Search Keyword Based System forProspective Threat Avoidance Using Big Data

Sunny Shah
Department Of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, Maharashtra, India
sunnyshah0588@gmail.com

Abstract—Every organization in the world collects information about their user for providing better service and this information is stored in data warehouses where they apply different algorithms to make their service best suited for theirusers. On other end government authorities for public safety are still following the old methods and protocols to solve crime in this modern age of computing.These authorities need to acquaint themselves with contemporary style for solving the crime and to be upto date with the technology around them. This paper proposes a new and advanced method to combine this two public service in a convenient way. So, by using the potential of computing we can make our neighborhood safer.

Keywords—search keyword, information sharing, threat, service, crime, big data, the internet, public protection agency

I. INTRODUCTION

The traditional definition of information sharing is a one-to-one exchange of data between sender and receiver[1]. We in our daily life follow this tradition by sharing our knowledge with others. A threat under the law can be communicated intent to inflict harm or damage to a person or property to force someone's compliance or to restrict his or her freedom. Threatening behavior (including the use of abusive or insulting words) is a generally a criminal offense punishable with imprisonment and/or a fine [14].

Internet giants have realized that information is power. Information about the Internet, information about innumerable trends, information about its user.Google, Facebook, YouTube, etc. magnet have been implementing data collection since a long time and by now alone googlehas collected about 10 Exabytes of data (10 billion gigabytes).Facebook, the world's most popular social networking site has 300 million photos, 2.5 billion pieces of content all which add up to 500 terabytes of data [2].

On other hand government agencies for public protection also have set up their databases for developing and protecting public from threats in society. This database stores information about convicted criminals and past crimes. But they can never know or predict the future crimes. So this makes those agencies a correctional system, not a prevention. Criminals are much smarter, and with the power of the internet in everyonce hand have made them, even more smarter.

This two different but same in nature organization have faced significant problems working together.The government authorities on another hand the have to subpoena the records of the criminal to prove him guilty in a court of law. Which usually takes a bit of time and it may happen that he can be released due to lack of evidence. So, this proves that the antiquated system lacks significant amount of efficacy that can help them to expedite the process of solving crime.

Contribution: The contribution of this work are as follows:

1) To the best of the author's knowledge, this is the first work to propose a way to share information among different agencies and private companies for threat avoidance using search keyword.

2) We have shown how to use the latest technology of big data, information sharing through the cloud to prevent thethreat.

3) Have proved how this paper can be mapped in the real cases.

Organization: The rest of the paper is organized as follows: section II provides a background of big data, Google's information-gathering channels, and Google's data collection machine. In section III we have described the challenges faced by authorities in solving crime in the advent of technology.In section IV we describe the conceptual model of search keyword based information sharing and two case study.In section V we provide some related work and in section VI we end with a conclusion.

II. BACKGROUND

In this section, we first present the overview of the big data then explain Google's information gathering channels and how they use.

A. Big Data

Data that's *too big*, *too fast*, or *too hard* for existing tools to process is the most famous definition of big data by Madden et al [3]. Previously it was characterized by 3 vs but now as we have gone on in depth we discovered 4 new v's which makes total 7 of them as follows: volume, velocity, variety, variability, veracity, visualization, and value [4].

Volume: What used to be measured in Gigabytes (GB) is now measured in Zettabytes (ZB) or even Yottabytes (YB).This is how much data we are acollecting and we have. The invention of new technologies plus the increase in growth of IOT (Internet of Things) is creating data at an exponential rate. In 2009 we had at most.79 ZB but it is expected that by 2020 we will gather 73.5 ZB data out of which 1/3 of data produce will live in or pass through the cloud [4].

Velocity: The speed at which data is accessible or the frequency of incoming data that needs to be processed. Every minute people upload 100's of hours of videos on YouTube. In addition, every minute over20 million photos are viewed,more than 200 million emails are sent, and almost 2.5 million queries on google are performed [5].

