Data Privacy
CS 510 Mid Term Project Report

Yiming Zhang
Computer Science
Portland State University
Portland, Oregon, USA
ymzhang@pdx.edu

ABSTRACT

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 [1]. The growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. In this project, the details of data mining will be introduced, and how it influences people's daily life in both good ways and bad in terms of data privacy. In addition, we will look into how big companies like Google, Amazon and Facebook collect and use your data. Last but not least, a case study of COVID-19's impact on data privacy, protection and security will be examined.

CCS CONCEPTS

- Data mining
- Security and Privacy
- Sensitive Information
- Anonymization

KEYWORDS

Data mining, web application, data, mobile, privacy, security, COVID-19

ACM Reference format:

Yiming Zhang. 2020. Data Privacy. In *Proceedings of CS 410/510: Explorations of Data Science. Portland, OR, USA, 9 pages.*

1 Introduction

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 applied almost anywhere [2].

Amazon.com provides purchase recommendations for each user by using collaborative filtering algorithms, they say "People buying this book also buy other books" [3]. Some newborn tech companies, such as Ditto Labs Inc., use software to scan publicly posted photos [4]. 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 are present, like whether the person is smiling, and what the scene's context is [5].

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 [6]. 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.

Artificial intelligence, statistical computation and logistic regression are the basic algorithms for data mining [7], 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 [8]. However, on the other hand, people concerned about their personal information may be invaded, analyzed and used unknowingly [9].

2 Process of Data Mining

The term data mining is often known as another term: knowledge discovery from data (KDD) which indicates the

goal of the mining process [10]. To obtain useful knowledge from data, the following steps are performed in an iterative way. (See Figure 1)

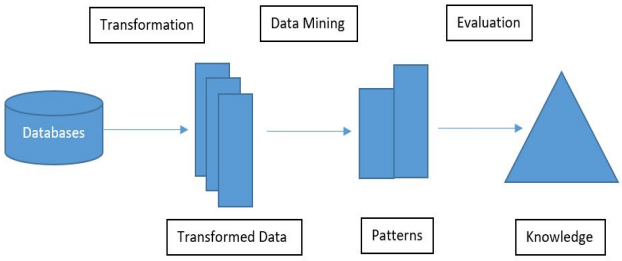

Figure 1: Overview of Data Mining process

Step 1: Data preprocessing. Basic operations include data selection (to retrieve data relevant to the KDD task from the database), data cleaning (to remove noise and inconsistent data, to handle the missing data elds, etc.) and data integration (to combine data from multiple sources).

Step 2: Data transformation. The goal is to transform data into forms appropriate for the mining task, that is, to find useful features to represent the data. Feature selection and feature transformation are basic operations.

Step 3: Data mining. This is an essential process where intelligent methods are employed to extract data patterns (e.g. association rules, clusters, classification rules, etc).

Step 4: Pattern evaluation and presentation. Basic operations include identifying the truly interesting patterns which represent knowledge, and presenting the mined knowledge in an easy-to-understand fashion [11].

3 Privacy Concerns and Ethics

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 even creepy sometimes. People have shown increasing concern about the privacy threats posed by data mining. For instance, you just searched some items on Amazon.com and you closed all the tabs. When you opened Amazon.com again, surprisingly, the items related to what you just searched appeared on recommendations. This is because the Amazon page reads your cookies and displays related advertisements (See Figure 2). 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. 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 [12-15].

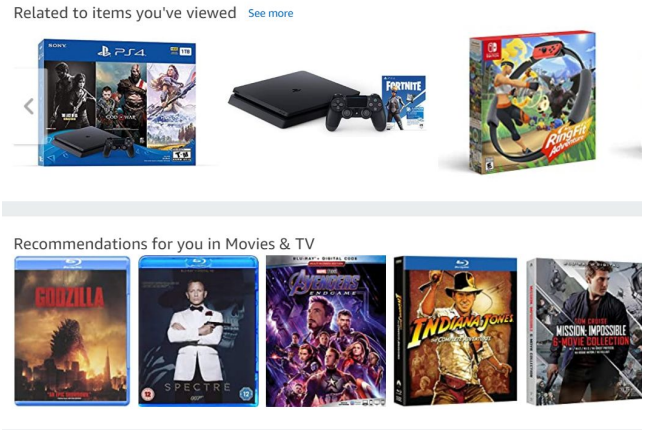

Figure 2: Web Personalization on Amazon.com

Individual's privacy may be violated due to the unauthorized access to personal data. the undesired discovery of one's embarrassing information, the use of personal data for purposes other than the one for which data has been collected [16]. For instance, the U.S. retailer Target once received complaints from a customer who was angry that Target sent coupons for baby clothes to his teenage daughter. However, it was true that the daughter was pregnant at that time, and Target correctly inferred the fact by mining its customer data [17]. From this story, we can see that the conflict between data mining and privacy security does exist.

