2015 IEEE International Conference on Computational Intelligence & Communication Technology

De-Identification of Textual Data using Immune System for Privacy Preserving in Big Data

Amine Rahmani[1]

GeCoDe laboratory, department of informatics sciences, Dr. Tahar Moulay university of Saida

aminerahmani2091@gmail.com

Abdelmalek Amine[2]

GeCoDe laboratory, department of informatics sciences, Dr. Tahar Moulay university of Saida

amine_abd1@yahoo.fr

Mohamed Reda Hamou[3]

GeCoDe laboratory, department of informatics sciences, Dr. Tahar Moulay university of Saida

hamoureda@yahoo.fr

***Abstract*—With the growing observed success of big data use, many challenges appeared. Timeless, scalability and privacy are the main problems that researchers attempt to figure out. Privacy preserving is now a highly active domain of research, many works and concepts had seen the light within this theme. One of these concepts is the de-identification techniques. De-identification is a specific area that consists of finding and removing sensitive information either by replacing it, encrypting it or adding a noise to it using several techniques such as cryptography and data mining. In this report, we present a new model of de-identification of textual data using a specific Immune System algorithm known as CLONALG.**

Keywords—de-identification, privacy preserving, big data, immune systems, CLONALG

I. INTRODUCTION.

One of the advantages of big data's services is the ability of sharing and publish data over the network. Those data can be sorted in two major categories: normal like books and other textual documents, and sensitive information such as names, medical books, and social information generally. Those last requires a high tier of protection for its importance and sensitivity because if it will be linked together, it forms a total or partial presentation of their owner; which leads to identify him even if this data do not contain any explicit identifiers. The aggregation of this information can presents a unique identity of the person as like as the fingerprint. In addition, the data, once are stored on the web, it becomes accessible and treatable by a third party and, therefore, by other people who shared the same resources which make the privacy an essential aim to ensure. That's what gives birth to a new domain known as Privacy Preserving Data Publishing (PPDP) which offers a set of methods and techniques for protection of users' privacy. Many deeds are performed within this arena and a lot of approaches are published and used for that, these approaches can be covered on three essential groups:

- Heuristic based approaches in which a set of works are done using data mining algorithms in the form of adaptive modification of selected data. This is based on the fact that the selective data modification is an NP-hard problem so that this group of methods is addressed to the complex problems.
- Cryptography based approaches that are represented by a secure multiparty computation where the privacy is guaranteed basing on a probabilistic function in order to ensure that at the end for multiparty computations neither party can knows except its own input and the final results of computation.
- Perturbation and re-construction of data in which the proposed approaches consist of ensuring data by re-constructing randomly the distribution of data on such aggregated level.

One of the techniques of PPDP is the de-identification in which such system consists to detect and remove any information leads to the individuality of such user through his own data. In this work we propose a new approach based on Immune system in order to ensure privacy by detecting and modifying the information leading to identity of users so that we start, in the rest of the paper, with a presentation of basic concepts such as PPDP and its techniques focusing on de-identification and modification technique. Then we pass to the presentation of our idea and its results. And finally, we finished with the discussion of results and the final conclusion.

II. BASIC CONCEPTS

A. Privacy preserving data publishing

A data publisher is typically a data collector that consists in collecting data from Different sources, then pass it to a data miner or publish it to the public which can include an attacker.

The Fig 1 shows the point of view of (Fung, Wang, Chen & YU, 10) about a data publisher.

978-1-4799-6023-1/15 $31.00 © 2015 IEEE
DOI 10.1109/CICT.2015.146

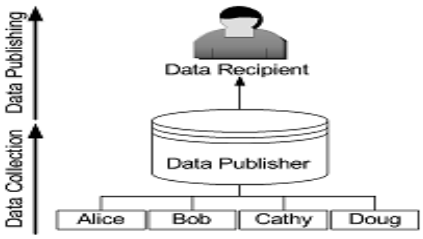

Fig. 1. A typically example of data publisher (Fung, Wang, Chen & YU, 10)

PPDP is a general term that refers to the set of approaches and algorithms that consist of protecting users' privacy within published data. However, PPDP is known also by privacy preserving data mining (PPDM) because of its ability to treat data and extract knowledge, even when some papers mentioned that the data mining concept in PPDP has a broad sense and not necessarily limited by classifying and building models.