Variety: Data can be XML to photos to videos, that states that data can be of any type either structured or unstructured or combination of both as semi-structure, which is one of the biggest challenge for big data. Moreover, to organize the data is also a challenging task, especially when the rate at which data is changing is massive [4].

Variability: The meaning of this is constantly changing which can have a huge impact on your data homogenization. It's like, a coffee shop may offer 6 different blends of coffee but if you get the same blend every day and it tastes different every day, that variability [4].

Veracity: It's about making sure the data is accurate, which requires theprocess to keep the bad data from accumulating in your system [4].

Visualization: Using graphs and charts to visualize a large amount of complex data is much more effective in conveying ameaning that reports and spreadsheets full of numbers [4].

Value: Data is the new oil and by now we have a lot. But data itself is not valuable at all. The value is the analyses done on the data and how the data is turned into information and eventually turning it into knowledge.McKinsey states that potential annual value of big data to the US Health Care is $ 300 billion, more than double the total annual health care spending of Spain [15].

B. Google's Information-gathering Channels

As of February 2016, Google Search is the most-used search engine in the US with 64% market share. With 4.5 billion active user and available in 123 languages almost more than half of the population on earth uses Google [6]. But most of the people don't know that Google gathers, even more, information than we know. [9]

Searches: Handling more than three billion searches each day [7][8].Google tracks all searches, and now with search becoming more and more personalizes, this information is bound to grow increasingly detail and user specific [9].

Website Analytics: Google analytics is by far the most popular website analytics package. Being free and supporting a number of advanced features, it's used by a large percentage of the world's website[9].

Ad Serving: AdWords and AdSense are cornerstones of the Google's financial success, but this also provide google with a lot of valuable data like which ads are people clicking on? Which product is running in themarket? All this is useful information [9].

Twitter: Now the Google has a direct access to all tweets that pass through Twitter after a deal made late last year [9].

YouTube: The world's most popular video site is owned by Google. Which gives Google a huge amount of information about its users' viewing habits [9].

This list of can goes on and on which proves how much we didn't know. Those logs are kept for 9 months and cookies aren't anonymized until after 18 months [9].

C. Google's Data Collection Machine

There are many different aspects of Google's data collection. Google generally log's the IP addresses that made a request, cookies are used for setting and tracking purpose, and if you are logged into Google account, what you do on Google-owned sites can often be coupled to you personally, not just your computer. So if you're using Google service it will know what you're searching for, which websites you visit, what videos you watch, which topic interest you, what you want to buy, and more [9].

It should be mentioned that Google isn't alone in doing this kind of data collection. Rest assured that Microsoft is doing similar things with Bing and Hotmail, to name just one example [9].

D. Analytics Server

Analytic Server is comprehensive enterprise analytics platform which fuses batch and real-time analytics of any source of data with predictive via machine learning. In abusiness context, this effectively helps enterprises to gather insightful information and make awell-informed decision, monitoring this insight via theuser-friendly dashboard. In abroader application, this allows for any cluster of theevent to be analyzed, monitored and even predicted [16][17].

III. CHALLENGES IN SOLVING CRIME

There is a revolution going on in the criminal activity. It creates amajor problem for law enforcement in almost every part of the world. The Internet is almost available to every citizen in the country and couple of country have made it as a basic citizen right but not always it has been used for right things. There are 3, 18,000 results available in just 0.56 seconds if searched "How To Make A Plastic Explosive" [10] and themajority of those result can have been used to make a real plastic explosive to commit a crime. While it's a challenge for the crime investigator and first responder to deal with this issue but it's always difficult to do so and in some major case the criminal goes free or the crime is unsolved.

As the technologies are growing it is the responsibility of the government to ensure their citizen's safety but due to many reasons they are not able to keep up with this age this is one of the major challenges faced by the authorities.

IV. INFORMATION SHARING MODEL

In this section, we first describe the conceptual model for information sharing and provide the methodology to use it and thentwo theoretical case studies based on this.

A. Information Sharing Conceptual Model

Based on the challenges, we propose an information sharing conceptual model to support public protection agencies using big data. The proposed model is presented in Figure 1.

