Privacy preserving Data Mining in e-government environment

Vugar Musayev
Institute of Information Technology of ANAS
iro@iit.ab.az

Abstract— **Creatinga central data warehouse of public records of E-government in order to apply intelligent data analysis methods for better decision support is very important. We have developed a conceptual framework for such a data mining application scenario. Considering the privacy problems, it is essential to maintain the privacy preserving data mining. We proposed a general overview of a privacy preserving data mining system and we gave very concise and general survey of privacy preserving methods and approaches.**

Keywords— **E-government; data mining; privacy preserving; data warehouse**

I. INTRODUCTION

Digital government formation brings new opportunities for fundamental change in governance and management besides the services and tools it proposes to state, business and citizens. One of the mainstream direction of research is about to emerge on the vast amount of data accumulating in e-gov databases. It is clear that data analysis and data mining methods have been employed for many years in analysis, classification, clasterization, anomaly detection and forecasting. However, only recently, in the beginning of the golden era of big data, it is becoming possible to have deep analysis on strongly related data of huge volumes.

On the other hand, data mining has also negative impression in legislative, social and ethical context. The important question arises that is it possible to save the privacy in data mining process. There might be no satisfactory ultimate solution to the privacy preservation in data mining. However, quite valuable research has created a new direction, namely Privacy Preserving Data Mining (PPDM). PPDM is to ensure privacy by manipulating data while maintaining the quality data mining.

It should be noted that data mining of public records means data mining of personal, mostly private data. However,the ultimate goal is not discovering any hidden private aspects of life of citizens. The main goal is to provide society and the state with information based on analysis of actual data. In many PPDM methods, loss of data is inevitable for the sake of privacy. In our case, it does not mean a loss if we really do not need to keep sensitive part of data open or accessible. Hence, PPDM perfectly fits for our research.

The paper begins with the problem statement in detail and makes a very short introduction to the potential of data mining in the problem domain.It continues with a breif introduction to privacy preserving data mining, related research directions and methods.

II. PERSONAL DATA INFRASTRUCTURE IN E-STATE.

E-state and its corporative e-environments gather the personal data, either biometric or non-biometric in various databases. An important property of personal information databases is that they have different corporative security policies. This is actually a dynamic online environment.Considering that every citizen can be identified uniquely, it is possible to link all the personal data spread over variousdatabases maintained for different e-services of e-government. With this huge system of databases, it is possible to obtain extremely important tools for public and private organizations. Data mining and other intelligent analysis tools together withdecision support systems based on this linked system of personal data will serve for society and efficient management of resources.

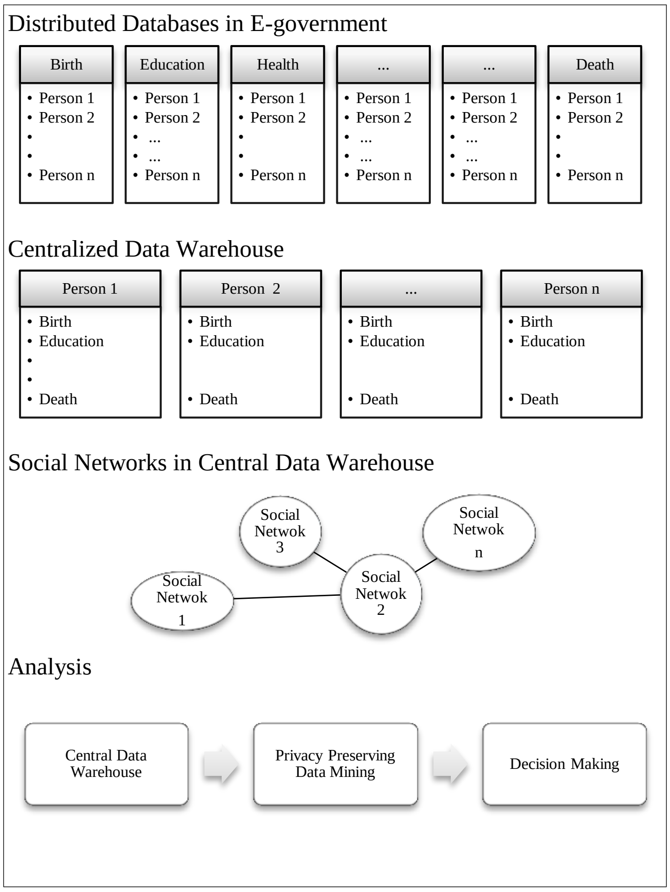


Fig. 1 Privacy preserving data mining in digital government

III. DATA MINING

Data mining (DM), an interdisciplinary branch of computer science, is theanalysis step of the "Knowledge Discovery in Databases" [1]. It is anautomated process of discovering patterns in large data sets byemploying methods of artificialintelligence, machine learning, statistics, and database systems [2]. Mining large amountsof data yields patterns like groups (cluster analysis), unusual set of data (anomaly detection) and relations (association rule mining). Results of DM may be used for further analysis and decision-making.