B. De-identification

The de-identification is a word that refers to the fact of removing or hiding data in order to prevent any disclosure of identifiable data. This concept can be seen as a special topic for researches where the goal is to define an effective approach that ensures the maximum of de-identification without forgetting the reciprocal operation known as a re-identification in case when the owner desires to access the information. The efficiency of such approach resides in the re-identify data in real time when the owner decides to grow his own data.

C. K-anonymity

K-anonymity or de-identification of data refers to the ability of sharing data over the web without compromising the privacy of users. Otherwise, the k-anonymity is a security concept that consists to maximise the use of data while limiting the risk of disclosure of information that lead to the identity of the corresponding entity even if it is not an explicit identification.

Formally, let RT be a table with n attributes (A1... An) And QIRT the quasi-identifier (Sweeny, 02) corresponding to RT. Such system verifies the k-anonymity if and only if for each sequence of values in RT [QIRT], it appears at least k time in RT [QIRT].

D. Immune system

As a biological aspect, the immune system is the core of the defence system of the human being. The basic purpose of this concept is to find the infected cells in the physical structure; so that it either neutralizes it or destroys it depending to the characteristics of the cubicle and the way that it entered by to the body. The fact of choosing between neutralize or destroy cells is established on an existed memory on each defensive cell that we called antigens. This memory has an initial value once it is borne by the human; then it will be updated by the time using medical techniques by adding new antibodies characters to the memory via medicaments.

Technically, the artificial immune system is a set of bio-inspired algorithms of the innate immune system of the human being in which the goal is to build a robust and powerful information process that is applied to tackling complex problems of informatics skills and fields such as combinatorial optimization and security issues.

III. RELATED WORKS

Many works were done inside the field of de-identification. In (Kerschbaum & Oertel, 11), the authors present an idea that consists of the detection of anomalous events using immune system and graph theory science in order to secure the privacy of users according to their compartments. Other works that interpret the bulk of works were interested by medical books because of its sensitivity. The application of de-identification algorithms in this field was on structured data using terms that are recognized in the medical field. In (Uzuner, Luo, & Szolovits, 07) the authors present a state of the art on de-identification of medical records by evaluating some known systems in this topic such as HIPAA as part of the i2b2 (Informatics for Integrating Biology to the Bedside) projects. In (El Emam, Dankar et al, 09), the authors propose a new de-identification algorithm named OLA (Optimal Lattice Anonymization) and evaluate it by comparing its efficiency on medical data with three other algorithms: datafly, Samarati, and Incognito. In (Fernandes, Cloete et al, 13), the authors propose and discuss their new de-identification procedure on mental health records by evaluating the obtained results from applying this procedure on two different datasets: CRIS and MIST.

Other works were interested by domain of image treatment such as the written document represented in (Newton, Sweeny & Malin, 03) and (Goss, Sweeny et al, 08) where the authors propose their models for privacy protection of facial images in shared data sets. The primary end of these works is how to de-identify pictures while conserving their utility. In (Du & Ling, 11) the authors demonstrate their mind for protecting privacy of license plate images named IPCB (Inhomogeneous Principal Component Blur) so that they discuss the efficiency of this idea by comparing it with several previous works and also by variating some parameters. In (Gedik & Liu, 08) the writers describe in their paper, a new customizable k-anonymity model for shelter of privacy of users' location through location based services. Their model is grounded on two major features: a customizable framework for k-anonymity and spatio-temporal cloaking algorithm.

In addition, some other works were done within big data, such as in (Gardner, Xiong, Li & Lu, 09), where the inventors propose a prototype for de-identification of data. Their model allows the treatment of the main problem of big data and shared data generally, the heterogeneity of data so that this model permits to de-identify both structured and unstructured data forms.

IV. OUR IDEA

Consider the system SONIC presented in (Boukorca, Bellatreche, since & Faget, 13) as a case study. The

inventors of this system describe in their paper how they combined between the electronic design automation (EDA) domain of integrating circuits technics and the optimization domain in order to build a system of multi-query optimization. They consist of representing graphically the system of queries. The idea is simple, the queries sharing the same aim will be grouped in the same query. The Fig 2 indicates the general idea of those inventors:

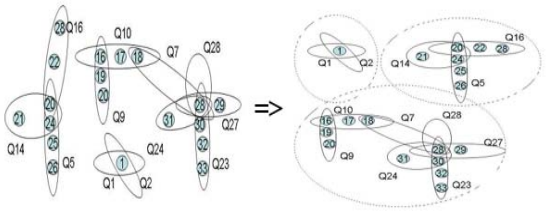

Fig. 2. General idea of the system presented in (Boukorca, Bellatreche, Senouci & Faget, 13)

This hypothesis still true in term of optimization advantages but not in the security standpoints. Nevertheless, this idea presents a major impediment resides in the fact that these queries can come from different sources; which leads to sharing users' secrete information. There where mechanisms and technics of de-identification can present a solution.