Users normally search on the internet using some kind of search engine, and the companies which provide this services usually records the data and maintain those data on their server and some data on the user's system in form of cookies. But before the data passes to the data warehouse of the company this data passes to the server often called analytics server. This server is configured with analytics program to be run on those datasets which pass through that server anduse machine learning to make their service more efficient, to make amarket decision, do research, and refine its product. With all this data at fingertips, they can group data together in avery useful way. Sometimes not just per user or visitor, but also can examine trends and behaviors of entire cities or countries.

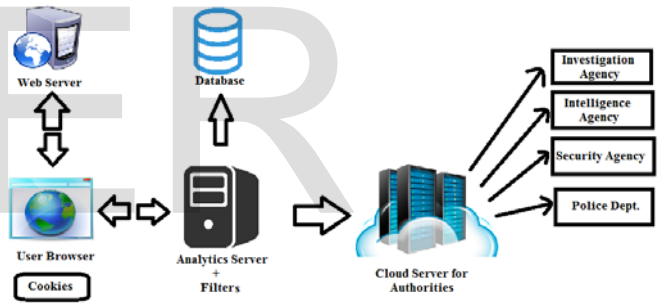

Figure 1: An information sharing conceptual model

Now if this analytics server has a special filter which is capable of machine learning and has configured with some search pattern-based algorithms so if a user searches or does something on theinternet which matches the pattern of a criminal or crime it can pass the IP and the search pattern to the government authorities. This data that the analytics server passes goes to the cloud server where every government organization can have access to those and whichever agencies got the jurisdiction can take the case. Then, this agency can keep a track of the person or group its searches and if they feel its activity to be suspicious they can apply for the search warrant. This way they can prevent the crime before it can see the light of day and can make anarrest of the criminal. Here it will be beneficial that when we are using thecloud to store data so every agency can have access to the case and also it is going to be cheap for every agency because they don't have to acquire a server and have no extra maintains charges.

This model is not just for the search engines but it can also be used with social sites like Facebook, Twitter, Instagram, etc. Today's world everyone is obsessed with this sites and apps and if we can make those analytics server's filter customized to this sites this can make a lot of difference and add another layer to identify criminals.

Using the current technique it would be never possible for this agency to do this kind of analytics and make anarrest because they lack this types of tools and technique to use those tools. But after this model in action they can make more progress and the crime rate connected with the internet may go down.

B. Case Study based on model

In this section, we provide two theoretical case study to understand how this model can work better than thecurrent system.

1. Case Study 1

In this theoretical case study, we are supposing that a man in connection with ISIS want's to damage the government property killing the people nearby so he starts to build a plane. Not every criminal knows how to create a plastic c4 explosive, so he goes to the internet and assuming he is using a search engine to look for the materials to make an explosive. After finding the materials he looks for the planes of public places using online maps and 360' tour of the public spots. He searches for the peak time when the crowd is full or the pattern when the security guards change the post, etc. Creating a plane now he can execute the plan.

In this scenario the public protection agency with the current system will never be able to stop this crime before it has been committed but if implied the information sharing model the analytics server with filter will first analyses the pattern of the search like he searched for the materials which can be used to make c4 explosive then searched for some circuit then the place and map of the place this kind of searches will alert the system and will flag his IP and send this search information and IP to the public protection agency's cloud. The nearest agency to that IP now can keep track of more suspicious activity and if they find any they can make an arrestbefore the crime happens.

So this proves how this model can be used to fight crime before it has committed and avoided aprospective threat to society.

2. Case Study 2

In this theoretical case study, we suppose that a man has been driving a car and its late night. He was drunk and had blood alcohol level more than the legal limit and in that situation he crashed the car and killed someone, seeing that he shocks and flee the scene.But it is a person tendency that he cannot forget that and searches the internet for anaccident in that place where he crashed the car and killed someone then he searches for something that can lead authorities back to him in news, also searches for a lawyer, and have also filed aninsuranceclaim for thecar.

