A Study on Identifying and Removing Sensitive Attributes in Data Mining

***Kannasani Srinivasa Rao*[1], *M Krishnamurthy*[2]**

[1] Department of Computer Science and Engineering, Hindustan University, India

[2] Department of Computer Science and Engineering, KCG College of Technology, India

E-mail: srinu532@gmail.com, mkrish@kcgcollege.com

Abstract

Data Mining is a technique which extracts useful information and data, like sequential patterns and the trends from the large amount of databases. The space to yourself responsive input data and the output data that is often used for selecting be worthy of defence against exploitation. In this paper we describe work in progress of our research project on how and to what amount authorized and moral rules can be incorporated in data mining algorithms to prevent such exploitation. For that purpose, fact sets in the field of public safety are used, made available by law and honesty departments. The centre is on preventing that selection rules turn out to distinguish particular groups of people in immoral or against the law ways. Important questions are how existing lawful and moral rules and principles can be translated in format understandable for computers and in which way these rules can be used to guide the data mining process. Furthermore, the technical potentials are used as response to plan solid directions and recommendation for formalising *legislation. This will additional clarify how existing moral and lawful standards are to be applied on new technology and, when required, which new ethical and legal principles are to be developed. Opposing to previous attempt to protect space to you in data mining, we will not focus on (a priori) access limiting method concerning input data, but rather focus on (a posterior) accountability and clearness. Instead of limiting access to data, which is more and more hard to put into effect in a world of computerized and interlinked databases and information networks, rather the question how data can and may be used is stressed.*

Keywords: *Data mining, space, exploitation, departments, information networks*

INTRODUCTION

The goal of data mining is to extract useful information, such as sequential patterns and trends, from huge amounts of data [1–3]. In their battle beside offence and violence, many governments are assembling large amounts of data to increase nearby into methods and activities of suspect and possible suspect. This can be very helpful, but regularly at least part of the data on which data mining is applied is secret and isolation sensitive. Examples are medical data, financial data, etc. This raise the question how isolation, mainly of those who are blameless, can be ensure when applying data mining. Furthermore, the results of data mining can direct to choice, stigmatisation and disagreement [4]. False positives and false negatives are often inevitable, resultant in the fact that people are frequently being judge on the base of description that are right for them as group member, but not as persons as such [5]. In the background of open safety, false positives may result in investigate blameless people and false negatives may imply criminal stay on out of scope.

A prior defence may be realised by protective input data and access to input data. However, remove key attribute such as name, address and social security number of the data subject is inadequate to assurance isolation; it is frequently still possible to exclusively identify particular persons or entity from the data, for example by combine different attributes. Since, the results of data mining are often used for choice, a posterior protection is also attractive, in order to make sure that the output of data mining is only used within the forced moral and lawful frameworks. This implies, for example, that data mining results on terror campaign, where data was collected within widespread authority of secret services, cannot be used just like that for robbery or car theft, where data was collected within limited authority of the police.

The aim of the project described in this paper is to investigate to what extent legal and ethical rules can be integrated in data mining algorithms. The focus will be on the security domain. Key questions are: “How can legal and ethical rules and principles be translated in a format understandable for computers?” and “In which way can these rules be used in the data mining process itself?” A typical example of such an ethical and legal principle in this context concerns anti-discrimination. To reduce unjustified discrimination, it is not allowed to treat people differently on the basis of ethnic background or gender. Self-learning data mining algorithms can study models to expect criminal activities of existing and future criminals, based on past data. However, these models may include unfairness of particular groups of people for unfairness in the past that can be found in past datasets or for unequal datasets. To avoid such incident, self-learning algorithms must be ‘made aware’ of active authorized boundaries or policy.

