www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 12 December 2017, Page No. 22312-22314
Index Copernicus Value (2015): 58.10, 76.25 (2016) DOI: 10.18535/ijecs/v6i12.11

Big Data Mining approach to E-governance: Mining Census data for enhanced citizen centric policy making

Smita Khot [1], Pranav Ghildiyal[2], Prasanna Dubey[3], Mitali Patil[4], Pooja Matere[5]

[1]Professor, Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune
[2,3,4,5]Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune

Abstract: ***With Internet and technology making exponential rise in its presence in recent years, the role of public has increased in governments across the states or country and the people are becoming more aware about policies being made by the government and expressing their views about them, suggesting improvement and expressing happiness about good policies. When it comes to policy making, it is influenced by multiple factors, the policies must be made keeping the masses in mind, and assessing its impact on the masses, a policy can be termed as being a good one, if and only if it is benefiting the masses. The process of policy making will become easy and the policies would turn out to be of more benefit and help the masses if it they are made after taking into consideration the needs and requirements, which can be done if we have enough data to analyze which is in sufficient magnitude to reflect the current status of the citizens of the country. It's not a challenge to get the data, but it would be huge in size and challenge will be in analyzing it. As humans can't analyze such amounts of data, Data Mining comes to rescue here, the data can be collected and mined on various parameters to understand the current state, the citizens of the country are and hence policies can be made which would directly reflect in nations progress.***

Keywords: Census Data Mining, Big Data, Knowledge Discovery, Privacy-Preserving Data Mining, Map Reduce

1. Introduction

Census data is the data of citizens of the country which as a whole can reflect the state of a nation, it's progress, the needs and requirement, which when used by the government while deciding new policies to be made, or while improving the currently existing ones will give better results as the states of the whole country is being taken into consideration, but this data is going to be huge, it's not going to be just data, it's going to be what we call big data and it will not be possible for government officials to process and analyze this data without any support of some software system. We very much need a software architecture which is capable of handling, processing such big amounts of big data. The data is primarily going to be unstructured hence, structured database would make the system inefficient, so the data will be stored in unstructured i.e., flexible schema supporting database like MongoDB. Interface to access, process and analyze this data will be through java based application. As a stand-alone application is going to the main interface between the data and analysis or access, hence it will help maintain secureness and confidentiality as only the decided system and designated official(s) will be able to access this data, which is major requirement given the nature and origin of data, hence the privacy of such Census data will be maintained.

2. Problem Statement

The major problem that today's government face while drafting new policies or improving existing ones is the lack of ability to take into consideration the current complete state of people, so to eliminate this shortcoming of analysis we can perform data mining on the census data collected by the Government officials, this system will help the government in making better citizen oriented policies for the development and progress of nation.

3. Literature Review

Census can provide the fundamental population data of the whole nation. The census data is rich with hidden knowledge which can be used to assess the current condition of the nation and its citizens. Most of the national policies are constituted straightly based on the population status, hence, this huge amount of data can be used to develop effective policies for the citizens of the country, to improve their life and hence it turn speed up the nations progress i.e., to provide services for country's social and economic development. According to them classification, is one of the important Data Mining techniques, when it comes to Census or Government Data Mining [1].

Data mining can extract implicit, previously unknown, and potentially useful information from the data. They say that Census is a significant investigation of national condition and national power, strongly believe that Data Mining in census data has very high learning value. When it comes to Mining Big Data, it might not be always possible to mine detailed data, hence though details is lost by generalization, but the generalized data may be more meaningful and easier to interpret. Hence by generalizing the Big Data, concept hierarchies can be made, which can be mined to extract useful information [2].

They highlight the impact of big data on current scene of the society. They say that the term big data occurs more frequently now than ever before. A large number of fields and subjects, ranging from everyday life to traditional research fields involve big data problems. Big data incorporates endless amount of information. In many industries, it is growing, providing a means to improve and streamline business. Big data has changed the world in terms of predicting customer behavior. The actual challenge of big data is not in collecting it, but in managing it as well as making sense of it. While working on big data, it is crucial to determine whether the benefits outweigh the costs of storage and maintenance. They also

review recent research in data types, storage models, privacy, data security, analysis methods, and applications related to network big data [3].

Big Data concerns with large-volume, complex, growing data sets with multiple, autonomous sources. They believe that big data can prove out to be very useful, only if we can harness the data, to extract the hidden knowledge that it contains. They propose a theorem by the name of HACE theorem (Big Data starts with large-volume, **H**eterogeneous, **A**utonomous sources with distributed and decentralized control, and seeks to explore **C**omplex and **E**volving relationships among data) [4].