In anormalscenario, the authorities will have to go by protocol before deciding this case as homicide or accident. They will search the screen and if they didn't find any evidence so they follow the normal procedure in which they will find any motive or revenge factor and like this, the investigation will go on which will take a lot of humans as well as technical resources which can be applied to other cases. If the model is implemented the server will match the keyword and the crime in the area of that keyword and if they match server will flag the IP to the authorities and if they can link this thing, he will be arrested and will solve the crime and also will save a lot of resources.

With this two case study's it can be proven that is amodel is not only useful for pre-crime but also for solving post-crime.

V. RELATED WORK

NYPD (New York Police department) was one of the first police department which implemented the system called Domain Awareness System (DAS) which gives real-time information about happenings in the city. The City has approximately 3,000 Closed-Circuit TV cameras connected to the Domain Awareness System. Through which investigators will have immediate access to information through live video feeds, and instantly see suspect arrest records, 911 calls associated with the suspect, related crimes occurring in the area and more, investigators can map criminal history to geospatially and chronologically reveal crime patterns [11].

Like digital fingerprints, we leave small clues about our lives all over the Internet and social media is no exception. Posts on Facebook, YouTube, Twitter and other social media sites commonly lead to arrests and convictions for theft, DUI, drug offenses, assault and battery, white-collar crimes and sexual assault. Because police are using Facebook to solve crimes and catch criminals [12].

Canada's Multi-Agency Situational Awareness System (MASAS) is an information aggregation system that facilitates sharing situational awareness within the public safety community. Information shared relates to incidents and planned events. It includes public alerts, risks to responders, and community profiles [13].

VI. CONCLUSION

The use of theinternet have increased and in this age, it has become an urgent need to be upto date with this technology. Since the government authorities are for the public protection they should also keep up with this changes and for doing so they have to implement some tools that can help them to do their jobs much more efficiently and in an effective manner so it is important to implement the conceptual model and integrate this model with the other real-time tools to get the best out of it.

REFERENCES

[1] www. wikipedia.org, "Information_sharing,"https://en.wikipedia.org/wiki/Information_sharing.

[2] Daniel Price, "Surprising Facts and Stats about the Big Data Industry," http://cloudtweaks.com/2015/03.

[3] S. Madden, "From databases to big data," IEEE Internet Computing, vol. 16, no. 3, pp. 4–6, 2012.

[4] Ashley Devan, "The 7 V's of Big Data,"https://www.impactradius.com/blog/7-vs-big-data/,April7,2016.

[5] Mark van Rijmenam," Why the 3V's Are Not Sufficient to Describe Big Data," floq.to/5Yai6.

[6] Lela, Adam (March 16, 2016). "comScore Releases February 2016 U.S. Desktop Search Engine Rankings". ComScore.com. Retrieved June 27, 2016.

[7] "Digital Indians: Ben Gomes". BBC News. Retrieved June 28, 2016.

[8] "Almost 12 Billion U.S. Searches Conducted in July". SearchEngineWatch. September 2, 2008.

[9] Tech Blog,"How Google Collects Data About You And Internet,"http://royal.pingdom.com/2010/01/08/.

[10] Google.com, "How To Make Plastic Explosive".

[11] Press Release "New York City's Domain Awareness System",goo.gl/pLZAHM.

[12] Brian Hughes," Police are Using Facebook to Solve Crimes: Is yourPrivacy at Risk? "https://smallbiztrends.com/2015/12/.

[13] CanOps, "Multi-Agency Situational Awareness System," http://www.canops.org/?page=AboutMASAS

[14] http://www.businessdictionary.com/definition/threat.html

[15] BaselKayyali,"Revolution In US Health Care", http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care

[16] "Data Analytics Server ",WSO2,http://wso2.com/products/data-analytics-server/.

[17] "SPSS Analytic Server', by IBM,http://www-03.ibm.com/software/products/en/spss-analytic-server

[18] "iSecure",byDelopt,http://www.delopt.co.in/video-analytics-server.html.