SIAT: ON THE USE OF SIDE INFORMATION FOR MINING TEXT DATA

S. Manjula, K. Kavitha

***Abstract*—In many text mining applications, side-information is available. The side-information is along with the text documents. It contains different kinds, such as document provenance information, the links in the document, user-access behavior from web logs, or other non-textual attributes which are included into the text document. The attributes may contain a excellent amount of information for clustering purposes. It can be insecure to assimilate side-information into the mining process using in this cases. Because it can either improve the quality of the representation for the mining process, or can add noise to the process. In the proposed system we implement the SIDE INFORMATION ANALYZER TOOL(SIAT) to check the each and every rating manipulation. Its checks either the information is fully or partially viewed. It based on their rating. It also to check whether from same IP address with multiple logins giving rating constantly.**

***Index Terms*—Mining applications, Side information, SIAT.**

I. INTRODUCTION

The issue of text clustering appears in the context of many applications domains such as the web, social networks, and other digital applications. The rapidly enlarging the amount of text data in the context in the context. These huge online collections has led to an interest in integrate scalable and effective algorithms.

A large amount of work has been done in recent years on the problem of clustering in text collections [1], [2], [3], [4], [5]. These collections in the database and information retrieved communities. In many application domains, a large amount of side-information is also associated along with the documents. Because text documents typically occur in the context of a variety of applications in which there may be a large amount of other related database attributes or meta-information which may be functional to the clustering process.

In this side information we track user access behavior of the documents, the user-access behavior may be represent in the form of web logs. For each document, the meta information may communicate to the browsing behavior of the dissimilar users. Many text documents carry the links among them, which can also be served as attributes. Such links to contain a lot of useful information extract purpose. In previous case, such attributes may often to provide perception about the correlation among documents in a way which may not be simply attainable from raw content [6] [7] [8].

Several web documents have meta-data analogous with them which communicate to different kinds of data such as the other information about the origin of the document. In other cases, data such as possession, location, or secular information may be informative for mining purposes. In a number of networks and user-sharing applications, documents may be quite collection [9-15].

In the existing model they we implement the tool is called Side Information Analyzer Tool (SIAT). It checks login networks and their location for all the users [16] [17].

II. SYSTEM ANALYSIS

A. Mining Applications with Side Information

Many mining application to contains the self information is available along with it. Such side-information maybe of different kinds, such as document provenance information, the links in the document, user-access behavior from web logs, or other non-textual attributes which are embedded into the text document. Many forms of text contains a large amount of meta-information. Which is used to improve the clustering process. The clustering proves to combine a two techniques, interactive portioning techniques and probability estimation process usage of different kinds of side information. This approach used to design both clustering and classification algorithm. However, the relative importance of this side-information may be difficult to estimate, basically when some of the information is noisy [18-25].

It monitor whether the time logged in and logged out matches the stipulated time for viewing any messages and giving appropriate rating. It checks whether the time logged in and logged out matches the specified time for viewing any messages and giving appropriate rating [26].

It checks whether the time logged in and logged out matches the stipulated time for viewing any messages and giving appropriate rating. Its checks either the information is fully or partially viewed. It based on their rating. Thus by doing this the attacker is also detected and given counter measure to their attacks [27].

B. Problems Using Self Information

It can be insecure to integrate side information into the mining process. Because it can either upgrade the quality of

***S.Manjula**, Mailam Engineering College, Mailam, Tamil Nadu, Chennati.*
***K.Kavitha**, Mailam Engineering College, Mailam, Tamil Nadu, Chennati.*

the representation for the mining process, or can noise to process. Side information will either be neglected or collected and dissipated. Sometimes mining application to taking up time perform their task. It is major disadvantage of it [28-35].

C. SIAT Tool

SIDE INFORMATION ANALYZER TOOL(SIAT) to monitor each and every rating manipulation. It monitor the whether the time logged in and logged out matches the specified time for viewing any messages and giving appropriate rating. It checks either the information provided is fully or partially viewed thus accordingly their ratings given. It also checks whether the login in network which is providing rating is either from same location or its location varies each and every time from vast geographical location. It also checks whether from same IP address multiple logins giving ratings continuously. Thus by doing this the attacker is also detected and given counter measure to their attacks.

III. USAGE OF SIAT TOOL

We implement SIDE INFORMATION ANALYZER TOOL(SIAT) to monitor each and every rating manipulation. It monitor whether the time logged in and logged out matches the stipulated time for viewing any messages and giving correct rating. It check either the information provided is fully or partially viewed thus accordingly their ratings given. It also checks whether the login in network which is providing rating is either from same site or its varies each and every time from huge geographical location. It also to monitor the whether from same IP address multiple logins. Its giving ratings continuously. Thus by doing this the attacker is also detected and given counter measure to their attacks.

A. Benefits of SIAT Tools

It reasure the valid rating while being noticeable online so it intercept attack before existing. SIDE INFORMATION ANALYZER TOOL(SIAT) is execute so that the it examine or evaluate the rating by the user from all the features and then only it left to observed by others.