In this paper, we will describe work in progress of our research project on new data mining tools that are non-discriminating. Even though our research is not finished yet, we offer some preliminary results. Furthermore, we think our research approach and issues involved may be interesting for people working in the fields of data mining, information security, discrimination and privacy. First we start with pointing out the legal and ethical issues involved in data mining and profiling, particularly risks concerning discrimination. Then we explain, by discussing some of our previous research results, that simply removing sensitive attributes in databases does not solve (most of) these problems. We continue by explaining another approach is needed and what approach we have chosen in our research. This new method focus on clearness and responsibility, rather than on right to use to data, as we expect data access to be hard to keep in many situation. To more explain this, we compare remove sensitive data from databases to keep away from inequity with removing identifiably from databases to avoid space to you infringements. Next, we discuss the data sets we are using and the (Dutch) lawful structure we are using.

DISCRIMINATION AND OTHER RISKS

The investigate for sequential patterns and associations in data by means of data mining may give overview of huge amount of data, make possible the handling and retrieve of information, and help the examine for immanent organization in nature. More closely related to the goals of exacting users, collection profile resultant from data mining may improve effectiveness (achieving more of the goal) and good organization (achieving the goal more easily). Data mining may be useful in many different areas [6]. In the clash against offence and violence, analyse large amounts of data may help to increase nearby into methods and activities of suspect and possible suspect. In rotate, this may help police and justice departments to deal with this (potential) suspect more effectively for example, by redirect support, people and notice to these particular groups. Focussing on aim group may be very helpful to set priority, but it may also reason error. Because of inaccurate and unfinished data sets, restricted consistency of data mining results and understanding errors, risk profile may not always be totally consistent [7]. As a result, risk profile may contain false positives, i.e., people who are part of the group describe in the risk profile but who do not share the group character as an individual, and false negatives, i.e., people who are not part of the group described in the risk profile even though they comprise the risk the summary try to explain.

The next example may explain this. Take a (fictional) aerodrome danger summary for terrorists consisting of people with a large black challenge, tiring conventional Islamic fashion, and received from Pakistan. When selection for examination would be based on such an outline, it is very likely that most or all people in this group are not terrorists at all. These are the false positives: they are wrongly selected on the basis of their outline. Additionally, there may be terrorists who do not share this character, who will not be notice by the outline. These are the false negatives: they are incorrectly not selected on the basis of their profile. Note that real terrorists would try to keep away from corresponding such a profile, for instance, by flake their challenge, exhausting different clothing, and arriving from different airports.

When selecting persons or group of people on exacting character, this may be not needed or unwarranted or both. Selecting for jobs on the basis of gender, cultural surroundings, etc., is considered immoral and, in many countries, prohibited by (anti-discrimination) law. When risk profiles construct by company, governments or researchers become 'public knowledge', this may also lead to stigmatisation of particular groups. Unfairness and stigmatisation on a large scale may also result in polarisation of (different groups of) society [8].

REMOVING SENSITIVE ATTRIBUTES

It may be not compulsory that removing responsive attribute from databases is the greatest solution to avoid immoral or illegal sharp data mining results. If responsive attributes such as gender, cultural background, spiritual background and sexual preference, unlawful and medical records are deleted from databases, the resultant pattern and relatives cannot be perceptive anymore, it is sometimes argue. However, recent research in our project has shown that this hypothesis is not correct [9]. Classification models frequently make prediction on the basis of training data. If the training data is partial towards definite groups or classes of objects, e.g., there is ethnic unfairness towards black people; the educated model will also show biased behavior towards that particular community. This partial attitude of the educated model may lead to influenced outcome when tagging future unlabeled data objects.

For example, the whole time the years, in a certain administration black people might methodically have been without from jobs. As such, the past employ information of this company regarding job application will be influenced towards giving jobs to white people while denying jobs from black people. Simply remove the sensitive attributes from the training data in the learning of a classifier for the classification of future data objects, but is not enough to solve this problem, because often other attributes will still allow for the recognition of the discriminate community. For example, the traditions of a person might be strongly related with the postal code of his housing area, leading to a classifier with indirect racial biased behavior based on postal code. This effect and its utilization is often referred to as redlining, stem from the practice of deny or increasing services such as, e.g., mortgages or health care to people in certain often culturally strong-minded areas. The term redlining was coined in the late 1960s by centre of population activists in Chicago. The authors also support this claim: even after removing the sensitive attribute from the dataset unfairness persist [10].

