Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 82 (2016) 43 – 48

Procedia Computer Science

Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia

Big data analytics for behavior monitoring of students

Abdul Rauf Baig[a,*], Hajira Jabeen[b]

[a]College of Computer & Information Sciences, Al-Imam Mohammed Ibn Saud Islamic Univ. (IMSIU), Riyadh, Saudi Arabia
[b]AKSW, Institute for Applied Informatics, University of Leipzig, Leipzig, Germany

Abstract

Security threat from senseless terrorist attacks on unarmed civilians is a major concern in today's society. The recent developments in data technology allow us to have scalable and flexible data capture, storage, processing and analytics. We can utilize these capabilities to help us in dealing with our security related problems. This paper gives a new meaning to behavioral analytics and introduces a new opportunity for analytics in a typical university setting using data that is already present and being utilized in a university environment. We propose the basics of a system based on Big Data technologies that can be used to monitor students and predict whether some of them are becoming prone to deviant ideologies that may lead to terrorism.

Peer-review under responsibility of the Organizing Committee of SDMA2016

Keywords: business intelligence; business analytics; big data; student behavior monitoring; behavioral analytics;

1. Introduction

Behavioral analytics is an emerging area of research[1,2]. It focuses on the "how" and "why" of behaviors taking a holistic and human view of data. Traditionally, the term has been reserved for data collected for E-commerce, but in this paper we broaden its scope and use it to include data from all available resources in order to predict if someone is becoming prone to deviant ideologies leading to terrorist tendencies. This is a sensitive but very important topic as new challenges involving terrorism and killing of innocent civilians are emerging and we have to do all that is possible to deal with the situation. This includes turning to rapidly evolving new technologies to help us.

In a university setting, a sizeable amount of personal and academic data is usually available for each student and it can be augmented with data from other sources that we will discuss in later sections. Various tools and techniques[4]

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: raufbaig@ccis.imamu.edu.sa

Peer-review under responsibility of the Organizing Committee of SDMA2016
doi:10.1016/j.procs.2016.04.007

are already available for automatic monitoring and analysis of data collected regarding the students' activities. The recent developments in data technology allow us to have scalable and flexible data capture, storage, processing and analytics. We can now utilize these powerful capabilities to integrate a large quantity of data from multiple sources and use them to extract worthwhile information.

Even though the ideas propounded in this paper can be easily applied to other situations and settings, but we limit the scope of our discourse to monitoring of student activities at the university level. Particularly, we note that the idea can be easily extended to include teachers and staff and that they can be more dangerous as they easily influence dozens or even hundreds. Furthermore, with some modifications the endeavor can be at the level of Ministry of Higher Education to include all universities or even at the level of Ministry of Education to include schools (K-12). In fact, this type of effort can be easily modified for any organization or enterprise large enough to justify the cost and benefits.

The rest of the paper is organized as follows. Analytics, particularly Big Data analytics pertaining to educational sector data is discussed in the next section. In Sections III, IV, and V, we discuss the sources of data, the possibilities of capturing, storing, and analyzing behavioral analytics of students, respectively. In Section VI we give our concluding remarks.

2. Educational sector behavioral analytics

In the educational sector setting, the most relevant analytics would be that which supports pedagogy. There are several examples of efforts made in this regard. In[3] the authors report student performance prediction, intelligent course recommendation, data driven learning analytics, and personalized learning. An interesting example is given in[4] regarding data analysis of students from 2000 high schools of Colorado, USA. The total number of students analyzed was 860,000 and the aim of the analysis was to discover a student's readiness for college or career. In[5], the authors propose an architecture for using Big Data in education.

Our idea, though described in an educational setting, does not focus on pedagogy. Analytics pertaining to student behavior aims at monitoring a student to observe if he or she is deviating from normal behavior. The normal behavior is defined in context of a particular issue that is in focus. Judging from the news items and newspaper articles one of the concerns of recent times is the ideological deviation leading to terrorism, particularly among young people. We focus on the monitoring and anomaly detection of students with regards to this issue.