Considering the data mining activates carried out by governments, companies and research projects, the followings are the main types relevant to our research.

A. Medical data mining

Supreme Court of the United States,permitted the pharmacies to share information with other companies. This decision was based on "freedom of speech" [3].

B. Spatial data mining

Application of DM methods to spatial data to extract patterns with respect to geography has great potential. Employing DM and GISjointly is very important in e-state environment, since most data in public databases have geographic components.

C. Sensor data mining

Sensors are increasingly common in even everyday life.In the context of Internet of Things, sensor data will be essential in public governance. Even a small fraction, e.g. medical purpose sensor data provides live map of health status of a country or any sub region. Besides the simple statistics, intelligent analysis will be much useful on sensor data.

D. Visual data mining

Regardless of the source,huge amount of available data is analyzed visually for extracting patterns. Visual data mining is faster and much more intuitive[4] .

E. Surveillance

Data mining is efficiently used in U.S. government programs, such as the Total Information Awareness (TIA) program; Secure Flight, Analysis, Dissemination, Visualization, Insight, Semantic Enhancement (ADVISE)[5]; and the Multi-state Anti-Terrorism Information Exchange (MATRIX) [6]. These programs were withdrawn on the basis of violationto the 4th Amendment to the United States Constitution. However, some sub programs of them continue [7].

F. Subject-based data mining

"Subject-based data mining" could be considered as the social network analysis of a certain individual or a group. "Subject-based data mining uses an initiating individual or other datum that is considered, based on other information, to be of high interest, and the goal is to determine what other persons or financial transactions or movements, etc., are related to that initiating datum" [8].

IV. PRIVACY PROBLEMS IN DATA MINING

Citizen centric data mining on the public databases requires serious ethical considerations. Issues of privacy, legality, and ethics should be addressed. Otherwise many projects like TIA, will eventually be terminated on the basis of privacy concerns.

One of the steps before the actual data mining, data should be prepared. The preparation step is open to a risk of uncovering private information. For instance, in data preparation step, data aggregation is performed to combine data in order to facilitate the data mining phase. This might cause the unique identification of private, individual data [9].Thensomebody with access to the aggregated data set can easily identify certain individuals even if the data was anonymous.

There is a strong recommendation that an individual should be made aware of the followings before data are collected [9]:

- The purpose of the data collection and any (known) data mining projects,
- How the data will be used,
- Who will be able to mine the data and use the data and their derivatives,
- The status of security surrounding access to the data,
- How collected data can be updated.

V. PRIVACY PRESERVING DATA MINING SYSTEMS

Privacy-preserving data mining has three main stages (Fig. 2). At the first stage, data providers submit their private data to data warehouse. In the second phase, data warehouse server supports online analytical data processingin order to transform the raw data into aggregate data. This helps the next stage, data mining, to process the large amount of data quickly. The essential point here is that, data mining servers have limited access to the data warehouse.

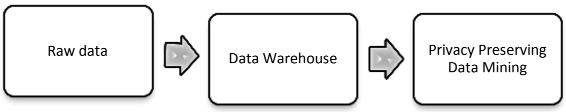
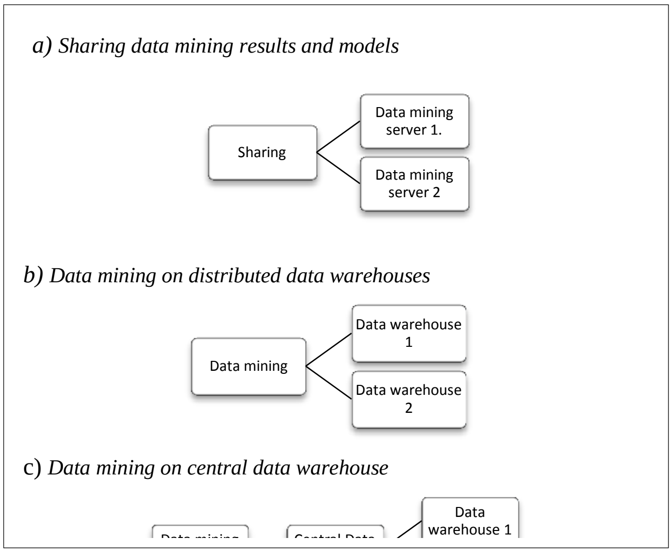


Fig.2 Main stages of privacy-preserving data mining

In order to ensure privacy preserving in e-gov data analysis, we foresee three approaches. In the first one, e-gov data mining system is composed of several data mining servers, and only the data mining models and result are shared among different servers (Fig. 3a). In the second approach, all data mining servers have access to all the data warehouses(Fig. 3b). In the last approach we propose a central data warehouse that collects citizen related data from the other public domains (Fig. 3c).

