, M. Dejori, and H.-P. Kriegel proposed a paper "Hierarchical Bayesian Models for Collaborative Tagging Systems"[15] .In this paper, they reduce the data complexity in these systems by finding meaningful topics that serve to group similar users and serve to recommend tags or resources to users. They propose a well-founded probabilistic approach that can model every aspect of a collaborative tagging system. By integrating both user information and tag information into the well-known Latent Dirichlet Allocation framework, the developed models can be used to solve a number of important information extraction and retrieval tasks. Here meaningful recommendations are studied which will be important for us.

X. Li, C.G.M. Snoek, and M. Worring, proposed a paper "Learning Social Tag Relevance by Neighbor Voting" [16]. In this paper they work on intuition of user like they propose a neighbor voting algorithm which accurately and efficiently learns tag relevance by accumulating votes from visual neighbors. Under a set of well-defined and realistic assumptions, they prove that their algorithm is a good tag relevance measurement for both image ranking and tag ranking. This process helps us to understand the rating system using which content is promoted to particular user.

S. Marti and H. Garcia-Molina proposed a paper "Taxonomy of Trust: Categorizing P2P Reputation Systems" [17].The field of peer-to-peer reputation systems has exploded in the last few years. Their goal is to organize existing ideas and work to facilitate system design. They present taxonomy of reputation system components, their properties, and discuss how user behavior and technical constraints can conflict. In their discussion, they describe research that exemplifies compromises made to deliver a useable, implementable system. In this system to analyze the user to user relation they propose a basic idea of user reputation system.

K. Bischoff, C.S. Firan, W. Nejdl, and R. Paiu proposed a paper "Can All Tags Be Used for Search?"[18]. This paper is the first to present an in-depth study of tagging behavior for very different kinds of resources and systems - Web pages (Del.icio.us), music (Last.fm), and images (Flickr) - and compares the results with anchor text characteristics. They analyze and classify sample tags from these systems, to get an insight into what kinds of tags are used for different resources, and provide statistics on tag distributions in all three tagging environments. Since even relevant tags may not add new information to the search procedure, they also check overlap of tags with content, with metadata assigned by experts and from other sources. They discuss the potential of different kinds of tags for improving search, comparing them with user queries posted to search engines as well as through a user survey. The results are promising and provide more insight into both the use of different kinds of tags for improving search and possible extensions of tagging systems to support the creation of potentially search-relevant tags.

P. Heymann, G. Koutrika, and H. Garcia-Molina proposed a paper "Can Social Bookmarking Improve Web Search?" [19].Social bookmarking is a recent phenomenon which has the potential to give us a great deal of data about pages on the web. One major question is whether that data can be used to augment systems like web search. To answer this question, over the past year we have gathered what we believe to be the largest dataset from a social bookmarking site yet analyzed by academic researchers. Their dataset represents about forty million bookmarks from the social bookmarking site del.icio.us. They contribute a characterization of posts to delicious: how many bookmarks exist (about 115 million), how fast is it growing, and how active are the URLs being posted about (quite active). They also contribute a characterization of tags used by bookmarkers. They found that certain tags tend to gravitate towards certain domains, and vice versa. They also found that tags occur in over 50 percent of the pages that they annotate, and in only 20 percent of cases do they not occur in the page text, back link page text, or forward link page text of the pages they annotate. They conclude that social bookmarking can provide search data not currently provided by other sources, though it may currently lack the size and distribution of tags necessary to make a significant impact

J. J. Golbeck proposed "Combining Provenance with Trust in Social Networks for Semantic Web Content Filtering"[20]. In this paper, they present a two level approach to integrating trust, provenance, and annotations in Semantic Web systems. They describe an algorithm for inferring trust relationships using provenance information and trust annotations in Semantic Web-based social networks. Then, they present an application, Film Trust that combines the computed trust values with the provenance of other annotations to personalize the website.

H. Polat and W. Du proposed a paper "Privacy-Preserving Collaborative Filtering Using Randomized Perturbation Techniques"[21].Collaborative filtering (CF) techniques are becoming increasingly popular with the evolution of the Internet. To conduct collaborative filtering, data from customers are needed. However, collecting high quality data from customers is not an easy task because many customers are so concerned about their privacy that they might decide to give false information. They propose a randomized perturbation (RP) technique to protect users' privacy while still producing accurate recommendations.

H. Polat and W. Du proposed "SVD-Based Collaborative Filtering with Privacy"[22].Collaborative filtering (CF) techniques are becoming increasingly popular with the evolution of the Internet. Such techniques recommend products to customers using similar users' preference data. The performance of CF systems degrades with increasing number of customers and products. To reduce the dimensionality of filtering databases and to improve the performance, Singular Value Decomposition (SVD) is applied for CF. Although filtering systems are widely used

by E-commerce sites, they fail to protect users' privacy. Since many users might decide to give false information because of privacy concerns, collecting high quality data from customers is not an easy task. CF systems using these data might produce inaccurate recommendations. In this paper, they discuss SVD-based CF with privacy. To protect users' privacy while still providing recommendations with decent accuracy, they propose a randomized perturbation-based scheme.

H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar proposed a paper "On the Privacy Preserving Properties of Random Data Perturbation Techniques"[23].They present the theoretical foundation of this filtering method and extensive experimental results to demonstrate that in many cases random data distortion preserve very little data privacy. They also point out possible avenues for the development of new privacy-preserving data mining techniques like exploiting multiplicative and colored noise for preserving privacy in data mining applications.