The attacker is perceive and their system is given counter attack so that they won't to be perform further attack in future. The attacker will be ostracize so that only with proper explanation for their existing act they will be able to get connections.

III. METHODS

In proposed system we follow several kinds of modules. These are all described below.

A. Generate Client and Server Communication

Client and server systems are to develop in this module. Then they construct the communication channel in between them. They communicate each other pass information to each other. In this module the client can request services from the server the server. Then those clients can provide the ratings in the standing system.

In client and server communication the major part is registration. The client and sever to register the information to their own admin. In that registration client and server to give their own IP address, port numbers and their names.

Every server to give response to the all client requests. The server to maintain that information to all client requests. The client also to maintain their registration information, and also to monitor the rating of the system.

B. Social Reputation System Model

To designing a social reputation system which initiate texts, images, videos, audio and publishing them in online. In this module we produce a reputation system that contains products. Those products with their viewing time and or with strength is given by the sender.

The reputation system to maintain the rating of all the client. The rating is manipulated based on the score. The rating will maintaining based on every client rating. The priority wise the user will be ordered in this system [36] [37].

C. Client Usage

In this system we initiate the clients through users. The users to requesting their rating for repeating system according to that activity. The huge reputation to maximize scores. If they needed for giving prioritization to the products. The client to provide the rating that product will be prioritized accordingly. If any one user to given the corrupted data the system will analyze the user data. If they find any dishonest rating given by the user those system will rating and to given less priority.

IV. IMPLEMENTING INFORMATION ANALYZER TOOL(SIAT)

In the proposed system we implement the side information analyzer tool (SIAT). It's analyze whether any repeated scores are given by user. Repeated scores honest or dishonest data to given by user. In case any reputation rating given by any honest or dishonest users, the SIAT to determines it. For SIAT implementation we enlarge the parameters of this tool.

SIAT tool will monitor the user data completely. If the user is approved, and also the user to uses the product appropriately then only the tool to give the rating to the user.

If the SIAT tool to found either the dishonest rating or unauthorized user to uses the product, the rating of that user to be rejected without any information. At that time the user will lose their data permanently. In case, the attacker tried to attack reputation system, at that time the tool will be given to counter attack.

SIAT tool to audit all the users information. If any dishonest rating to induced by user it will be attacked by the system. In case, the user to attacked repeatedly the system will take the counter attack.

V. CONCLUSION

In this proposed model the server to manage and monitor the users rating and also the reputation system. The client to viewed the reputation scores, which given by server. The scores will monitored based on their scores. It will prioritized. The user to give any dishonest rating it will attack the reputation system. For this purpose we create tool called side information analyzer tool. This tool to monitor the reputation system, and client rating. If any dishonest rating to given by user the tool will reject the rating.

REFERENCES

[1] C. C. Aggarwal and H. Wang, Managing and Mining Graph Data. New York, NY, USA: Springer, 2010.

[2] C. C. Aggarwal, Social Network Data Analytics. New York, NY, USA: Springer, 2011.

[3] C. C. Aggarwal and C.-X. Zhai, Mining Text Data. New York, NY, USA: Springer, 2012.

[4] C. C. Aggarwal and C.-X. Zhai, “A survey of text classification algorithms,” in Mining Text Data. New York, NY, USA: Springer, 2012.

[5] C. C. Aggarwal and P. S. Yu, “A framework for clustering massive text and categorical data streams,” in Proc. SIAM Conf. Data Mining, 2006, pp. 477–481.

[6] C. C. Aggarwal, S. C. Gates, and P. S. Yu, “On using partial supervision for text categorization,” IEEE Trans. Knowl. Data Eng., vol. 16, no. 2, pp. 245–255, Feb. 2004.

[7] C. C. Aggarwal and P. S. Yu, “On text clustering with side information,” in Proc. IEEE ICDE Conf., Washington, DC, USA, 2012.

[8] R. Angelova and S. Siersdorfer, “A neighborhood-based approach for clustering of linked document collections,” in Proc. CIKM Conf., New York, NY, USA, 2006, pp. 778–779.

[9] J. Chang and D. Blei, “Relational topic models for document networks,”in Proc. AISTASIS, Clearwater, FL, USA, 2009, pp. 81–88.

[10] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, “Scatter/Gather: A cluster-based approach to browsing large document collections,” in Proc. ACM SIGIR Conf., New York, NY, USA, 1992, pp. 318–329.

[11] Dhillon, S. Mallela, and D. Modha, “Information-theoretic coclustering,” in *Proc. ACM KDD Conf.*, New York, NY, USA, 2003, pp. 89–98.

[12] P. Domingos and M. J. Pazzani, “On the optimality of the simple Bayesian classifier under zero-one loss,” *Mach. Learn.*, vol. 29, no. 2–3, pp. 103–130, 1997.

