Name: Haomin He
Due: 03/22/2018 11:59:59 pm
Instructor: Niru Bulusu

Internet Privacy and Tracking - Data Mining

Introduction:
In recent years, the practice of data mining has become a popular computing technology. Data mining is a means to summarize useful information by analyzing big sets of data. Some newborn tech companies, such as Ditto Labs Inc., use software to scan publicly posted photos. For example, they might look for images of individuals holding a Dr. Pepper drink to determine what logos are in the picture, what facial expressions present, like whether the person is smiling, and what the scene's context is (Dwoskin & MacMillan, 2014). Artificial intelligence, statistical computation and logistic regression are the basic algorithm is tools for data mining, which make it possible to not only rely on numerical data sets, but also other types of datasets, such as text and photo. Data mining plays an important role in analyzing customer information and helps companies relate to their consumers better. However, on the other hand, people concerned about their personal information may be invaded, analyzed and used unknowingly (Christiansen, 2011). In this paper, the details of data mining will be introduced, implemented, and how it influences people's daily life in both good ways and bad.

Background:
Data mining is a young, important, and increasingly popular field, with the first paper appearing only around 1992. Since then, database researchers have started working on huge amounts of data and scalable algorithm computations. Now data mining can be apply almost anywhere. The website *Amazon.com*, provides purchase recommendations for each user by using collaborative filtering algorithm, they say "People buying this book also buy other books" (Winslett & Braganholo, 2012).

In the structured data, people are usually looking for different patterns and interpret the hidden meaning from the numbers, whether to find clusters, regression or evolution. The methods need raw data to support basic calculations, and personal data is sourced from both public databases and private sectors, this is where the ethics of data mining comes into the place (Winslett & Braganholo, 2012).

Implementation:
Sainani performed a calculation example of logistic regression, one of the most important data mining methods in order to show how it works with raw data and delivers insightful information. In this example, we want to know whether a customer accepted a personal loan or not based on other factors, like where they live, whether they have a credit card and their family size. The outcome of logistic regression has only two levels. Either the customer accepts the personal loan or rejects to purchase personal loan (Sainani, 2014). There are unique IDs for each entry. And there are features in customers, we call them variables: age, experience on taking loans, income level, ZIP code, family size, CCAvg (a

statistical measurement number), education level, took personal loan before or not (use numerical number 1 or 0 to represent this concept), have securities account, CD account, and online account or not, own a credit card or not. If the variable type is about 'Yes' or 'No', binary number is used. 1 means 'Yes, the customer do have this feature, 0 means 'No, this customer don't have this feature'.

ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard
1	25	1	49	91107	4	1.60	1	0	0	1	0	0	0
2	45	19	34	90089	3	1.50	1	0	0	1	0	0	0
3	39	15	11	94720	1	1.00	1	0	0	0	0	0	0
4	35	9	100	94112	1	2.70	2	0	0	0	0	0	0
5	35	8	45	91330	4	1.00	2	0	0	0	0	0	1
6	37	13	29	92121	4	0.40	2	155	0	0	0	1	0
7	53	27	72	91711	2	1.50	2	0	0	0	0	1	0
8	50	24	22	93943	1	0.30	3	0	0	0	0	0	1
9	35	10	81	90089	3	0.60	2	104	0	0	0	1	0
10	34	9	180	93023	1	8.90	3	0	1	0	0	0	0
11	65	39	105	94710	4	2.40	3	0	0	0	0	0	0
12	29	5	45	90277	3	0.10	2	0	0	0	0	1	0
13	48	23	114	93106	2	3.80	3	0	0	1	0	0	0
14	59	32	40	94920	4	2.50	2	0	0	0	0	1	0
15	67	41	112	91741	1	2.00	1	0	0	1	0	0	0
16	60	30	22	95054	1	1.50	3	0	0	0	0	1	1
17	38	14	130	95010	4	4.70	3	134	1	0	0	0	0
18	42	18	81	94305	4	2.40	1	0	0	0	0	0	0
19	46	21	193	91604	2	8.10	3	0	1	0	0	0	0
20	55	28	21	94720	1	0.50	2	0	0	1	0	0	1

Then we do cross validation calculation in order to get the coefficients of each variables (There are built-in formulas in Excel already. People just need to learn how to use them). We plug in the coefficients into logistic regression formula, where we can get the probability of accepting the personal loan. As the example showing below:
<*p* stands for probability of getting a 'Yes, customers have personal loans'. *p*-head is where we plug in the calculated coefficients. MLE is Maximum Likelihood Estimation, which finds the estimation that maximizes the chance of obtaining the data we have. >

$$p = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2))}$$

$$\text{MLE: } \hat{\beta}_0 = -6.462,\ \hat{\beta}_1 = 0.038 \text{ and } \hat{\beta}_2 = 2.343$$

$$\hat{p} = \frac{1}{1 + \exp(6.462 - 0.038 x_1 - 2.343 x_2)}$$

The bigger number of probability (p-head) means variables X1 and X2 have stronger influence on accepting personal loan compare to other variable factors. Let's assume X1 represents income level and X2 represents experience. Marketers can target and promote their product specifically towards this market segmentation who has higher income and has experience on loans. This calculation seems a little bit complicated, but from this, companies can find insightful information from the raw data and deliver better products/services to the targeted consumers.