METHOD AND APPROACH

When removing sensitive attributes does not prevent unethical or illegal discrimination, other approaches are needed, since impartial classification results are desired or sometimes even required by law for future data objects in spite of having biased training data. In our project, we tackle this problem by introducing a new classification scheme for learning unbiased models on biased training data. Our method is based on massaging the dataset by making the least intrusive modifications which lead to an unbiased dataset. On this modified dataset we then learn a non-discriminating classifier. In order to understand this, the first step is to convert laws and rules into quantitatively quantifiable property against which the exposed models may be checked. Such a formalisation enable verify the correctness of existing laws and rules in the exposed models. We would like to strain here that it is not our goal to build up a complete programmed and automated system for a code of laws. Rather we want to explore how some selected current legislation, e.g., anti-discrimination rules, convert into constraint that can be broken computationally. For example, anti-discrimination rules may convert to the recognition of sensitive data attributes and actual quota. When this is successful, the research can be unlimited to other ethical and legal rules.

The second step is to combine these rules directly in the computerized quest for appropriate models. This may be by modify the original dataset that is learning from or modify the discovery strategy of the algorithms. This provides interesting challenges in the area of data mining research; existing models barely take into account ethical, moral and legal rules. As such, in the current condition the best that can be achieved is learning a model and verify it a posterior, and if the proof fails the model cannot be used [10]. Another problem lies in the computational complication of the confirmation. In the background of pattern mining it is, in general, computationally inflexible to measure what can be derived from the output [11, 12]. As such, it is not possible to guarantee privacy protection later. Recently, several methods have been developed to make sure privacy for data mining, but most of these methods are based on a numerical situation and do not provide severe guarantee are computationally very heavy [13]. In our project we investigate whether anti-discrimination rules can be used as constraint in the model assembly, aiming to remove or ignore a potential bias in the training data in accord of a disadvantaged community.

The third step is to provide feedback on what is efficiently possible, by draft recommendation on how moral and legal rules can be formalised to be useful in a computational background. In summary, our project contains three milestones:

1. Analyse the possibilities to convert laws and rules into a plan reasonable for computers that may be confirmed.
2. Integrating this formalisation in the current state-of-the-art algorithms for discover models.

3. Given that comment of the technical possibilities for real recommendation for formalising legislation.

We start with a dataset contain a sensitive attribute and a class label that is partial with relation to this responsive attribute. For example, the dataset contains information about the occurrence of sure crime types and the sensitive attribute indicated who insert the data. This data is split into training and a confirmation set to avoid that the effectiveness of the approach is tested against the same data that was used for creating the model. For the training set, the unfairness is removed. Two methods will be explored. In the first method we will try to remove unfairness from the training dataset. For this purpose, ranking methods can be used in order to predict the most likely sufferers of unfairness and to adjust their label. Based on the cleaned dataset, conventional classification methods can be employed. Notice that the combination of the unfairness removal step and the traditional model construction on the unbiased dataset can also be seen as a learning method applied directly on unfair data. The advantage of this first approach is that it is independent of the model type being used, whereas the disadvantage clearly is the vital dependence of the overall result to the ability to accurately identify unfairness without the occurrence of training data for this task. A second method is to set in the constraints extremely inside model construction methods. When learning a Naïve Bayes classifier, e.g., we can change the probabilities to reflect "non-discrimination." The advantage here is that this method allows for a better control of the discrimination share in the output models, whereas the disadvantage is that the method is less general than in the first approach. Both approaches will be explored.