[13] M. Franz, T. Ward, J. S. McCarley, and W. J. Zhu, “Unsupervised and supervised clustering for topic tracking,” in *Proc. ACM SIGIR Conf.*, New York, NY, USA, 2001, pp. 310–317.

[14] G. P. C. Fung, J. X. Yu, and H. Lu, “Classifying text streams in the presence of concept drifts,” in *Proc. PAKDD Conf.*, Sydney, NSW, Australia, 2004, pp. 373–383.

[15] H. Frigui and O. Nasraoui, “Simultaneous clustering and dynamic keyword weighting for text documents,” in *Survey of Text Mining*, M. Berry, Ed. New York, NY, USA: Springer, 2004, pp. 45–70.

[16] S. Guha, R. Rastogi, and K. Shim, “CURE: An efficient clustering algorithm for large databases,” in *Proc. ACM SIGMOD Conf.*, New York, NY, USA, 1998, pp. 73–84.

[17] S. Guha, R. Rastogi, and K. Shim, “ROCK: A robust clustering algorithm for categorical attributes,” *Inf. Syst.*, vol. 25, no. 5, pp. 345–366, 2000.

[18] Q. He, K. Chang, E.-P. Lim, and J. Zhang, “Bursty feature representation for clustering text streams,” in *Proc. SDM Conf.*, 2007, pp. 491–496.

[19] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, “An evaluation of feature selection for text clustering,” in *Proc. ICML Conf.*, Washington, DC, USA, 2003, pp. 488–495.

[20] Q. Mei, D. Cai, D. Zhang, and C.-X. Zhai, “Topic modeling with network regularization,” in *Proc. WWW Conf.*, New York, NY, USA, 2008, pp. 101–110.

[21] R. Ng and J. Han, “Efficient and effective clustering methods for spatial data mining,” in *Proc. VLDB Conf.*, San Francisco, CA, USA, 1994, pp. 144–155.

[22] G. Salton, *An Introduction to Modern Information Retrieval.* London, U.K.: McGraw Hill, 1983.

[23] H. Schutze and C. Silverstein, “Projections for efficient document clustering,” in *Proc. ACM SIGIR Conf.*, New York, NY, USA, 1997, pp. 74–81.

[24] F. Sebastiani, “Machine learning for automated text categorization,” *ACM CSUR*, vol. 34, no. 1, pp. 1–47, 2002.

[25] M. Steinbach, G. Karypis, and V. Kumar, “A comparison of document clustering techniques,” in *Proc. Text Mining Workshop KDD*, 2000, pp. 109–110.

[26] Y. Sun, J. Han, J. Gao, and Y. Yu, “iTopicModel: Information network integrated topic modeling,” in *Proc. ICDM Conf.*, Miami, FL, USA, 2009, pp. 493–502.

[27] W. Xu, X. Liu, and Y. Gong, “Document clustering based on nonnegative matrix factorization,” in *Proc. ACM SIGIR Conf.*, New York, NY, USA, 2003, pp. 267–273.

[28] T. Yang, R. Jin, Y. Chi, and S. Zhu, “Combining link and content for community detection: A discriminative approach,” in *Proc. ACM KDD Conf.*, New York, NY, USA, 2009, pp. 927–936.

[29] T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: An efficient data clustering method for very large databases,” in *Proc. ACM SIGMOD Conf.*, New York, NY, USA, 1996, pp. 103–114.

[30] Y. Zhao and G. Karypis, “Topic-driven clustering for document datasets,” in *Proc. SIAM Conf. Data Mining*, 2005, pp. 358–369.

[31] Y. Zhou, H. Cheng, and J. X. Yu, “Graph clustering based on structural/attribute similarities,” *PVLDB*, vol. 2, no. 1, pp. 718–729, 2009.

[32] S. Zhong, “Efficient streaming text clustering,” *Neural Netw.*, vol. 18, no. 5–6, pp. 790–798, 2005.

[33] Banerjee and S. Basu, “Topic models over text streams: A study of batch and online unsupervised learning,” in Proc. SDM Conf., 2007, pp. 437–442.

[34] Dhillon, “Co-clustering documents and words using bipartite spectral graph partitioning,” in *Proc. ACM KDD Conf.*, New York, NY, USA, 2001, pp. 269–274.

[35] Jain and R. Dubes, *Algorithms for Clustering Data.* Englewood Cliffs, NJ, USA: Prentice-Hall, Inc., 1988.

[36] McCallum. (1996). *Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering* [Online]. Available: http://www.cs.cmu.edu/ mccallum/bow

[37] Silverstein and J. Pedersen, “Almost-constant time clustering of arbitrary corpus sets,” in Proc. ACM SIGIR Conf., New York, NY, USA, 1997, pp. 60–66.

S. Manjula B.tech (IT) in University College of Engg Villupuram, Doing M.E(CSE) in mailam engg college.

K. Kavitha Did MCA and ME(CSE) in Arunai Engg College, Thiruvanamalai. And worked as lecturer in prince sri venkaleshwara bhatmawathi engg college for 2years and now she is working as a assistant professor in Mailam Engg College.