Our approach is framed of two major processes: de-identification and re-identification, in the remainder of this section we will identify our system step by step

A. De-identification

It is the main process that allows the protection of identity within users' data. This one is used to detect and replace any information that can be considered as an identifier so that this process is working on five steps: tokenization, codification, detection, storage and replacement.

1) Tokenization:

In this tone we use bag of words where we decompose both information of the user and data to be de-identified in a set of words.

2) Codification:

In this step each word from the identity is decomposed to set of characters, then each one is replaced by its ASCII code so that each word is considered as antigen; the same operation is applied to the words of the data to be de-identified. The words of these data will be considered as antibodies. The accompanying instance demonstrates a distinctive model of codification:

John => [J, o, h, n] => [74, 111, 104, 110]

3) Detection:

This step allows the detection of identifiers from the data using distance metric that is calculated between each antigen and all the antibodies as the following pseudo-code indicates:

Algorithm of detection

For each antigen AG ∈ list of antigens do

For each antibody AB ∈ list of antibodies do

Compute D the distance between AG and AB

If (D< threshold) then pass to storage step

Else pass to the next antibody

End for

End for

Pass to the replacement step

End

Algorithm. 1. Detection of identifiers

4) Storage:

This step consists to store the recognised antibodies in the memory. As we cited above, the natural immune system has an initial value that is born with the owner. In our system this initial value is created at the inscription of user in our system so that the information that he fills up in inscription formula will be transformed to antigens and stored in the memory. After that by the time when new antibodies are recognized the memory will be updated by adding those antibodies that at the end represents some variants of the words considered as antigens.

The way that words are stored is similar to the MapReduce principle in which the word represents the key and its positions represent the address of storage and that is called mapping in big data.

5) Replacement:

After finishing from detecting and storing identifiers the de-identification step starts by perturbing and replacing every word from the memory in the original data as follows:

Algorithm replacement

For a set of iterations do

Choose randomly two antibodies from the memory

Concatenate the chosen antibodies

Mutate the result for number of times

Choose a random position and split the result on two new antibodies

Decoding the new antibodies to new words

Replace the old chosen words with the new ones

End for

Algorithm. 2. Replacement process after detection of identifiers

As the algorithm 2 shows, the replacement step works as follows: The antibodies, once are recognised, it will be stored in memory of the owner. Then for a number of iterations. In each one, two antibodies will be randomly selected and concatenated. Afterward, and fro several times, two random components will be chosen from the concatenated word and permuted. Then a random position will be taken in which the outcome will be split in two sections. The first one replaces the first chosen antibody and the second replaces the second single. Ultimately, the replacement antibodies will be decoded to replace the identifiers in the original text.

B. Re-identification

This process consists to get back the de-identified documents to their original form by re-entering each word to his situation. The way to do that, as easily as the mathematical function is made out in storage step. The operation of cutting is executed in this step in which each antibody in the memory is a key that identifies for each position which word must come along. After determining the positions of antibodies, the system decodes those antibodies to generate the genuine words that will be added to the text files.

V. EXPERIMENTS AND RESULTS

We carried on a set of experiments by evaluating our model basing on three essential criteria: performance of our system in detecting and removing identifiers, uncertainty of our example and data utility after de-identification.

A. In term of performance

We managed a set of experiments basing on several parameters such as the number of detected and removed identifiers depending on the number of words in the original text, and number of specified identifiers by the user (antigens). As well as we mention in algorithm 1 above, the decision of such antibody if it equates to an identifier or not is relied to a doorway that we set before. And for more official results we conducted our experiment using different distance features: Euclidean, Manhattan, MINKOWSKI, COSINUS, and TCHEBYCHEV. We apply our system on five different texts with several sizes starting from the smallest one to the biggest and defining for each one a set of identifiers that represent antigens. The identifiers are general information of the users, such as name, email, business, birthday, ISBN… etc.

In our system we use bag of words for text representation but even with that we don't need to stem and lemmatization because we compute the distance between the ASCII codes of the words. The closest words to each other can be with the minimum of likeness between their characters. Here where the threshold can play an important role.