TRANSPARENCY AND ACCOUNTABILITY

For the first and the second landmark, ethical and legal research will be used as input for the technological research, for the third landmark, this will be vice versa. Since, in the past, much research has been done on a priori privacy protection (such as privacy enhancing technologies and privacy preserving data mining, the focus will be on posterior protection [13]. The current starting point of a priori privacy protection, based on access limitations on personal information, is increasingly insufficient in a world of automatic and interlinked databases and information networks, in which persons are quickly losing hold on who is using their information and for what purposes, mainly due to the easiness of copying and disseminate information [14]. In most cases, the output is much more applicable, since this is the information that is being used for management. Therefore, it seems useful to balance a priori access controls with a posterior responsibility. Such responsibility requires clearness [15].

Clearness and responsibility focus on the use of data rather than on the access to data. Clearness refers to nearby in the history of collecting, processing and interprets both the raw data and the results of data mining. Responsibility refers to the possibility to check whether the rules of collecting, processing and interpret both the raw data and the results of data mining were adhere to. New design principles for data mining may indicate how data can and may be used. To additional limit the scope of the research field, the focus will be on integrate rules on clearness and responsibility regarding unfairness. In the context of unfairness responsiveness, new constraints can be imposed on the distribution over different population groups of the predictions by a model on future data. This target allocation can be significantly different from the allocation of the training data.

ANONYMITY AND PRIVACY

Most of the classification models deal with all the attributes equally when classifying data objects and are unaware towards the understanding of attributes. When the goal is to avoid or minimize unfairness, it may be useful to found the sensitivity of attributes. However, when the goal is to avoid or minimize privacy infringements, it may be useful to establish the identifiably of attributes. From a technological perspective, identifiably is the possibility to single out a particular person from a group of potential candidates on the basis of a piece of information. Not every characteristic is equally useful for determining a person's identity. Similarly, not every characteristic is equally useful for selecting a person, nor is it equally allowed from an ethical or legal perspective. Here, an interesting corollary can be made between sensitivity and discrimination on the one hand and individuality and privacy on the other hand. In most classification models, all attributes are treated equally. From a normative viewpoint, however, there may be reasons to treat certain attributes with unwillingness. Some data are directly individual (e.g., name, address, and zip code), some data are indirectly individual (e.g., profession, residence), and some data are non-individual or anonymous (e.g., date of birth, gender). However, data from these categories may be combined and subsequently result in higher degrees of individuality.

Similarly, some unfairness sensitive data are directly selective with regard to equal treatment when used for selection purposes (e.g., gender, ethnic background, sexual preferences, political views, union membership, criminal records, and medical records), some data are indirectly selective (e.g., income, social-economic status, zip code, first name, profession and level of education) and some data are not selective (e.g., shoe size, number of kids, and, in most cases, age). Again, data from these categories may be combined and subsequently result in higher degrees of compassion.

Secrecy may help to prevent privacy intrusion, but it may not prevent data mining and profile. Next to techniques that provide (more) secrecy, techniques that limit link ability may be useful [16]. Link ability may be limited by restricting (the type of) access to the data items to be linked. This may be done with the help of multilevel security systems, requiring three types of information controls: access controls, flow controls and inference controls [17]. Access controls handle the direct access to data items. Information flow controls are concerned with the approval to distribute information. Inference controls are used to ensure that released statistics when combined do not disclose confidential data. As mentioned before, our previous research has shown that access controls are less successful with regard to sensitive data and anti-unfairness. However, as described in Section 4, we will investigate the use of flow and inference controls in our approach, as limiting (the influence of) understanding of indirectly selective attributes may provide more protection against selective data mining results.

