Analysis of Location-Based Tweets Related to Covid-19 on Social Networking Services

Ayahiko Niimi Faculty of Systems Information Science Future University Hakodate, Japan

Abstract

The present study aims to determine whether a privacy-preserving data mining method can be effectively applied in data mining for a social networking service (SNS). Data mining with privacy protection is a technology that is used to discover relevant knowledge from large datasets while protecting users' personal and sensitive information. The growing popularity of SNSs in recent years has raised concerns about user privacy, as SNS collects personal data from users such as address and birth date. It is now possible to provide secure personalized services to SNS users by implementing privacy preserving data mining on the personal information collected by an SNS. In a previous study, we considered using anonymized data mining to protect people's privacy. By this approach, all input information is anonymized while performing data mining. We examined whether the anonymization approach can be applied to data that can be partially anonymized, such as the SNS data, and how many users can be identified by the anonymization approach. However, previous research did not include an analysis of actual SNS data. In the current study, we examine tweets about Covid -19 and extract personal information from the content. We investigated whether the posting location could be estimated by examining the frequency of words in the posted content, with the correct answer data being the posting position of the tweet with location information. We analyzed tweets by state in the United States. According to the survey results, the top keywords in the posted content are place names. We confirmed the necessity of privacy protection data mining for SNS that we are proposing.

1. Introduction

Cybersecurity The social networking services (SNSs) utilizing personal information such as addresses and birthdays have been widely used in recent year. It is possible to provide services by performing data mining on personal information stored in an SNS. Data mining is a technique for extracting useful knowledge from large amounts of accumulated data. However, such a technique increases the risk of personal information leakage during data processing. As a result, studies on privacy-preserving data mining are being Conducted [1–3]. Data mining with privacy protection is a technology that extracts relevant knowledge from large amounts of data while protecting personal and confidential information. The privacy-preserving data mining has gained considerable attention. However, studies considering the possibility of identifying an individual when the secret information is combined with publicly available information that cannot be hidden are scarce.

The Author have proposed privacy protection data mining for SNS data such as SNS where some information has already been published [1]. However, previous research was not an analysis using actual SNS data. In the present study, we examine tweets about Covid-19 and extract personal information from the posted content. We investigated whether the posting location could be estimated by examining the frequency of words in the posted content, with the posting position of the tweet with location information serving as the correct answer data. We analyzed tweets by state in the United States. The top keywords in the posted content, according to the survey results, are place names. The latter is a confirmation of necessity of privacy protection data mining for SNS that is proposed.

2. Privacy-Preserving Data Mining

This section describes the types of data handled, main methods used in privacy-preserving data mining, and relationship between this study and privacy preserving data mining.

2.1. Data Types

Attribute data that can directly identify a specific individual, such as an individual number introduced in Japan or a social security number introduced in the United States, is called identifier data. Attribute data that can indirectly identify a specific individual, such as gender, birthday, and address, in combination with other attribute data is called quasi-identifier data.

2.2. Method Assuming a Third Party (Ideal Model)

In this method, a trusted third party (TTP) that, does not leak any information aggregates data and

performs data mining. This method is considered to be the safest. However, TTP installation is often unrealistic as it must be performed by the government or a reliable institution.

2.3. Anonymization

In the anonymization approach, data is processed and data mining is performed to avoid the identification of a particular individual. Specifically, processes such as the deletion of the identifier, integration of multiple variable values of the quasiidentifier into one category and converting a variable value to an ID are performed. Thus, even if the identifier is deleted, there is a possibility that the individual can be indirectly identified with the combination of other data. To achieve anonymity, it must be designed to achieve anonymity definitions such as anonymity and diversity[4], [5]. Anonymity is the property in which there are at least amount of data with the same number of attribute values. Diversity is the property in which there are at least variations in the attribute values of confidential data in the anonymity data.

2.4. Randomize

In randomization, random noise is added to personal information, and data mining is performed. Specifically, processes such as adding random noise to variable values, random exchange with the data of other individuals, and replacing variable values with random values are performed. Randomization is a lossy operation in which the restoration of original data is difficult; thus, privacy is protected. Computational cost for this method is low; however, accuracy and safety are statistical. Moreover, the higher is the degree of randomization, the higher is the safety but the less accurate are the results.