Positive Side of Data Mining:
In today's technological society, many fraud activities happened due to advanced technology. One typical fraud is the ATM phone scams that attract victims to transfer their money into fraudulent accounts. Thanks to data mining technology, by applying Bayesian

Classification and Association Rule, it is possible to identify and detect fraudulent accounts, in order to reduce victims' losses (Li, Yen, Lu & Wang 2012).

Negative Side of Data Mining:

We have to admit that data mining tools do make our lives easier, but sometimes the tools reveal too much about our personal information which make us feel unsafe, uncomfortable and creepy. For instance, you just searched some books on Amazon.com and you closed all the tabs. You opened your Facebook page. Surprisingly, the books you just searched appeared on your Facebook feed. You probably had the feeling like "What?! How do you know?!" Actually this is because Facebook page reads your cookies and displays related advertisement. This is also called Web Personalization or Web Mining. Web mining is a concept that gathers all techniques, methods and algorithms used to extract information and knowledge from data originating on the web (web data). A part of this technique aims to analyze the behavior of users in order to continuously improve both the structure and content of visited web sites (Velásquez, 2013). This technique may help the user feel comfortable when they visit a site through a personalization process. However, to some extent, the website may infringe the privacy of those who visit it.

Another disadvantage of data mining is the accuracy of data gathered. If the data gathered is not accurate enough, marketers are more likely to implement wrong business strategies and leads to business losses. As research shows, some participants in online data collection applications are distrustful and unreliable to the data collector. The reason why the respondents refuse to provide truthful data is because they are in fear of personal information leakage and collusion attacks. In order to get relatively accurate data, companies need to employ cryptographic and random shuffling techniques to preserve data accuracy (Ashrafi & Ng, 2009).

Other Possible Ways:

There are alternative authentic ways to get to know consumers. All business must serve its customers with the best they can. In order to serve well, it must understand their need and desires. Since data mining methods are quite controversial, we can use methods like customer storytelling, market research, customer feedback and social media. Companies should listen to customer stories and learn from them. By truly caring about storytelling, business leaders can better serve their customers and their companies (Gorry & Westbrook, 2011).

Privacy concerns have become an important issue in data mining. A popular way to preserve privacy is to randomize the dataset to be mined in a systematic way and mine the randomized dataset instead (Xia, Yang & Chi, 2004). Since different people see privacy differently, the way they evaluate what information should be private varies (depends on people). As a result, a new method is invented, called the privacy preserving association rule mining algorithms, which allows different attributes to have different levels of privacy that is, using different randomization factors for values of different attributes in the randomization process (Xia, Yang & Chi, 2004).

Avoid Personal Information Leakage From User Side and Company Side:

Modern technology platforms offer a multitude of useful features to their users but at the same time they are highly privacy affecting (Bal, Rannenberg & Hong, 2015). Internet is a complicated place, people should learn how to protect themselves and be able to tell what is good, what is bad and make good choices accordingly. As research found, personal data are oftentimes gathered in exchange for rewards, such as free gifts and access to certain content (Park, Campbell & Kwak, 2012). Reward-seeking is salient in decision making. People need to be aware of the situation that they provide personal information and get something in return. In addition, people need to learn new protection technology on the market in order to prevent information leakage. For example, privacy-preserving implicit authentication system is a new invention that achieves implicit authentication without revealing information about the usage profiles of the users (Shahandashti, Safavi-Naini, & Safa, 2015).

The prevention of personal data leakage can also be done by the company side while they are gathering data. We call this method as privacy-by-design paradigm that is to address the emerging and growing threats to online privacy. The main idea is to inscribe the privacy protection into the design of information technologies from the very start. This paradigm represents a significant innovation with respect to the traditional approaches of privacy protection because it requires a significant shift from a reactive model to proactive one. In other words, the idea is preventing privacy issues instead of remedying to them (Monreale, Rinzivillo, Pratesi, Giannotti & Pedreschi, 2014).