DATA SETS

The research aims at using data sets from the Dutch Central bureau of Statistics (CBS) and the Dutch Research Centre of the Ministry of Justice (WODC). CBS holds detailed data regarding different demographic parameters of persons, such as living conditions, neighbourhood, country of birth, and country of birth of parents. Combined with data from the department of justice, models were constructed linking demographic parameters to potential criminal behaviour. From this research a strong correlation between ethnicity and criminal behaviour becomes apparent. This correlation, however, is misleading; people do not become criminals because of their ethnicity, but the ethnicity rather is a carrier of other information such as socio-economic position, job, and education. We investigate to what extent our discrimination-aware data mining approach can help answering the following questions: "If we divide the data with relation to the characteristics of which we assume that ethnicity is a carrier, can we identify accurate models within the data selections, i.e., does ethnicity still add accuracy to the model?"

With traditional methods this question is difficult to answer due to the redlining effect; simply removing the ethnicity attribute will not resolve the problem: in general it is to be expected that other demographic attributes in the dataset still allow for an accurate identification of the ethnicity of a person. Another issue that was raised is that "expectedly, some information in police databases is inconsistent because of the subjective nature of some qualifications. E.g., the type of crime is not always straightforward to determine and may, as such, depend on the person who entered the information or on the organisation where the data is entered. Is it possible with the discrimination-aware data mining approach to filter out such effects in model construction?" Both CBS and WODC cooperate with us in this project, since they hope of find answers to these questions.

Concerning the third milestone, it is important to distinguish that the legal and ethical frameworks are also subject to progressing development, due to the rapid technological changes. It is often not clear how ethical and authorized values should be applied to new technologies; in addition, the question can be pose whether new ethical and legal values should be developed for new technologies and its applications. The above mentioned check whether technology complies with the ethical and legal frameworks is therefore only a 'photograph'. Both ethics and law can be further developed with input of existing technological possibilities and realistic developments and applications. From this perspective, the vulnerabilities of those involved can be further explicated and suggestions for change can be made to deal with this. This assessment will not only include vulnerabilities of data subjects, particularly those who are innocent, but also the vulnerabilities of organisations in the security domain, such as police and justice departments, who are in a constant need of intelligence to perform their tasks.

LEGAL ASPECTS

In this Section, we will discuss the Dutch legal situation with regard to profiling and data mining and its consequences regarding discrimination and privacy. First we will discuss the Dutch Equal Treatment Act and then we will discuss the European Personal Data Protection Directive, which was implemented in the Netherlands and other EU member states.

Equal Treatment Act

In the Netherlands, the Algemene Wet Gelijke Behandeling (AWGB, the Equal Treatment Act) clarifies what discrimination is and when this is not allowed. The AWGB may be considered to be based on Article 1 of the Dutch constitution. Article 1 AWGB distinguishes between direct and indirect distinctions. Direct distinctions are distinctions on the basis of religion, philosophy of life, political orientation, race, gender, nationality, sexual orientation, or civil status. These characteristics are not always considered sensitive data in data protection law (see next subsection). Indirect distinctions are distinction on the basis of other characteristics resulting in direct distinctions. The latter guard against redlining. The AWGB does not apply to indirect distinction that may be 'neutrally acceptable'. Other cases in which the AWGB does not apply include, for example, so-called positive unfairness, i.e., making a difference with the purpose of reducing or abolish existing inequalities. Cases, in which distinction is banned, whether direct or indirect, include contribution jobs; dismiss employees, promoting employees, offering products and services, etc. It should be mentioned that the AWGB deals with distinction that actually have a consequence, not with distinctions that may occur.

It could be argued that, when using group profiles, no distinctions are made among group members. However, the AWGB also prohibits the use of some characteristics for decision-making. For instance excluding all Muslims from insurance may be equal treatment within the group, but is generally considered unfair discrimination and, therefore, prohibited. Hence, with regard to the data mining algorithms we are developing, there are two main legal challenges.

The first challenge is to determine whether our data mining results are indirectly discriminating. Direct discrimination is rather straightforward to determine, but indirect discrimination is less clear. For instance, a group defined by zip code that is 90% Indian may result in de facto discrimination, but is it also de facto discrimination when the group is 50% Indian? The second legal challenge is related to fairness, rather than to accuracy. Even when group profiles are completely reliable, we may not consider them fair to use. When particular characteristics are used for selection, treatment and decision-making, do we consider these characteristics fair to use for such purposes?