2.5. Encrypt

In encryption, data is encrypted, and data mining is performed. Secret calculation, which is one of the encryption approaches, is a technology that performs calculations such as statistical analysis and machine learning, while maintaining the confidentiality of personal information. In addition, it only outputs the results[6]. With data encryption, privacy is protected as the data is randomized and encrypted during the secret calculation. However, the computational cost of this method is high, but accuracy and security are more stringent than the aforementioned methods due to encryption.

2.6. Relationship with This Study

This study is related to anonymization. In the existing method, all data is anonymized and data

mining is performed. In this study, we consider the effect of anonymization of only a part of the data and performing data mining to reproduce the data published on the Internet, such as personal information of SNS.

3. Tweet Dataset for Covid-19

The Author considered Twitter as SNS data. Since Twitter makes daily tweets, it is thought that many tweets about the address, which is private information, are included. In the current study, we focused on tweets about Covid-19.

Covid -19, which is currently circulating worldwide, is on its way to becoming a historic pandemic if the spread of infection is not halted. Several studies and countermeasures are being conducted to prevent Covid-19 infection, but infection is still occurring. The spread has not been stopped. Information such as the details of Covid-19 and the spread of infection has been disseminated in the media daily, and lifestyles such as infection prevention and refraining from going out have changed. The United States has become the country with the highest number of Covid-19 infections. California has the highest number of infections in the United States, followed by Texas, Florida, New York, and other populous states. It has the greatest number of inhabitants [5]. It is assumed that this is simply due to the large population of these states, but when the ratio of infected people per 100,000 population is compared, North Dakota, states with relatively low populations such as South Dakota, and Wisconsin take the top positions [6]. We assumed that there were differences in people's movements in the provinces and other regions to prevent infection.

In this experiment, the tweets used in the experiment were collected using the Coronavirus (Covid-19) Tweets Dataset provided by IEEE [7]. Before performing the analysis, we perform text cleaning of the body of the tweet. The number of tweets used is counted one word at a time from the tweets that have been text cleaned. In addition, two or three consecutive words are examined. The entire United States was used as the criterion for dividing the tweets for the experiment. Furthermore, the time series is distinguished, and the analysis is conducted for the entire period from March 2020 to November 2020, as well as by month. As a result, the experimental results from each state are representative of the entire United States. Moreover, the reason for dividing the period by month is that it is believed that the content of the tweet may change during the period when the spread of infection is notable.

In the current study, we decided to use the Coronavirus (Covid-19) Geo-Tagged Tweets Dataset provided by the IEEE Data Port [7]. This dataset uses the words and hashtags related to Covid -19 by

the IEEE. This is a collection of English tweets containing Covid-19. These tweets contain geotagging information, which is indispensable for conducting this research. All tweets use this dataset. This solved the problem of collecting tweets by crawling, which was the problem of not being able to retrieve tweets efficiently because it was not possible to know whether or not they contained geotags. The body, latitude, longitude, city, state, county, country, and tweet posting date and time were obtained from the existing ID using the Twitter API. Among the tweet IDs listed in the dataset, the retweet ID is also included. It is included, and if all tweets are acquired, the results of experiments such as the frequency of appearance of words may change. Therefore, this time the retweeted post uses the tweet ID of the original tweet to which it is quoted. We tried to refer to it. This solved the problem that the same tweet was spread by the retweet function. This method collected 234,033 tweets with a collection period of approximately 1 month. The Author have already reported on the analysis by country[8].

4. Experiment

First, the Author investigated the relationship between the number of tweets by the state in the United States and the number of infected people per 100,000 population. This is because we thought that if the number of posts was biased, it would affect the subsequent experiments. Here is a link between the number of tweets by state and the number of infected people per 100,000 population. Table 1 summarizes the ranking of the number of tweets by state. The Author investigated whether there was a relationship between the number of tweets by state and the number of infected people per 100,000 population. The number of tweets appears to approximately correlate with the state's population. However, California, New York, Florida, and Texas, which have the most tweets, have the highest infection rate per 100,000 population. Is not particularly high in the United States. On the other hand, North Dakota, South Dakota, Wisconsin, and other states with a high number of infected people per 100,000 population have the most tweets, ranking 49th, 47th, and 31st, respectively. There were no outcomes. As a result, no correlation was discovered between the number of infected people per 100,000 population and the number of tweets.