Ideological deviation, if left unchecked, can make the student wander into the realm of unlawful activities. The monitoring and prediction could provide us with early detection capability that can be backed up by information, guidance, advice, and early feedback to the person concerned and also can be used to notify the relevant authorities. That may help to improve the student's behavior and also make him/her realize where he/she is heading and its consequences. A good and reliable model of a "normal" student could also be beneficial for academic and other purposes. From the society's point of view, the biggest benefit of all is that this system, if in place, can save innocent lives and curb other negative fallouts of terrorist activities, in addition to saving the future of our precious youth.

In the following, we give a perspective on how the analytics has evolved in recent years to allow us the implementation and exploitation of our idea.

Informally, analytics can be defined as the process of getting insights from the available data. According to one prevalent view point analytics has three levels[4]. At the basic level we have descriptive or reporting analytics. It is about knowing what has happened and why it has happened. The next level is predictive analytics[6] that attempts to foretell what is going to happen. The third level builds upon the first two levels and recommends actions to deal with a given situation. It is called prescriptive analytics.

Analytics allows organizations to take actions based on data driven insights rather than only on experience and intuition of their executives. It helps businesses to be more effective and efficient in their operations, management, and strategic planning.

Analytics is not a recent discipline. Ever since we had databases, we have been able to query them. However, with databases we have the problem that critical data we need for reports and analysis is scattered among numerous different applications and systems (PC, mobile phones, etc.) residing on many different platforms, in many different formats, in many different organizations (internal and external) and geographies.

The early 1990s introduced us to the concept of data warehouses (DWH)[4]. With DWH we can bring selected data from each of our different applications together into a single store of information and that is where we run our reports and do our analysis. The DWH provides us with one-stop servicing for our data needs. As data is brought into the DWH (Extract), the transformation (or unification) of the data (Transform) has to be done. The extract phase requires rigorous requirement analysis. We have to decide what to bring into the warehouse based on our type of reports and analytics required. The Load phase occurs in multiple phases. We have to update the DWH on a regular basis to keep the data relevant. The data is stored in a DWH according to a rigid structure that we specify. This helps us to slice and dice our reports and do all kinds of dimensional analysis. The data warehouse allows us to do OLAP (Online Analytical Processing) operations and we can create different types of reports and perform analysis. On top of these we have the option of applying data mining applications for better insights. Even though very useful, DWH has limitations. One limitation of DWH is that we can only bring selected data into it. A significant upfront effort is required to determine what would be stored in a DWH. As our requirements change over time we find it difficult to keep our DWH in alignment with our needs. Another limitation the data is limited to structured data. We cannot have audio, video, tweets, blogs. They don't fit in very well in the current structure of most DWH. Third limitation is that DWH is not instantaneous (batch oriented). In the present times the nature of potentially useful data has changed and DWH solution is no longer enough to cope with it.

The data available nowadays is known by the popular term of Big Data[7,8] and is characterized by 3 Vs. They are volume, velocity, and variety. Volume refers to the size of the data. It is now common to have data sources generating several terabytes of data. For example, in personal capacity, with the widespread availability of smartphones, we are taking digital photos, making videos, surfing, and doing many other things that were not possible even 10 years ago. The next catalyst seems to be the Internet of Things and 10 years from now the data generated can be mind boggling. Velocity refers to how fast data is being produced. The data production rate has consequences for its processing and storage requirements. Common examples include videos being watched on YouTube, the like button being clicked on Facebook all over the world, and tweets being sent. When Internet of Things will become common place data will be streaming in from several places even in a small apartment. We don't have the leisure of doing ETL on it and storing it in a DWH for future analytical processing. Variety means that data can comprise of a variety of forms (video, images, audio, text) and can be unstructured. The data can be in the form of tweets, blogs, and comments. This implies that it cannot be easily stored in a relational format and also that we have to use non-traditional methods to analyze or query it.