We evaluate our results using six criteria: average number of detected identifiers which are the antibodies (DI), average of number of replaced identifiers (RI). The cited measures are counted by words. We use also average of percentage of success (PS) which is the rate of number of removed identifiers on the number of detected ones and finally average of taken time (T) of de-identification counted by milliseconds.

After many experiments in which we used several thresholds and several texts with different sizes for each distance, we select the best doorstep to be represented in the following results. The table 1 and figure 3 represent the results of the best threshold for each text.

TABLE. 1. RESULT OF DE-IDENTIFICATION PROCESS

Distance Measure	Best threshold	DI	RI	PS	T
EUCLIDEAN	130	2 530	2 313	91.4	280.413
MANHATTAN	100	2 569	2 357	91.7	70.003
COSINUS	120	2 574	2 246	87.2	76.0034
MINKOWSKI	100	2 672	2 469	92.4	209
TCHEBYCHEV	150	2 627	2 499	95.1	77.2002

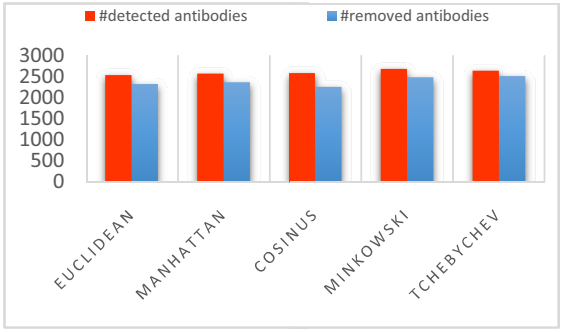

Fig. 3. result of de-identification process

From the outcomes shown in table 1 and Fig 3, we can clearly notice that the number of detected identifiers is guided by two principal parameters. The number of initial identifiers (antigens) and the value of threshold. We acknowledge that the best results in term of detected and removed identifiers is done using TCHEBYCHEV distance with 95.1% of success rate. Meanwhile, the worst results are given using COSINUS distance with no more than 87.2% of rate of success. However, the reason of that not all the detected identifiers are removed is that because of the random choose in replacement step. In term of time of computing, we notice that the use of MANHATTAN distance takes less time than the others while the MINKOWSKI distance takes longer.

B. In term of uncertainty

This criteria is a known criteria in which the evaluation of such model is performed by testing if a removed information can still be predicted or not.

Expect at the size of detected identifiers equals to 'n' and the medium size of an identifier equals to 'm'. The quality of identifiers that will limit the used characters on uppercases and lowercase letters and numbers which gives 62 possibilities for each font. Consequently, such

attack using dictionary or brute force attack, we got 62* (n*m)! Possibilities to construct a word.

However, this technique of replacing identifiers in our system represents a major advantage resides in his big complexity of finding the original words that carries on to the current replacements. This complexity is an exponential one because of the iterative behaviour based on nested loops.

C. *In term of data utility*

Our system consists to detect and take away information that lead to an identity independently of the context of data. Only the threshold does, otherwise, if an optimal threshold is chosen the removed information generally neither possess a weight nor influence other technics and algorithms such as indexation and retrieval models because this mechanism doesn't change the context of the information.

VI. CONCLUSION AND FUTURE WORKS

This report represents our proposition of Immune System based approach for discovering and removing information that can precede to the identity closure of the data's owner. After we confront many experiences basing on various metrics such as thresholds and distances, we arrived to a close that the bio-inspired methods generally and immune system specifically can represent an excited tools for de-identification technics. In addition, the mechanism of detecting identifiers on such information is not only depend on the size of the original data and a number of pre-defined identifiers, but also by the content of each one of them. Some other significant detail concerning the threshold, we reason that the value of the threshold plays an important role on detection step in such way. If we touch on a small one some variants of the identifiers will not be noticed, instead if we set up a very big threshold our system will get in charge some words as identifiers so that it is not which can, consequently, influences the data's context.

As a future work we will define an approach that allows to look for the optimal threshold basing on the substance of the identifiers in which our organization will be able to find only the identifiers. In other hand we will search for new approaches to ameliorate this system. Also, we will define other models using other bio-inspired techniques in order to get more performance of de-identification methods.

References

[1] Ahcène, B., Ladjel, B., Sid-Ahmed, B. S., & Zoé, F. (2013). SONIC: Scalable Multi-query OptimizatioN through Integrated Circuits. 24th International Conference, DEXA 2013 (pp. 278-292). Prague, Czech Republic: Springer.