According to them, the growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. An emerging topic in data mining, known as privacy preserving data mining (PPDM), has been extensively studied in recent years. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. Current studies of PPDM mainly focus on how to reduce the privacy risk brought by data mining operations, while in fact, unwanted disclosure of sensitive information may also happen in the process of data collecting, data publishing, and information (i.e., the data mining results) delivering. They identify four different types of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker. And discuss privacy concerns and the methods that can be adopted to protect sensitive information for each type of user [5].

Taking reference of Educational Data Mining, they tell us that though Data Mining of Educational data provides new insights for a better education system. However, sharing or analysis of educational data introduces privacy risks for the data subjects, mostly students, this holds true, that too in even greater proportion, when it comes to census data and mining of census data. Many initiatives and regulations protect personal data privacy in domains such as health, commerce, communications and most regulations do not enforce absolute confidentiality which would cause more harm than good but rather protect individually identifiable data that can be traced back to an individual with or without external knowledge. This gives rise to a wide range of studies primarily focusing on de-identifying private data with as little harm to its information content as possible, in an attempt to preserve both the privacy and usefulness of the data, should be considered with respect to various scenarios. Research on data privacy has formally defined and enforced privacy primarily in two scenarios: (1) Sharing data with third parties without violating the privacy of those individuals whose (potentially) sensitive information is in the data. This is often called privacy-preserving data publishing. (2) Mining data without abusing the individually identifiable and sensitive information within. This is often called privacy-preserving data mining or disclosure control [6].

Present us with Incremental Association rule mining approach. They tell that lot of study has been done in the context of preserving privacy of raw data and sensitive data, but has focused on one time mining. As per them, No work till now, has been published on incrementally mining association rules with privacy protection when the data upon which mining occurs is quantitative and subject to change. They study this problem against the backdrop of supply chain management. They, taking reference of this supply chain management, present uses of Discrete Wavelet Trans- form (DWT) to mask the original data in such a way that a majority of the original association rules are preserved [7].

According to them, it might not be always possible to store Big Data at a single place. Hence the data storage has to be distributed. Privacy being another crucial factor along with security restricts sharing or centralization of data. Privacy-preserving data mining has emerged as an effective method to solve this problem. Though distributed solutions have been proposed that can preserve privacy while still enabling data mining. However, while perturbation based solutions do not provide stringent privacy, cryptographic solutions are too inefficient and infeasible to enable truly large scale analytics to face the era of big data. Previous work on random decision trees (RDT) show that it is possible to generate equivalent and accurate models with much smaller cost, which can be exploited the fact that RDTs can naturally fit into a parallel and fully distributed architecture, and develop protocols to implement privacy-preserving RDTs that enable general and efficient distributed privacy-preserving knowledge discovery [8].

Propose Pincer-Search as an efficient Algorithm for Discovering the Maximum Frequent Set, from data. Knowledge Discovery of data can be defined as the non trivial process of identifying valid, potentially useful, and ultimately understandable patterns in data. Mining of frequent Set can discover trends hidden in the data. They tell that typical algorithm for finding the frequent set, operate in bottom-up, breadth-first fashion. The computation starts from frequent 1-itemsets at the bottom and then extends one level up in every pass until all maximal (length) frequent item sets are discovered. The efficiency of such algorithms decreases when any of the maximal frequent set becomes longer. In data mining applications the maximum frequent item sets could be long. To solve this problem, they present a novel Pincer-Search algorithm, which searches for Maximum Frequent Set from both bottom-up and top-down directions. They say that it performs well even when the maximal frequent item sets are long [9].

As new data and updates are constantly arriving, the results of data mining applications become stale and obsolete over time. Incremental processing is a promising approach to refreshing mining results. Given the size of the input big data, it is often very expensive to rerun the entire computation from scratch. Incremental mining utilizes previously saved states to avoid the expense of recomputations from scratch. They propose i2MapReduce, as a novel incremental processing extension to Map Reduce, the most widely used framework for mining big data [10].

4. Mathematical Model

Let S be the system.
S= {- - - -}
Identify I as input.
I= {a, b, c, d}
Where
a -> Census Data
b -> Current Job
c -> Salary
d -> Mobile Number
S = {I}
Identify O as output.
O= {i}
Where

i -> Pictorial Representation of requested information in the form of Bar Graphs, Pie Charts etc as applicable
S = {I, O}
Identify A as case of success.
A= {j}
j -> Requested information is represented pictorially in the form of Bar Graphs, Pie Charts etc, and mathematical figures.
S = {I, O, A}
Identify F as case of failure.
F= {k, l}
Where
k -> Improper data preprocessing
l -> Data connection error.
S = {I, O, A, F}