The Author investigated the frequency of occurrence of words by state. Some of the results of this experiment are shown in Table 2. In most states, the word for a place name in the United States ranked first in the number of appearances. Even in the second place, the number of appearances was the highest in the results obtained in the experiments targeting all tweets such as "pandemic" and "distance." As a result, many words were found, and no characteristic words that are often used only in the United States were found. Table 3 shows the results of 2-gram. From this result, it can be seen that "social distance" is included in the top 5 in almost all states. Also, the usage of the word representing the place name is high as in the case of one word. Finally, Table 4 shows the results of 3-gram. In the case of 3gram, the number of appearances is different from the past, and there are many states that do not represent place names. Also, like New Hampshire, "get out and walk", "walk local", "walk alone" There were some states where such coined words and hashtags appeared. Words that appeared mainly in New York and Illinois, such as "soda" and "bottle". which were not found in one word or 2-gram in California, are appearing. In Georgia, "Atlanta hairstyle list" or something like a strange coined word such as "atlantanails" has emerged. According to the experimental results, frequently used words frequently represent place names in the state. Because the location name is included in the tweet, it was discovered that anonymizing the address, which is private information, is difficult.

5. Conclusions

The Author have proposed privacy protection data mining for SNS data such as SNS where some information has already been published [1]. However, previous research did not include an analysis of actual SNS data. In the present study, we examine tweets about Covid -19 and extract personal information from the content. The Author investigated whether the posting location could be estimated by examining the frequency of words in the posted content, with the correct answer data being the posting position of the tweet with location information. From the survey results, it was found that the top keywords in the posted content are place names. We confirmed the necessity of privacy protection data mining for SNS that we are proposing.

6. References

[1] Niimi A., and Arakawa, T. (2020). "Privacypreserving data mining with partially anonymized data in social networking service," in World Congress on Internet Security (WorldCIS-2020), London, UK, (Virtual Conference), December. p. 5pages.

[2] Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006). "L-diversity: privacy beyond k anonymity," in 22nd International Conference on Data Engineering (ICDE'06), April. pp. 24–24. [3] Sweeney, L. (2002) "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570. http://www.worldscientific.com/doi/abs/10.1142/S02 18488502001648 (Access Date: 01 October 2021).

[4] Cramer, R., Damg°ard, I., and Nielsen. J. B. (2001). "Multiparty computation from threshold homomorphic encryption," in Advances in Crypto logy — Eurocrypt 2001, B. Pfitzmann, Ed. Berlin, Heidelberg: Springer. pp. 280–300

[5] Covid -19 (2021). United States Cases by County — Johns Hopkins University. https://coronavirus.j hu.edu/us-map. (Access Date: 01 October 2021).

[6] JHU CSSE. (2019). Covid-19 dataset. https://gith ub.com/CSSEGISandData/COVID-19/tree/master/cs seCovid19data (Access Date: 01 October 2021).

[7] IEEE Dataport (2021). Corona Virus (Covid-19) Geo-Tagged Tweets Dataset. https://ieee-dataport.o rg/openaccess/coronavirus-covid-19-tweets-dataset. (Access Date: 01 October 2021).

[8] Niimi A. and Sasaya, K. (2021). "Frequency analysis of phrases using geotagged tweets for COVID-19," in COVID-19 Challenge International Virtual Summit, London, UK, (Virtual Conference), March. p. 2.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 20H01728.