All of the above three factors are driving the development of Big Data solutions. An enterprise can start looking for alternate, non-traditional (i.e. non-relational database) solutions when faced with a challenge by any one of these factors. Big Data model also addresses almost all of the limitations of DWH and conventional databases[9]. We can have a lot of data at our disposal (theoretically infinite). The data is loaded without transformation. We store now and use later. It is also not constrained by structured data only. We allow semi and unstructured data. Furthermore, big data allows streaming as opposed to batch update. As soon as a tweet occurs, it will be accounted for.

3. Sources of Big Data for behavioral analytics

Behavior analytics is a complicated endeavor because several factors could influence a student's behavior. These factors include family, friends, habits, and interests. Data pertaining to these factors may not easily be available. Furthermore, it may not be ethical or legal to collect some of the available data. Moreover, the data collection should not entail unreasonable cost and effort.

Now let's have a look at the data that is available in a typical university and can be used to monitor and analyze the behavior of a student.

Traditional Databases: Traditional data sources are the existing relational databases, data warehouses, data marts, or any other information system producing structured data. In this category we have information about students, courses, exam´s marks, etc. There may be databases operational in the university's restaurants, medical centers, gymnasium, worship places, etc. Existing databases may be augmented by those that are missing to include additional information such as schedule of courses, allocation of class rooms and laboratories, opening times of buildings, teachers' office hours, etc. From these databases we can not only find out what courses a particular student is taking but also where he/she is supposed to be on a given time and day.

Personal Data: This data can be in digital or non-digital form. Digital data includes e-mails, cell phone calls, text messages, digital photos, audio material, video material, online purchases, and credit card usage. Non-digital data can be in the form of paper documents, handwritten notes, paper based photographs, newspaper cuttings, etc. Personal data is usually unstructured or semi-structured and is illegal or unethical to obtain until a strong suspicion prompts the authorities to procure it.

Web Digital Trail: Many of our everyday web based actions leave a digital trail. Student web activities can be monitored if they are connected to the web through the university networks and Wi-Fi zones. Record of pages that have been viewed while surfing and meta-data of e-mails are common examples.

The commonly used sources are web mining and text mining (logs from servers), opinion mining from social networks, and data collected from public portals.

A lot of data is publically available from social networking sites such as Facebook, LinkedIn, and Twitter. Examples: photos and videos uploaded, comments, messages, clicking of like button. Such data can be monitored by techniques such as opinion mining. Problems regarding these sources are that data is unstructured and the amount of available data may be too little for an individual to give any meaningful insight. On the other hand, collectively, the data may be huge for all the students combined together.

Outdoor Activities Data: Several sources are in control of the university administration, for example:

- Data about cars entering or leaving a parking area.
- Video surveillance.
- Information from gymnasium, restaurants, university worship places, etc.
- Internet of Things (IoT) data.

Surveillance videos, parking sensors, room access authorization systems, and several other systems are usually found all over the university. Data from these systems are monitored and stored in isolated manner and is usually available for local analysis for detecting security breaches and any other breaking of rules.

Other sources are also present but may not be available for integration in a university owned system. Examples are data regarding cell-phones pinging to the cell towers to check where they are and GPS systems locating a car or a phone. Cell phones ping to the cell towers to check where they are.

An emerging trend is that of Internet of Thing (IoT). It promises to be a widespread phenomenon in a few years. The IoT devices will be economically feasible to be placed in doors, windows and electricity switches. The devices will be present all over the building units and sub-units within a building. This would include gymnasium, car parking, hospital, restaurants, etc. Based on these devices, we will be having systems for automatic turning the lights on or off, automatic usage analysis of rooms and corridors, etc. It will be a step towards having energy saving and Green friendly environment. The data from IoT devices can potentially be used for behavioral analytics also.

Integration of Data: The available data listed above, when present in an integrated manner, can be used to extract models of average (or normal) behavior. We can attempt to monitor a student based on his/her background information, profile, historical performance, demographics, and current interests and activities. For example, from a single tweet or even from a single “like” on a social networking site we can increase or decrease the likelihood of proving a hypothesis that we have made about a person.