All we have discussed above is the concerns from data providers. On one hand, the provider should be able to make his very private data which contains information that he does not want anyone else to know, inaccessible to the data collector. On the other hand, if the provider has to provide some data to the data collector, he wants to hide his

sensitive information as much as possible and get enough compensation for the possible loss in privacy [18].

From data collectors' view, the data collected from data providers may contain individuals' sensitive information. Directly releasing the data to the data miner will violate data providers' privacy, hence data modification is required. On the other hand, the data should still be useful after modification, otherwise collecting the data will be meaningless. 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. Therefore, the major concern of data collectors is to guarantee that the modified data contain no sensitive information but still preserve high utility [19-21].

4 Privacy Laws

The privacy law in Europe is rather strong in order to strengthen the rights of the consumers. However, the U.S. - E.U. Safe Harbor Principles, developed between 1998 and 2000, currently effectively expose European users to privacy exploitation by U.S. companies. As a consequence of Edward Snowden's global surveillance disclosure, there has been increased discussion to revoke this agreement, as in particular the data will be fully exposed to the National Security Agency, and attempts to reach an agreement with the United States have failed [22].

In the United States, privacy concerns have been addressed by the US Congress via the passage of regulatory controls such as the Health Insurance Portability and Accountability Act (HIPAA). The HIPAA requires individuals to give their "informed consent" regarding information they provide and its intended present and future uses. U.S. information privacy legislation such as HIPAA and the Family Educational Rights and Privacy Act (FERPA) applies only to the specific areas that each such law addresses. The use of data mining by the majority of businesses in the U.S. is not controlled by any legislation [23].

4 Google Data Collection

Google is not only the world's largest digital advertising company [24] but also provides No.1 web browser [25], the No.1 mobile platform, and the No. 1 search engine world wide [26]. Google has more than one billion monthly active users in its video platform, email service, and map application. Google collects a tremendous amount of data about people's online and real-world behaviors via its various products and then uses it to target people with paid advertising [27].

Google collects user data in two ways: active and passive. The obvious way is "active", in which users directly and consciously communicate information to Google. For example, people sign in to Google's widely used applications such as YouTube, Gmail, Search etc everyday. Less obvious ways for Google to collect data are "passive" which means that Google uses some applications to gather information from users without their notice. Google's passive data gathering methods arise from platforms (e.g. Android and Chrome), applications (e.g. Search, YouTube, Maps), publisher tools (e.g. Google Analytics, AdSense) and advertiser tools (e.g. AdMob, AdWords). We as users usually overlook the extent and magnitude of Google's passive data collection.

4.1 A Day In the Life of A Google User

A study [28] has done an experiment to check how data is collected by Google on a normal day. The experiment was designed in a way that a researcher carried an factory reset Android mobile phone device and configured as a new device to avoid prior user information associated with the device. A new Google account was created and the researcher then went about a normal day using the mobile phone associated with the new Google account.

The study used two tools to check the data collection by Google: My activity [29] and Takeout [30]. My activity was used to show data collected by Google from any Search-related activities, use of Google applications (e.g. YouTube video plays, Maps search, Google Assistant), visits to 3rd-party web pages (while logged in to Chrome), and clicks on advertisements. The Google Takeout tool provides a more comprehensive information about all historical user data collected via Google's applications (e.g.

it contains all past email messages on Gmail, search queries, location collection, and YouTube videos watched). The key information collection events are shown in Figure 3.

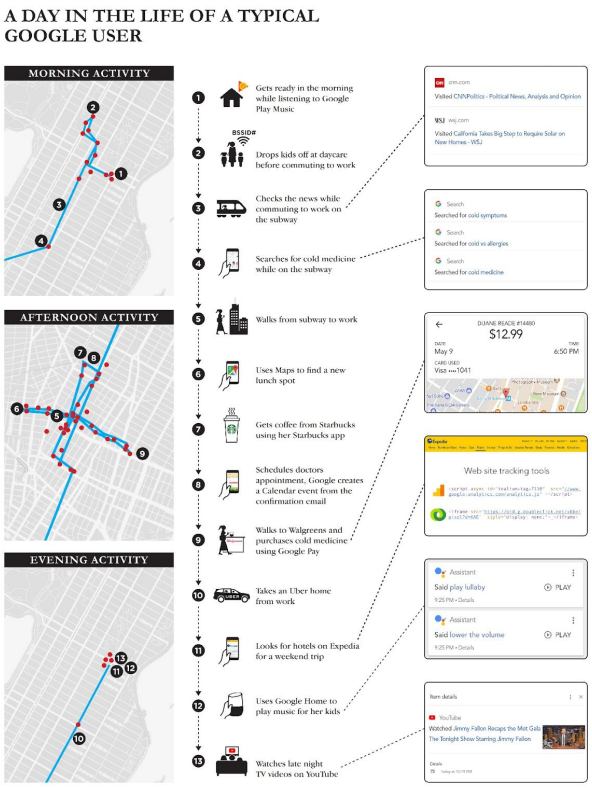

Figure 3: A typical day in the life of a google user [28].