Each of these approaches have challenging problems in both privacy and efficiency aspects. In each of the proposed approaches privacy preserving must be handled in every step of the process. In public organizations centralized data warehouse combined with privacy preserving data mining systems may even be more trustworthy than decentralized privacy preservation policies.



a) Sharing data mining results and models

b) Data mining on distributed data warehouses

c) *Data mining on central data warehouse*

Fig. 3 Three approaches to data mining on public records.

VI. Privacy preserving data mining methods

Privacy preserving data mining is a recent direction in research and has the following properties and challenges.

*Definition of Anonymity:*Different privacy-preserving tasks have led to several anonymity definitions.For instance, both k-anonymity and andt-closeness methods prevent the true identification in a different way.

*Method of k-anonymity:*Main idea in k-anonymity is to ensure that a certain record cannot be distinguished from at least other (k-1) records [10]. This is an important method to remove the possibility of recovery of unidentified data point from the openparts of the record. For example, address and gender can identify a data point even if the identification is hidden.

Method of Randomization: The randomization method perturbs data to make individual data points unrecoverable. Only aggregate distributions are open for data mining.In *Additive Perturbation*, random noise is added to thedata records. In *Multiplicative Perturbation*, the random projection or randomrotation techniques are used in order to perturb the records [11,12].

Quantification of Privacy: It is important to have a quantitative measure of security fordifferentprivacy-preserving methods. Quantification is established by measuring the risk ofdisclosure for perturbation ata certain level.

Association Rule Mining Privacy: Association rule mining privacy has two aspects. First problem is to extract correct association rules while keeping the privacy. The second problem is to achieve the privacy of the mined association rules which is called associationrule hiding [13].

Query Auditing: Thesemethods modifyor restrict the results of queries in order to preserve privacy in queries [14].

Cryptographic Methods for Privacy: Severaldata mining servers may share private data. In this case cryptographic protocols should be employed for sharing the information among several parties [15].

Personalized privacy-preservation: In personalized privacy-preservation, different records have a different level of privacy [16,17].

High Dimensionality Problem: In most cases, data points have very high dimensions. PPDM algorithms on these type of data sets have the challenge of efficiency and performance.For example, k-anonymization method has NP complexity [18].

Conclusions

E-government environment maintains many public records of citizens. All the records are accumulated over time and there are strong social networks inherent in the separate databases. For instance, an individual has natural social networks in education and military service.

There is a possibility of discovering new social networks unknown before the data mining, or social network analysis. The collection of separate databases can be gathered together into a central data warehouse and data mining methods can be performed on it. We have introduced this new research problem. The related data mining issues such as spatial data mining, subject based data mining, medical data mining, sensor data mining, surveillance are mentioned briefly. We introduced the challenges of privacy concerns in data mining process. We proposed a general overview ofa privacy preserving data mining system and we gave very concise and general survey of privacy preserving methods and approaches.

References

[1] Fayyad, Usama. Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases".

[2] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30.

[3] June 24, 2011. "Pharmaceutical industry: Supreme Court sides with pharmaceutical industry in two decisions - Los Angeles Times". Articles.latimes.com. Retrieved 2012-11-07. Text By David G. Savage "

[4] Keim, Daniel A.; Information Visualization and Visual Data Mining

[5] Government Accountability Office, Data Mining: Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risks, GAO-07-293 (February 2007), Washington, DC

[6] Secure Flight Program report, MSNBC

[7]"Total/Terrorism Information Awareness (TIA): Is It Truly Dead?". Electronic Frontier Foundation (official website). 2003

[8] National Research Council, Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment, Washington, DC: National Academies Press, 2008

[9] Think Before You Dig: Privacy Implications of Data Mining & Aggregation, NASCIO Research Brief, September 2004

[10] Samarati P., Sweeney L. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression.IEEE Symp. on Security and Privacy, 1998.

[11] Agrawal R., Srikant R. Privacy-Preserving Data Mining. ACM SIGMOD Conference, 2000.

[12] Agrawal D. Aggarwal C. C.On the Design and Quantification of Privacy Preserving Data Mining Algorithms.ACM PODS Conference, 2002.

[13] Verykios V. S., Elmagarmid A., Bertino E., Saygin Y.,, Dasseni E.: Association Rule Hiding.IEEE Transactions on Knowledge and Data Engineering, 16(4), 2004.

[14] Blum A., Dwork C., McSherry F., Nissim K.: Practical Privacy: The SuLQ Framework. ACM PODS Conference, 2005.

[15] Pinkas B.: Cryptographic Techniques for Privacy-Preserving Data Mining. ACM SIGKDD Explorations, 4(2), 2002.

[16] Aggarwal C. C., Yu P. S. On Variable Constraints in Privacy Preserving Data Mining.ACM SIAM Data Mining Conference, 2005.

[17] Xiao X., Tao Y..Personalized Privacy Preservation.ACM SIGMOD Conference, 2006.

[18] Meyerson A., Williams R. On the complexity of optimal k-anonymity.ACM PODS Conference, 2004.