Personal Data Protection

In Europe, the collection and use of personal data is protected by a European Directive (sometimes referred to as the privacy directive), which has been implemented in national law in the member states of the European Union. Privacy principles that are safeguarded in this directive are [18]:

- The group constraint opinion, state that there should be limits to the group of personal data and any such data should be obtain by legal and fair means and, where appropriate, with the knowledge or permission of the data subject.
- The data quality principle, state that personal data should be applicable to the purposes for which they are to be used and to the degree essential for those purposes, should be correct, complete and kept up to date.
- The purpose specification principle, state that the purpose for which personal data are collected should be particular and that the data may only be used for these purpose.
- The use limitation principle, state that personal data should not be disclose, made available, or else used for purposes other than those particular, except with the permission of the data subject or the power of law.
- The security safeguards principle, state that sensible safety measures should be taken against risk of loss, unauthorized access, demolition, etc. of personal data.
- The openness principle, state that the data subject should be able to know about the survival and nature of

personal data, its purpose and the uniqueness of the data organizer.

- The individual participation principle, state that among others, that the data subject should have the right to have his personal data erase, rectify, complete or amend under sure situations.
- The accountability principle, state that the data organizer should be responsible for comply with measures supporting the above principles.

These privacy principles for fair information practices are based on the concept of personal data, which is defined in article 2 sub a of the directive as: 'data concerning an identified or identifiable natural person'. In other words, this legislation and the principles above are applicable to databases that contain personal data. Although sometimes databases are anonym zed in order to avoid this legal framework, these principles ensure to some extent proper privacy protection, as they address both (a priori) access controls and (a posteriori) transparency and accountability. The European Directive contains a command with increased data protection for special categories of data. Processing personal data informative cultural or cultural source, political opinion spiritual or philosophical beliefs, trade-union membership and data regarding health or sex life is prohibited, if not an explicitly mentioned exception is made in the Directive. Nevertheless, there has been a lot of criticism on the European Directive. It has been described as abstract and general and difficult to enforce. European data protection authorities acknowledge the inadequacy of the current directive. For instance, the British Information Commissioner's Office (ICO) did a review and the European Commission has indicated its need for research on different approaches [19, 20].

CONCLUSION

With regard to the increasing amounts of data that are being collected and processed, data mining is an important technology to extract useful information from data. In the context of fighting crime, these patterns and trends may provide insight and may be based for decision-making and allocation of police forces, funds, etc. However, the use of data mining is controversial because it may negatively affect personal and civil rights, most importantly privacy and non-discrimination. In our research we try to find data mining algorithms in which it is prevented that selection rules turn out to discriminate particular groups of people. Our first research results have shown that simply removing sensitive attributes such as ethnic background from the databases do not solve this problem, since it is to be expected that other demographic attributes in the dataset still allow for an accurate identification of the ethnicity of a person. Therefore, we use a different approach, focusing on massaging the dataset by making the least intrusive modifications which lead to an unbiased dataset. For this, we analyse current anti-discrimination and privacy legislation and use this as input for the data mining algorithms. In this approach we compare discrimination sensitive attributes with privacy-sensitive attributes and try to find ways to limit link ability between directly and indirectly discriminating attributes similar to the existing methods to limit link ability between identifying and non-identifying attributes. The legal part of the research consists of determining when exactly the data mining results are indirectly discriminating and, even when group profiles are completely reliable, of determining which attributes are considered ethically and legally fair to use. The technological possibilities are used as feedback to formulate concrete directives and recommendations for formalising legislation. Instead of limiting access to data, which is increasingly hard to enforce in a world of automated and interlinked databases and information networks, rather the question how data can and may be used is stressed.