4. Capturing and storage of Big Data for behavioral analytics

After having identified the sources of data, we can now turn our attention to its capturing, storage, and analytics. Our data sources indicate that we are dealing with Big Data and therefore require technologies that would allow us to collect, store, and process large amounts of data having a large variety and some of it being generated at large speeds. Furthermore, they should be flexible enough to accommodate incremental additions of functionalities. When dealing with Big Data, a recent popular trend is to use the Hadoop platform[10,11] with the help of cloud computing services[12].

Hadoop: It is a framework that enables distributed processing of a large number of data sets across a large number of computers. Hadoop is a powerful platform for working with Big Data. It consists of many different modules (more than 150).

Hadoop Distributed File System (HDFS): It is a file-system that is distributed, scalable and portable. It is used to store large files (can be in gigabytes and terabytes) across many servers. Thus, Hadoop can have hundreds or even millions of separate files that are spread across many computers (can be in thousands) and all are connected through the software to each other.

MapReduce: It is another critical part of Hadoop that performs the distributed processing. This is a process consisting of mapping and reducing. Mapping splits a task and its related data into many pieces so that they can be sent to several different servers for being processed in parallel. The reducing process takes the results from the different computers and combines them to give a single result.

Pig: It is a platform in Hadoop that is used to write MapReduce programs. It uses its own language, called Pig Latin Programming Language.

Hive: This is a data warehouse within Hadoop. It can be used for queries, data summarization, and analysis. It uses HiveQL (an SQL like language) for queries.

Other components are also available. Among the more commonly used are HBase (a NoSQL database), Storm (allows processing of streaming data), Spark (allows fast in-memory processing), and Giraph (used for analyzing social network data).

Distributed storage & other Cloud Services Over the last few years, cloud storage, or storage over the internet, has become a preferred method[12]. It is attractive for Big Data because the storage space is scalable. It can be rented on an as-needed basis. There is no real up-front cost for installation or maintenance, and the level of redundancy that cloud providers can offer makes data nearly impossible to lose or damage. It is also possible to tailor cloud storage for things like scalability (how much storage do you need), redundancy (how many backup copies do you want), and speed (how fast you want to be able to pull information out). One of the more popular cloud storage vendor is Amazon. They have the Simple Storage Service, better known as Amazon S3.

Online storage is one element of a larger group called Infrastructure as a service (IaaS). In fact, the cloud service providers are able to provide many more computer related services known by the umbrella term of cloud computing. Infrastructure as a service (IaaS) is an online version of the physical hardware of computer. It can be thought of as a hosting layer. It includes disk drive, servers, memory, and network connections. Another service, Platform as a services (PaaS) can be thought of as the building layer. It includes the middle layer software, like the operating system, components (like the Java run time) that allows higher level software to run. It also gives access to databases like Oracle and application servers like Microsoft IIS. Software as a service (SaaS) can be thought of as consumer layer of cloud computing. It includes web applications that run entirely through the browser like Gmail and Office 365. Data as a service (DaaS) provides access to data which the data providers make available. Query as a service (QaaS) allows users to query the data that they have uploaded. It is as simple as using a web based search engine. The users pay for queries in addition to the payment for their data storage. Google's BigQuery is such a service and can be accessed from Google Cloud. An Apache project called Apache Drill is another example.

All the cloud computing services can play important roles in Big Data projects. They can provide the physical resources necessary to store and process the data, the software to interact with the data, or even the data itself. What they all have in common is the ability to shift the load off the consumer business and allow them instead to dedicate their own resources, their money, space, time, and energy to working with the data and getting the insight they need for their own projects and progress.

5. Behavioral analysis using Big Data

Several analytical tools have been well established. These include data mining, text analytics, web analytics, and predictive analytics.

- Data mining[13] is about finding unexpected patterns in data. These patterns might include unexpected associations between variables or people who cluster together in unanticipated ways.
- Text analytics is also a kind of data mining[1]. Its goal is to take the actual content of text data and find meaning and patterns in the words. Within text analytics, one of the most common tasks is sentiment analysis (determining how people feel about something).