In the activity shown in Figure 3, the study [28] shows that the number of "passive" data collection events outnumbered the "active" events by approximately two to one. The study also observes that Google analyzes the collected data to assess user interests, which it then applies to target users with appropriate ads. For example, Google provides a list of interests that it has inferred from a user's activities, available via the "topics you like" section in Google's Ad Personalization [31] web page.

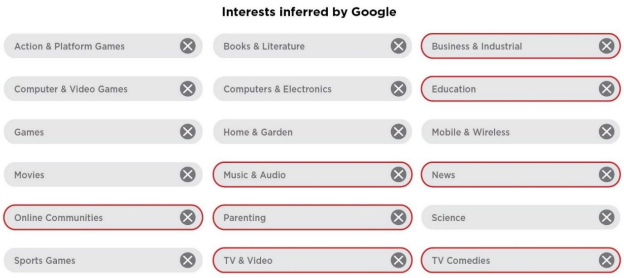

Figure 4: Google's assessment of user's interests at the end of the day [28].

Figure 4 shows a list of interests Google associated with the user's account after a day's use of activity. In total, Google associated 18 interests to the user, eight of which (shown by colored borders) closely matched the user's usage and activities.

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] 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.

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

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

[8] 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.

[9] 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.

[10] Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

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

[12] Seltzer, William (2005). "The Promise and Pitfalls of Data Mining: Ethical Issues"

[13] Pitts, Chip (15 March 2007). "The End of Illegal Domestic Spying? Don't Count on It". Washington Spectator. Archived from the original on 2007-11-28

[14] Taipale, Kim A. (15 December 2003). "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data". Columbia Science and Technology Law Review. 5(2). OCLC 45263753. SSRN 546782

[15] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques.San Mateo, CA, USA: Morgan Kaufmann, 2006

[16] L. Brankovic and V. Estivill-Castro, ``Privacy issues in knowledge discovery and data mining," in Proc. Austral. Inst. Comput. Ethics Conf., 1999, pp. 8999.

[17] R. Agrawal and R. Srikant, ``Privacy-preserving data mining," ACM SIGMOD Rec., vol. 29, no. 2, pp. 439450, 2000.

[18] Y. Lindell and B. Pinkas, ``Privacy preserving data mining," in Advances in Cryptology. Berlin, Germany: Springer-Verlag, 2000, pp. 3654.

[19] C. C. Aggarwal and S. Y. Philip, A General Survey of Privacy Preserving Data Mining Models and Algorithms. New York, NY, USA: Springer-Verlag, 2008.

[20] M. B. Malik, M. A. Ghazi, and R. Ali, ``Privacy preserving data mining techniques: Current scenario and future prospects," in Proc. 3rd Int. Conf. Comput. Commun. Technol. (ICCCT), Nov. 2012, pp. 2632.

[21] S. Matwin, ``Privacy-preserving data mining techniques: Survey and challenges," in Discrimination and Privacy in the Information Society. Berlin, Germany: Springer-Verlag, 2013, pp. 209221.

[22] UK Researchers Given Data Mining Right Under New UK Copyright Laws. Archived June 9, 2014, at the Wayback Machine Out-Law.com. Retrieved 14 November 2014.

[23] Judge grants summary judgment in favor of Google Books – a fair use victory". Lexology.com. Antonelli Law Ltd. Retrieved 14 November 2014.

[24] “Google and Facebook tighten grip on US digital ad market,” eMarketer, Sept. 21, 2017.

[25] “Market share or leading internet browsers in the United States and worldwide as of February 2018,” Statista, February 2018

[26] “Global OS market share in sales to end users from 1st quarter 2009 to 2nd quarter 2017,” Statista, August 2017.

[27] “Worldwide desktop market share of leading search engines from January 2010 to October 2017,” Statista, Feb. 2018.

[28] Douglas C. Schmidt, “Google Data Collection”, Digital Content Next, 2018.

[29] “My Activity,” Google, available at https://myactivity.google.com/myactivity

[30] “Download your data,” Google, available at https://takeout.google.com/settings/takeout?pli=1

[31] “Ads personalization,” Google, last accessed 2018, available at https://adssettings.google.com/authenticated