REFERENCES

1. Adriaans P., Zantinge D. *Data Mining*. Harlow, England: Addison Wesley Longman; 1996.
2. Fayyad U-M., Piatetsky-Shapiro G., Smyth P., et al. *Advances in knowledge discovery and data mining*. Menlo Park, California: AAAI Press / The MIT Press; 1996.
3. Mannila H., Hand D., Smith P. *Principles of Data Mining*. Cambridge, MA: MIT Press; 2001.
4. Custers B.H.M. *The Power of Knowledge; Ethical, Legal, and Technological Aspects of Data Mining and Group Profiling in Epidemiology*, Tilburg: Wolf Legal Publishers; 2004.
5. Vedder A. H. KDD: The Challenge to Individualism, In: Ethics and Information Technology. 1999; 1: 275–281p.
6. Meij J. *Dealing with the data flood; mining data, text and multimed*ia, The Hague: STT Netherlands Study Centre for Technology Trends; 2002.

7. Custers B.H.M. *Effects of Unreliable Group Profiling by Means of Data Mining*. In: G. Grieser, Y. Tanaka and A. Yamamoto (eds.) Lecture Notes in Artificial Intelligence, Proceedings of the 6th International Conference on Discovery Science (DS 2003) Sapporo. New York: Springer-Verlag. 2003; 2843: 290–295p.
8. Custers B.H.M. *The Risks of Epidemiological Data Mining*, In: Herman Tavani (Edn.) Ethics, Computing and Genomics: Moral Controversies in Computational Genomics. Boston: Jones and Bartlett Publishers, Inc; 2005.
9. Kamiran F., Calders T. *Classification without Discrimination*. In IEEE International Conference on Computer, Control & Communication (IEEE-IC4). IEEE press, 2009. Accepted for publication; 2009.
10. Pedreschi D., Ruggieri R., Turini F. *Discrimination-aware Data Mining*. In Proceedings of the 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2008.
11. Calders T. The Complexity of Satisfying Constraints on Transaction Databases. *Acta Informatica*. 2007; 44(7–8): 591–624p.
12. Calders T. Itemset Frequency Satisfiability: Complexity and Axiomatization. *Theoretical Computer Science*. 2008; 394(1–2): 84 –111p.
13. Lindell Y., Pinkas B. Privacy preserving data mining, *Journal of Cryptology*. 2002; 15: 177– 206p.
14. Westin A. *Privacy and Freedom*. London: Bodley Head; 1967.
15. Weitzner D.J., Abelson H., et al. *Transparent Accountable Data Mining: New Strategies for Privacy Protection*. MIT Technical Report, Cambridge: MIT; 2006.
16. Goldberg I.A. *A pseudonymous Communications Infrastructure for the Internet*, dissertation, Berkeley: University of California at Berkeley; 2000.
17. Denning D.E. *Cryptography and Data Security*. Amsterdam: Addison-Wesley; 1983.
18. Bygrave L.A. *Data protection law; approaching its rationale, logic and limits*. Information law series 10, The Hague, London, New York: Kluwer Law International; 2002.
19. Robinson N., Graux H., Botterman M., et al. *Review of the European Data Protection Directive*. Cambridge: RAND Europe; 2009.

20. Contract Notice 2008/S87-117940 regarding the Comparative study on different approaches to new privacy challenges, in particular in the light if technological developments, published on http://ted.europa.eu.

AUTHORS PROFILE

Kannasani Srinivasa Rao: Pursuing Ph.D. CSE from Hindustan University, Chennai. He finished his B.Tech. CSE - at RGM College of Engineering & Technology, Nandyal- AP. He finished his M.Tech. CSE - at JNTU Ananthapur. Area of Interest: Data Mining.

Dr. M Krishnamurthy is currently working as a Professor and Head of M.E (CSE) in the Department of Computer Science and Engineering, KCG College of Technology. He leads a research group of Computer Science and Engineering faculty in KCG. He has published more than 14 papers in refereed international conferences and journals. His primary work and research interests include large database mining techniques, incremental data mining techniques, and social network analysis and mining.