- Social network analytics examines network dynamics, identifies influential entities, and finds out interesting patterns of activities. Interestingly, the monitoring of criminal networks by law enforcement agencies are part of this.
- Predictive analytics use a range of techniques (e.g. neural networks) that try to predict future events based on past observations.

The collected data and the analytical tools can be all brought together and unified to answer a question, e.g.: "Is student X drifting towards being a terrorist"? This question can be converted to a well formulated hypothesis and our analytical effort would be to prove it or disprove it for each student on a perpetual basis. The workflow of this analytics would be detection of events followed by their processing through analytical models. The analytical models should give a likelihood of the hypothesis being true for a given student. If this likelihood is above a certain threshold, certain further actions can be taken (closer monitoring, guidance, warning, inspection of personal belongings, and so on).

In our case we know what we are looking for and we need a notification from the system when any event, or combination of events, occur that point towards that phenomenon. However, we need to specify the events and the criterion that would trigger such a notification. To understand this concept, a simple example is that of monitoring the physiological signs of a person (temperature, pulse, white cell counts, etc.) and raise an alarm if they fall outside the normal range under a given situation (e.g. a person sitting in a relaxed manner on a chair in an un-agitated mental state).

Because life and its events are complicated, we can also aim for anomaly detection in addition to monitoring of specific events. Anomaly detection is when a user wants to know when something unusual happens. We are looking for unusual activity without necessarily knowing what that might be.

A final precaution worth noting is that the quality of the incoming data and its relevance are as important as the reliability of the analytical models to avoid chaos instead of tranquility arising from this system.

6. Conclusion

This paper gives a new meaning to behavioral analytics and expands it from the limited realm of websites and E-commerce. We argue that enough data is available in a university environment that can be harnessed with the help of Big Data model and accompanying technologies to monitor and predict deviant behavior in students. In this paper, various sources of data are identified and a discussion regarding capturing and storage is presented. Opportunities and work flow for Big Data analytics are also discussed. The main characteristic of the whole idea is that it is built upon the data capturing and storage that is already being done in a university in one way or another for other purposes. The key concept is the available data's integration and unification for a special purpose.

References

1. Banerjee S, Agarwal N. Analysing collective behaviour from blogs using swarm intelligence. *Knowledge and Information Systems* 2012;**33(3)**:523-547
2. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. *Computational Science* 2011;**2(1)**:1-8
3. Cen L, Ruta D. Ng J. Big education: opportunities for big data analytics. In: *Proc. IEEE Intl. Conf. on Digital Signal Processing*. 2015. p. 502-506.
4. Sharda R, Delen D, Turban E. *Business intelligence and analytics: systems for decision support*. 10th ed. Pearson/Prentice Hall; 2015
5. Michalik P, Stofa J, Zolotova I. Concept definition for Big Data architecture in the education system. In: *Proc. IEEE 12th Intl. Sym. on Applied Machine Intelligence and Informatics (SAMI)*. 2014. p. 331-334.
6. Siegel E, Davenport TH. *Predictive analytics: the power to predict who will click, buy, lie, or die*. John Wiley & Sons; 2013.
7. Eaton C, Zikopoulos P. *Understanding Big Data: analytics for enterprise class Hadoop and streaming data*. McGraw-Hill; 2012.
8. Hurwitz J, Nugent A, Halper F, Kaufman M. *Big Data for dummies*. John Wiley & Sons; 2013.
9. Simon P. *Too big to ignore: the business case for Big Data*. John Wiley & Sons; 2013.
10. Lam C, Davis M, Gaddam A. *Hadoop in action*. 2nd ed. Manning Publications; 2016.
11. Schneider R. *Hadoop for dummies*. John Wiley & Sons; 2012.
12. Vaquero L. EduCloud: PaaS versus IaaS cloud usage for an advanced computer science course. *IEEE Trans. on Education* 2011;**54(4)**:590-598
13. Wu X, Zhu X, Wu GQ, Ding W. Data mining with Big Data. *IEEE Trans on Knowledge and Data Engineering* 2014;**26(1)**:97-107