Here is a real example of how healthcare facilities solve the dilemma of sharing patients' personal information. Electronic health records (EHRs) have been widely adapted at many healthcare facilities because they can improve the quality of patient care and increase the productivity and efficiency of healthcare delivery (Li, Bai & Reddy, 2015). It would be convenient if healthcare centers can share patients' information, but personal data leakage is a major concern here. To handle such disparate goals, researchers develop two adaptive distributed privacy-preserving algorithms based on a distributed ensemble strategy. The basic idea of the approach is to build an elegant model for each participating facility to accurately learn the data distribution, and then transfer the useful healthcare knowledge acquired on their data from these participators in the form of their own decision models without revealing and sharing the patient-level sensitive data, thus protecting patient privacy(Li, Bai & Reddy, 2015).

Conclusion:

As you can see, data mining is an indispensable analytical tool for answering both big and small business questions, and raw data is a necessary foundation for data mining implementation. On the contrary, this technology arouses privacy concerns as social networks are booming. People don't want to share their personal information without knowing it, they don't want their personal information to get misused or revealed (Wu, Huang, Yen, & Popova, 2012). But actually every company is keeping records of their customers. Also, information leaks are a harmful threat to people's normal lives. In recent years, there have been many personal information leak cases happened. Gradually, people are unwilling to provide personal information, which is a huge drawback to commerce, since companies don't have sufficient data to analyze and target their consumers. In

addition, there are possible ways to reduce risks of privacy invasion from both user side and company side, but extra methods and efforts are required. As a result, the utilization of data mining still stays controversial.

References

1. Christiansen, Linda. "Personal Privacy and Internet Marketing: An Impossible Conflict or a Marriage Made in Heaven?" Business Horizons 54.6 (2011): 509-14. Web.

2. Dwoskin, Elizabeth, and MacMillan, Douglas. "Smile! Marketing Firms Are Mining Your Selfies." WSJ. Web. 19 May 2014.

3. Gorry, G. Anthony, and Robert A. Westbrook. "Can You Hear Me Now? Learning from Customer Stories." Business Horizons 54.6 (2011): 575-84. Web.

4. Li, Shing-Han, David C. Yen, Wen-Hui Lu, and Chiang Wang. "Identifying the Signs of Fraudulent Accounts Using Data Mining Techniques." Computers in Human Behavior 28.3 (2012): 1002-013. Web.

5. Park, Yong Jin, Scott W. Campbell, and Nojin Kwak. "Affect, Cognition and Reward: Predictors of Privacy Protection Online." Computers in Human Behavior 28.3 (2012): 1019-027. Web.

6. Sainani, Kristin L. "Logistic Regression." Pm&r 6.12 (2014): 1157-162. Web.

7. Winslett, Marianne, and Braganholo, Vanessa. "Jiawei Han Speaks out On Data Mining, Privacy Issues and Managing Students." ACM SIGMOD Record SIGMOD Rec. 40.4 (2012): 28. Web.

8. Wu, Kuang-Wen, Shaio Yan Huang, David C. Yen, and Irina Popova. "The Effect of Online Privacy Policy on Consumer Privacy Concern and Trust." Computers in Human Behavior 28.3 (2012): 889-97. Web.

9. Ashrafi, Mafruz Zaman, and See Kiong Ng. Collusion-resistant Anonymous Data Collection Method. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '09 (2009): n. pag. Web.

10. Bal, Gökhan, Kai Rannenberg, and Jason I. Hong. Styx: Privacy Risk Communication for the Android Smartphone Platform Based on Apps' Data-access Behavior Patterns. Computers & Security 53 (2015): 187-202. Web.

11. Li, Yan, Changxin Bai, and Chandan K. Reddy. A Distributed Ensemble Approach for Mining Healthcare Data under Privacy Constraints. Information Sciences 330 (2015): 245-59. Web.

12. Monreale, Anna, Salvatore Rinzivillo, Francesca Pratesi, Fosca Giannotti, and Dino Pedreschi. Privacy-by-design in Big Data Analytics and Social Mining. EPJ Data Sci. EPJ Data Science 3.1 (2014): n. pag. Web.

13. Shahandashti, Siamak F., Reihaneh Safavi-Naini, and Nashad Ahmed Safa. Reconciling User Privacy and Implicit Authentication for Mobile Devices. Computers & Security 53 (2015): 215-33. Web.

14. Velásquez, Juan D. Web Mining and Privacy Concerns: Some Important Legal Issues to Be Consider before Applying Any Data and Information Extraction Technique in

Web-based Environments. Expert Systems with Applications 40.13 (2013): 5228-239. Web.

15. Xia, Yi, Yirong Yang, and Yun Chi. Mining Association Rules with Non-uniform Privacy Concerns. Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery - DMKD '04 (2004): n. pag. Web.