| Rank | State | # of Tweets | Rank | State | # of Tweets |
|------|----------------|-------------|------|---------------|-------------|
| 1 | california | 20679 | 26 | missouri | 918 |
| 2 | new york | 18029 | 27 | connecticut | 825 |
| 3 | florida | 7450 | 28 | alabama | 775 |
| 4 | texas | 7183 | 29 | kansas | 715 |
| 5 | georgia | 3881 | 30 | kentucky | 699 |
| 6 | new jersey | 3630 | 31 | wisconsin | 685 |
| 7 | maryland | 3196 | 32 | utah | 503 |
| 8 | illinois | 2919 | 33 | oklahoma | 428 |
| 9 | pennsylvania | 2757 | 34 | vermont | 382 |
| 10 | north carolina | 2408 | 35 | new mexico | 380 |
| 11 | massachusetts | 1995 | 36 | arkansas | 319 |
| 12 | ohio | 1950 | 37 | iowa | 316 |
| 13 | virginia | 1876 | 38 | nebraska | 315 |
| 14 | washington | 1834 | 39 | delaware | 290 |
| 15 | michigan | 1599 | 40 | mississippi | 278 |
| 16 | tennessee | 1568 | 41 | montana | 272 |
| 17 | arizona | 1498 | 42 | new hampshire | 269 |
| 18 | louisiana | 1433 | 43 | rhode island | 262 |
| 19 | indiana | 1378 | 44 | maine | 258 |
| 20 | oregon | 1373 | 45 | west virginia | 254 |
| 21 | nevada | 1340 | 46 | idaho | 243 |
| 22 | colorado | 1245 | 47 | south dakota | 149 |
| 23 | hawaii | 1099 | 48 | alaska | 118 |
| 24 | south carolina | 933 | 49 | north dakota | 88 |
| 25 | minnesota | 920 | 50 | wyoming | 73 |

Table 1. Number of Tweets by State

| Rank State | 1 | 2 | 3 | 4 | 5 |
|----------------|---------------|------------|------------------|------------------|-----------------|
| california | california | pandemic | quarantine | day | los |
| new york | new | york | pandemic | challenge | face |
| florida | florida | pandemic | miami | beach | quarantine |
| texas | texas | austin | pandemic | atx | all512 |
| new jersey | new | jersey | pandemic | social | distance |
| georgia | atlanta | georgia | pandemic | social | ga |
| maryland | maryland | pandemic | amp | day | quarantine |
| illinois | chicago | illinois | pandemic | amp | quarantine |
| pennsylvania | pennsylvania | pandemic | philadelphia | quarantine | amp |
| north carolina | north | carolina | amp | pandemic | home |
| ohio | ohio | pandemic | amp | social | quarantine |
| massachusetts | massachusetts | boston | pandemic | amp | home |
| virginia | virginia | pandemic | amp | dav | social |
| arizona | arizona | pandemic | phoenix | quarantine | tucson |
| tennessee | tennessee | nashville | pandemic | quarantine | dav |
| indiana | indiana | day | pandemic | social | distance |
| washington | washington | seattle | pandemic | dav | time |
| michigan | michigan | detroit | quarantine | nandemic | dav |
| colorado | colorado | denver | nandemic | home | time |
| oregon | oregon | nortland | pandemic | new | home |
| kaneae | state | united | pandenne | kaneae | case |
| louisiana | nola | ouarantine | 2020 | Railodo | louisiana |
| norrada | nota | quarantine | 2020 | nendomio | iouisialia |
| nevada | vegas | las | hermo | pandenne | social |
| connecticut | connecticut | pandemic | nome | den | us |
| south carolina | soum | carolina | distense | day | quarantine |
| alabama | alabama | social | distance | piease | pandemic |
| minesota | miniesota | pandenne | minieapons | amp | new |
| missouri | missouri | amp | music | snare | piay |
| hawan | hawan | pandemic | lockdown | social | honolulu |
| kentucky | kentucky | louisville | pandemic | tree | social |
| utan | utan | pandemic | tax | service | time |
| oklahoma | oklahoma | pandemic | like | us | time |
| vermont | quarantine | amp | vermont | ugcomedyshow | new |
| rhode island | island | rhode | pandemic | amp | providence |
| wisconsin | wisconsin | pandemic | ha | drink | milwaukee |
| lowa | 10Wa | des | moines | pandemic | amp |
| mississippi | mississippi | pandemic | new | oxford | canteen |
| new mexico | new | mexico | albuquerque | pandemic | quarantine |
| arkansas | arkansas | little | pandemic | rock | today |
| nebraska | nebraska | new | socialdistancing | lincoln | unitedstates |
| idaho | idaho | boise | end | north | social |
| montana | montana | ap | case | pandemic | county |
| new hampshire | new | hampshire | pandemic | socialdistancing | social |
| delaware | delaware | beach | 2020 | time | pandemic |
| west virginia | life | pandemic | stop | size | yourlifeyourway |
| maine | maine | pandemic | us | trolley | today |
| south dakota | south | dakota | fall | sioux | pandemic |
| alaska | alaska | anchorage | pandemic | world | day |
| wyoming | wyoming | mask | pandemic | covid19wy | apocalypse |
| north dakota | north | dakota | fargo | today | get |