Z. Huang, W. Du, and B. Chen proposed "Deriving Private Information from Randomized Data" [24].They propose two data reconstruction methods that are based on data correlations. One method uses the Principal Component Analysis (PCA) technique, and the other method uses the Bayes Estimate (BE) technique. They have conducted theoretical and experimental analysis on the relationship between data correlations and the amount of private information that can be disclosed based their proposed data reconstructions schemes. Their studies have shown that when the correlations are high, the original data can be reconstructed more accurately, i.e., more private information can be disclosed. To improve privacy, they propose a modified randomization scheme, in which they let the correlation of random noises "similar" to the original data.

J. Parra-Arnau, D. Rebollo-Monedero, and J. Forne proposed a paper "A Privacy- Protecting Architecture for Collaborative Filtering via Forgery and Suppression of Ratings" [25]. Recommendation systems are information-filtering systems that help users deal with information overload. Unfortunately, current recommendation systems prompt serious privacy concerns. In this work, they propose an architecture that protects user privacy in such collaborative-filtering systems, in which users are profiled on the basis of their ratings. Their approach capitalizes on the combination of two perturbative techniques, namely the forgery and the suppression of ratings. In their scenario, users rate those items they have an opinion on. However, in order to avoid privacy risks, they may want to refrain from rating some of those items, and/or rate some items that do not reflect their actual preferences. On the other hand, forgery and suppression may degrade the quality of the recommendation system. Motivated by this, they describe the implementation details of the proposed architecture and present a formulation of the optimal trade-off among privacy, forgery rate.

III. ANALYSIS

All of the above systems work on user privacy hiding while doing some semantic work. While working on tags analysis it is found that user to user or user to item relation is predictable. Hence to find solution on it tag suppression and user profile preservation techniques are used and analyzed to get desired result.

IV. CONCLUSIONS

This research conclude that there is need of modules in social networking system that has two filtering walls like one is for policy layer which considers user preferences and other one should privacy layer that work on tagging and user profile to preserve user privacy and sensible data protection.

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 Work sharing, pp. 126-144, 2008.

[6] R. Gross and A. Acquisti, "Information Revelation and Privacy in Online Social Networks," Proc. ACM Workshop Privacy Electronic Soc. (WPES), pp. 71-80, 2005.

[7] S.B. Barnes, "A Privacy Paradox: Social Networking in the United States," First Monday, vol. 11, no. 9, Sept. 2006.

[8] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forne´, "A Privacy-Preserving Architecture for the Semantic Web Based on Tag Suppression," Proc. Seventh Int'l Conf. Trust, Privacy, Security, Digital Business (TrustBus), pp. 58-68, Aug. 2010.

[9] J. Voß, "Tagging, Folksonomy & Co - Renaissance of Manual Indexing?" Computer Research Repository, vol. abs/cs/0701072, 2007.

[10] 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.

[11] 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.

[12] [12] E. Frı´as-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.

[13] Z. Yun and F. Boqin, "Tag-Based User Modeling Using Formal Concept Analysis," Proc. IEEE Eighth Int'l Conf. Computer Information Technology (CIT), pp. 485-490, 2008.

[14] Shepitsen, J. Gemmell, B. Mobasher, and R. Burke, "Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering," Proc. ACM Conf. Recommender Systems (RecSys), pp. 259-266, 2008.

[15] M. Bundschus, S. Yu, V. Tresp, A. Rettinger, M. Dejori, and H.-P. Kriegel, "Hierarchical Bayesian Models for Collaborative Tagging Systems," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 728-733, 2009.

[16] X. Li, C.G.M. Snoek, and M. Worring, "Learning Social Tag Relevance by Neighbor Voting," IEEE Trans. Multimedia, vol. 11, no. 7, pp. 1310-1322, Nov. 2009.

[17] S. Marti and H. Garcia-Molina, "Taxonomy of Trust: Categorizing P2P Reputation Systems," Computer Networks, vol. 50, pp. 472-484, Mar. 2006.

[18] K. Bischoff, C.S. Firan, W. Nejdl, and R. Paiu, "Can All Tags Be Used for Search?" Proc. 17th ACM Conf. Information and Knowledge Management (CIKM), pp. 193-202, 2008.
[19] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Can Social Bookmarking Improve Web Search?" Proc. Int'l Conf. Web Search Data Mining (WSDM), pp. 195-206, 2008.
[20] J. J. Golbeck, "Combining Provenance with Trust in Social Networks for Semantic Web Content Filtering," Proc. Int'l Conf. Provenance and Annotation of Data, pp. 101-108, 2006.
[21] H. Polat and W. Du, "Privacy-Preserving Collaborative Filtering Using Randomized Perturbation Techniques," Proc. SIAM Int'l Conf. Data Mining (SDM), 2003.
[22] H. Polat and W. Du, "SVD-Based Collaborative Filtering with Privacy," Proc. ACM Int'l Symp. Applied Computing (SASC), pp. 791-795, 2005.
[23] 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.
[24] 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.
[25] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forne´, "A Privacy-Protecting Architecture for Collaborative Filtering via Forgery and Suppression of Ratings," Proc. Int'l Workshop Data Privacy Management, Autonomous Spontaneus Security (DPM),pp.42-57, Sept. 2011.