5. Proposed System Overview

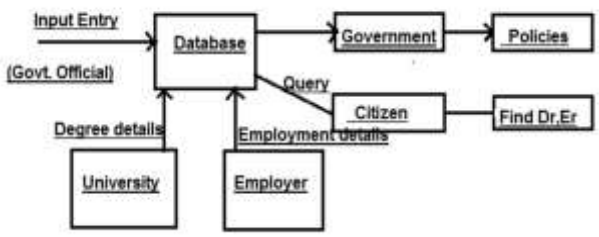

Figure 1: Proposed System Overview

Figure 1 shows the overview of the proposed system. The data that will be stored in the database will come from mainly three sources, these are I. Designated Government census officials, II. University and III. Employer. As shown in the figure the designated government census officials, will be the primary collectors of data. The university and employers which will have the unique identification details of the students and employees studying, working with them will be (responsible for) inserting the respective education degree (completed), job details for the students and employees with them using unique identification number as and when. The data stored in the database will be utilized in two ways first, by the government which can make use of this data to get updated knowledge about the state of the citizens of the country which than further be used to improve existing policies and to make new policies. Second, the citizen through web interface (limited access) can query this data to find nearby doctor, engineer or other professionals as per their need. Due to university and employers entering the details, the details that will be displayed when a citizen fires a query to find the practitioner, he can be ensured that he is getting verified and correct details. The university and employers will be responsible for the authentication, validation and correctness of all information or data that they are inserting into the database.

6. Advantages

The major advantage of the proposed system is that it will help government in Informed policy making, i.e., the government will be able to improve existing or create new policies taking into consideration the states of people. It will also help in Fake degree prevention, as the educational data available in the system will be entered by the respective universities. Also in a similar way fake job profiles can be prevented, as the employment details will be entered by the respective employers.

7. Application

The system can be used by the Government for estimating the current state of the people for better citizen centric policy making. The system will also help the citizens avail basic services like help of Doctor, Engineers and other professional when required.

8. Conclusion

With advances in technology, public awareness and for the betterment of nation's progress and speedy development, taking into consideration the nation's state is very important. The proposed system does exactly this, it gives government the facility to take into consideration the whole sate status of the nation by such census data mining. Additionally, the proposed system serves multiple good purposes like, preventing fake degrees, fake job profiles, access services of doctors, engineers and other professionals.

References

1. Bing Sheng and Sun Ghengxin, Data Mining in Census Data with CART, 3rd International conference on Advanced Computer Theory and Engineering (ICACTE), pp. V3-260- V-264, 2013
2. Sheng Bin and Gengxin Sun, The preprocessing in census with Concept Hierarchy, 2nd International Conference on Computer Engineering and Technology, Volume 1, pp V1-535 V1-538, 2010
3. Zhihan Lv, Houbing Song, Pablo Basanta-Val, Anthony Steed, Minho Jo, Next Generation of Big Data Analytics : State of Art, Challenges, and Future Research Topics, IEEE Transactions on Industrial Informatics, Volume 13, Issue 4, pp 1891-1899, 2014
4. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, Data Mining with Big Data, IEEE Transactions on knowledge and data Engineering, Volume 26, Issue 1,pp 97-107, January 2014
5. Lei Xu, Chunxiao Jhang, Jian Wang, Han Yung, and Yong Ren, Information Security in Big Data: Privacy and Data Mining, IEEE Access, 2014
6. Mehmet Emre Gursoy, Ali Inan, Mehmet Ercan Nergiz and Yucel Saygin, Privacy Preserving Learning Analytics : Challenges and Techniques, IEEE Transactions on Learning technologies, Volume 10, Issue 1, pp 68-81, 2017
7. Madhu V Ahluwalia, Aryya Gangopadhyay, Zhlyuan Chen and Yelena Yesha, Target Based Privacy Preserving and Incremental Association Rule Mining, IEEE Transactions on Services Computing, Volume 10, Issue 4, pp 633-645, 2017.
8. Jaideep Vaidya and Wei Fan, A random decision tree framework for privacy preserving Data Mining, IEEE transactions on dependable and secure computing, Volume 11, No. 5, pp 399-411, Sept/Oct 2014
9. Dao-l Lin and Zvi M. Kedem, Pincer-Search: An efficient Algorithm for discovering the maximum frequent set, IEEE Transactions on Knowledge and data engineering, Volume 14, No 3, pp 553-566, MAY/JUNE 2002
10. Yanfeng Zhang, Shimin Chen, Qlang Wang and Ge Yu, i2 map reduce: Incremental Map Reduce for Mining Evolving Big Data, IEEE Transactions on knowledge and data engineering, Volume 27, issue 7, pp 1906-1919, 2015