Table 2. Frequency of Word Occurrence by State

Table 3. 2-Gram by State

| Rank State | 1 | 2 | 3 |
|----------------|-------------------------------|------------------------------|--|
| california | los angeles | angeles california | corona california |
| new york | new york | face shield | empty soda |
| florida | social distance | miami florida | beach florida |
| texas | atx austin | austin all512 | coronavirus pandemic |
| new jersey | new jersey | social distance | face shield |
| georgia | atlanta georgia | social distance | dunwoody ga |
| maryland | covid19 coronavirus | coronavirus pandemic | social distance |
| illinois | chicago illinois | social distance | face shield |
| pennsylvania | social distance | philadelphia pennsylvania | wear mask |
| north carolina | north carolina | social distance | overall piedmont |
| ohio | new york | social distance | face shield |
| massachusetts | boston massachusetts | social distance | corona virus |
| virginia | social distance | virginia beach | richmond virginia |
| arizona | social distance | phoenix arizona | tucson arizona |
| tennessee | nashville tennessee | social distance | coronavirus self |
| indiana | social distance | distance day | hammond indiana |
| washington | seattle washington | social distance | hand sanitizer |
| michigan | social distance | quarantine day | acres mi |
| colorado | denver colorado | social distance | wear mask |
| oregon | portland oregon | social distance | wear mask |
| kansas | united state | benlegends urbangraffitisben | urbangraffitisbcn artdengroundmoviment |
| louisiana | nola 2020 | quarantine nola | new orleans |
| nevada | las vegas | vegas nevada | social distance |
| connecticut | social distance | home work | today 's |
| south carolina | south carolina | social distance | charleston south |
| alabama | social distance | corona churchstreeteast | churchstreeteast downtownmobile |
| minnesota | minneapolis minnesota | social distance | saint paul |
| missouri | consider music | music share | share covid19 |
| hawaii | social distance | honolulu hawaii | wear mask |
| kentucky | louisville kentucky | free insurance | insurance quote |
| utah | oso tax | tax service | salt lake |
| oklahoma | social distance | tulsa oklahoma | oklahoma city |
| vermont | pnandfriends pnisawesome | pnisawesome full | full episode |
| rhode island | rhode island | social distance | providence rhode |
| wisconsin | social distance | offhaus summer | summer covid |
| iowa | des moines | social distance | downtown des |
| mississippi | new oxford | oxford canteen | brothers divid |
| new mexico | new mexico | albuquerque new | social distance |
| arkansas | little rock | social distance | rock arkansas |
| nebraska | socialdistancing covid 19 | lincoln nebraska | google searches |
| idaho | boise idaho | north end | social distance |
| montana | ap montana | gallatin county | covid-19 cases |
| new hampshire | new hampshire | social distance | getoutandwalk walklocal |
| delaware | social distance | state fair | wear mask |
| west virginia | makelifehappen lifeexperience | lifeexperience pandemic | stop life |
| maine | covid-19 policies | review covid-19 | policies purchase |
| south dakota | south dakota | sioux fall | fall south |
| alaska | anchorage alaska | social distance | coronavirus covid19 |
| wyoming | covid19wy covid19 | covid19 apocalypse | apocalypse happyapocolypse |
| north dakota | north dakota | fargo north | get test |