[2] Andrea, C. F., Danielle, C., Matthew, T. B., Richard, D. H., Chin-Kuo, C., Richard, G. J., . . . Felicity, C. (2013). Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. BMC Medical Informatics and Decision Making .

[3] Andrew, J. M., Britt, F., Guergana, S., Isaac, S. K., & Ben, Y. R. (2013). Improved de-identification of physician notes through integrative modeling of both public and private medical text. BMC Medical Informatics and Decision Making.

[4] BENJAMIN, C. M., KE, W., RUI, C., & PHILIP, S. Y. (2010). Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys.

[5] Buğra, G., & Ling, L. (2008). Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms. IEEE Transactions on Mobile Computing, 1-18.

[6] Cyril, G., & Pierre, Z. (2013). Automatic De-Identification of French Clinical Records: Comparison of Rule-Based and Machine-Learning Approaches. MEDINFO (pp. 476-480). Copenhagen, Denmark: dblp.

[7] Daniel, A., Guillermo, N.-A., & Vicenc¸, T. (2010). Towards Privacy Preserving Information Retrieval through Semantic Microaggregation. ACM International Conference on Web Intelligence and Intelligent Agent Technology (pp. 297-299). IEEE.

[8] James, G., Li, X., Kanwei, L., & James, J. L. (2009). HIDE: Heterogeneous Information DE-identification. the 12th International Conference on Extending Database Technology: Advances in Database Technology (pp. 1116-1119). Saint Petersburg, Russia: ACM.

[9] Jiaqian, Z., Jing, Y., & Junyu, N. (2008). Web User De-Identification in Personalization. the 17th international conference on World Wide Web (pp. 1081-1082). Beijing, China: ACM.

[10] Jordi, S.-C., & Josep, D.-F. (2012). Probabilistic k-Anonymity through Microaggregation and Data Swapping. IEEE World Congress on Computational Intelligence (pp. 1-8). Brisbane, Australia: IEEE.

[11] Karen, T., Julie, K.-G., Tezeta, F. M., Chiriac, M., & Joel, M. (2010). De-identification of primary care electronic medical records free-text data in Ontario, Canada. BMC Medical Informatics and Decision Making .

[12] Khaled, E. E., Fida, K. D., Romeo, I., Elizabeth, J., Daniel, A., Elise, C., . . . Jim, B. (2009). A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. Journal of the American Medical Informatics Association , 670–682.

[13] LATANYA, S. (2002). Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 571-588.

[14] LATANYA, S. (2002). k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 557-570.

[15] Liang, D., & Haibin, L. (2011). Preservative License Plate De-identification for Privacy Protection. International Conference on Document Analysis and Recognition (ICDAR) (pp. 468 - 472). Beijing: IEEE.

[16] Pan, J., Deng, H., Song, Y., & Li, D. (2014). Potential Attacks against k-Anonymity on LBS and Solutions for Defending the Attacks. In Y. J. Hwa, S. O. Mohammad, Y. Y. Neil, J. James, & H. P. Jong, Advances in Computer Science and its Applications (pp. 877-883). Berlin Heidelberg: Springer.

[17] Patrick, S., Hongwei, T., Weining, Z., & Shouhuai, X. (2007). Privacy-Preserving Data Mining through Knowledge Model Sharing. the 1st ACM SIGKDD international conference on Privacy, security, and trust in KDD (pp. 97-115). San Jose, California: Springer.

[18] Ralph, G., Latanya, S., Jeffrey, C., Fernando, d. ,., & Simon, B. (2009). Face De-identification. In S. Dr. Andrew, Protecting Privacy in Video Surveillance (pp. 129-146). London: Springer .

[19] Uzuner, Ö., Yuan, L., & Peter, S. (2007). Evaluating the State-of-the-Art in Automatic De-identification. Journal of the American Medical Informatics Association, 550-563.

[20] Varun, B., Tyrone, G., & Carlos, M. (2012). Recommendation-based De-Identification A Practical Systems Approach towards De-identification of Unstructured Text in Healthcare. Eighth World Congress on Services (pp. 155-162). Honolulu, HI: IEEE.

[21] Weiyi, X., Raymond, H., Xiaofeng, D., Jiuyong, L., & Bradley, M. (2013). Efficient discovery of de-identification policy options through a risk-utility frontier. the third ACM conference on Data and application security and privacy (pp. 59-70). San Antonio, TX, USA: ACM..