Table 4. 3-Gram by State

| Rank | 1 | 2 | 3 |
|----------------|--|--|--|
| State | 1 | 2 | 3 |
| california | los angeles california | soda bottle face | empty soda bottle |
| new york | empty soda bottle | bottle face shield | face shield become |
| florida | en miami florida | dustproof mask washable | mask washable price |
| texas | atx austin all512 | local news coronavirus | news coronavirus pandemic |
| new jersey | empty soda bottle | soda bottle face | bottle face shield |
| georgia | greatthingsatlanta atlanta covid_19 | atlanta covid_19 covid19 | covid_19 covid19 atl |
| maryland | perry hall maryland | covid19 coronavirus pandemic | hot todd lincoln |
| illinois | empty soda bottle | soda bottle face | bottle face shield |
| pennsylvania | badstreet philadelphia pa | make best horrible | essential work may |
| north carolina | overall piedmont area | piedmont area traffic | charlotte north carolina |
| ohio | covid-19 pandemic quarantine | pandemic quarantine people | quarantine people granville |
| massachusetts | yeah nantucket stay | nantucket stay home | stay home safe |
| virginia | virginia beach virginia | coronavirus flattenthecurve cheflife | close tuesdays wednesdays |
| arizona | corona de tucson | de tucson arizona | free covid-19 test |
| tennessee | coronavirus self quarantine | self quarantine day | livestreammusic -wolff rock |
| indiana | social distance day | national kidney foundation | chronic disease coalition |
| washington | sure know rule | know rules know | rules know social |
| michigan | work home desk | home desk detroit | desk detroit mi |
| colorado | private suv 7 | suv 7 pax | 7 pax ideal |
| oregon | accurate amp date | amp date information | date information click |
| | bcnlegends | urbangraffitisben | artdengroundmoviment |
| kansas | urbangraffitisben | artdengroundmoviment | blegends |
| | artdengroundmoviment | blegends | acabose |
| louisiana | quarantine nola 2020 | new orleans louisiana | posted photo quarantine |
| nevada | las vegas nevada | 10yrs older need | older need wear |
| connecticut | drive home work | home work brought | today 's drive |
| south carolina | charleston south carolina | columbia south carolina | hilton head island |
| alabama | corona churchstreeteast | churchstreeteast downtownmobile | downtownmobile mobilealabama |
| | downtownmobile | mobilealabama | downtownmobileal |
| minnesota | saint paul minnesota | meals minnesotans need | minnesotans need covid19 |
| missouri | consider music share | music share covid19 | share covid19 edition |
| hawaii | mile trail run | apply online today | ewa beach hawaii |
| kentucky | free insurance quote | cincinnati metropolitan area | metropolitan area covid-19 |
| utah | oso tax service | salt lake city | lake city utah |
| oklahoma | missed would like | would like review | oklahoma city oklahoma |
| vermont | pnandfriends pnisawesome full | pnisawesome full episode | full episode 'pn |
| rhode island | providence rhode island | covid19 entertainment sales | beef entertainment corporation |
| wisconsin | offhaus summer covid | summer covid brew | covid brew ha |
| iowa | downtown des moines | des moines jowa | des moines des |
| mississippi | new oxford canteen | brothers divide i.w | divide i.w worsham |
| new mexico | albuquerque new mexico | new mexico covid-19 | santa fe new |
| arkansas | little rock arkansas | first baptist little | baptist little rock |
| nebraska | google searches 4 | walkalone social distancing nhscenery | mst covid19 unitedstates |
| idaho | pick dinner corona | dinner corona village | corona village meridian |
| montana | glacier national park | ap montana reports | vellowstone national park |
| | | | walkalone |
| new hampshire | getoutandwalk walklocal walkalone | walklocal walkalone socialdistancing | socialdistancing |
| delaware | lose track time | 2020 Jose track | healthy wear mack |
| ucus ware | makelifehannen lifeevnerience | 2020 IOSE URLA | lieunity wear union |
| west virginia | pandemic | stop life live | expectations yourlifeyourway |
| maine | covid-19 policies purchase | review covid-19 policies | policies purchase tickets |
| south dakota | sioux fall south | fall south dakota | midst worldwide pandemic |
| alaska | thanks mtn_peach guest | mtn_peach guest submission | guest submission myfavoritemask |
| | covid19wy | covid19 | |
| wyoming | covid19 | apocalypse happyapocolypse | apocatypse nappyapocotypse coronadiary |
| north dakota | fargo north dakota | think may expose | get test free |
| north theote | ingo north thatota | uning may expose | Ber rest mee |