7-17-2017 10:00 AM

Data Science Solution for User Authentication

Anas Ibrahim, *The University of Western Ontario*

Supervisor: Dr. Abdelkader Ouda, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Engineering Science degree in Electrical and Computer Engineering

Recommended Citation

Ibrahim, Anas, "Data Science Solution for User Authentication" (2017). *Electronic Thesis and Dissertation Repository*. 4871.
https://ir.lib.uwo.ca/etd/4871

Abstract

User authentication is considered a key factor in almost any software system and is often the first layer of security in the digital world. Authentication methods utilize one, or a combination of up to two, of the following factors: something you know, something you have and something you are. To prevent serious data breaches that have occurred using the traditional authentication methods, a fourth factor, something you do, that is being discussed among researchers; unfortunately, methods that rely on this fourth factor have problems of their own.

This thesis addresses the issues of the fourth authentication factor and proposes a data science solution for user authentication. The new solution is based on something you do and relies on analytic techniques to transfer Big data characteristics (volume, velocity and variety) into relevant security user profiles. Users' information will be analyzed to create behavioral profiles. Just-in-time challenging questions are generated by these behavioral profiles, allowing an authentication on demand feature to be obtained. The proposed model assumes that the data is received from different sources. This data is analyzed using collaborative filtering (CF), a learning technique, that builds up knowledge by aggregating the collected users' transaction data to identify information of security potential. Four use case scenarios were evaluated regarding the proposed model's proof of concept. Additionally, a web based case study using MovieLens public dataset was implemented. Results show that the proposed model is successful as a proof of concept. The experiment confirms the potential of applying the proposed approach in real life as a new authentication method, leveraging the characteristics of Big data: volume, velocity and variety.

Keywords: User Authentication, Big-Data Analytics, Knowledge-based authentication, Recommender Systems, Collaborative Filtering, Security.

Dedication

This thesis is dedicated to my mother for her constant encouragement and support throughout my life and educational career. I also dedicate this thesis to my sister, who has been my best friend, where she stood by my side whenever I needed her. I thank you for the love, support, and unwavering belief in me. Without you, I would not be the person I am today.

Acknowledgments

This thesis is the result of a hard work and very busy research program, which would not have been possible without the support of a many of people.

I am very thankful to Prof. Abdelkader Ouda, who provided thorough and helpful support with a close guidance throughout my masters’ research program. He has given me invaluable insight into many areas of study relating to software engineering. He has been my supervisor and my friend.

Many thanks to my friends and colleagues for their support and help during my masters’ research program.

List of Abbreviations

NIST	National Institute of Standards and Technology
DSA	Data Security based Analytics
BDA	Big data-driven Authentication
JitHDA	Just in time Human Dynamics
SaaS	Software as Service
FP	Frequent Pattern
KBA	Knowledge-based authentication
PR	Page Rank
RS	Recommender Systems
CF	Collaborative Filtering
SR	Social Recommendation
HF	Hybrid Filtering
(PR)	Precision and Recall
IDA	Innovative data authentication model

Table of Contents

List of Figures

List of Tables

Chapter 1

1. Introduction

User authentication is considered a key factor in almost any software system and is often the first layer of security in the digital world. NIST defines electronic authentication as the process of establishing confidence in user identities electronically presented to an information system [19]. Authentication is a process that consists of two steps: (1) Identification step, where presenting and identifier to the security system, and (2) Verification step, where presenting or generation authentication information that corroborates the binding between the entity and the identifier [20]. However, all systems that have been designed and implemented haven't proven to be fool proof against attacks and as a result users' information and money has been stolen.

Therefore, this shows the need to find and explore different models that overcome the security weaknesses the current approaches have. To help alleviate the perceived issues of authentication systems, it is worthwhile to investigate existing security issues in current authentication systems as well as how those issues have been resolved. By applying the experience gained from studying authentication systems, it will be possible to address the concerns and requirements of users, to ensure the privacy of data, as well as money.

Our work proposes a Big data based user authentication as a new approach that leverages the power of Big data analytics to develop a fertile field for the next generation user authentication. This new approach relies on “something you do”-based verification methods, where the users’ dynamic behaviors are analyzed in order to generate real-time uniquely identifiable information to generate unique dynamic challenging questions to authenticate them. Hence, creating an “authentication on demand” system that will be used when needed.

1.1 Hypothesis

The hypothesis is that a new solution should be proposed because the current authentication methods aren’t foolproof against attacks. To confirm whether the hypothesis is true or not, we should investigate the current authentication methods, and prove the needs of a new authentication model. The next step is to attempt to develop a new model that will address the concerns of users and meet the security standards of digital systems.

1.2 Methodology

To achieve the desired authentication model, different steps will be taken. First an investigation of current authentication models will be completed focusing on systems and approaches that are already in use in the real world. The authentication systems will be analyzed for their strengths and weaknesses as well as addressing those with weaknesses. A new approach and model will then be developed that will meet the needs of the digital systems. The developed model will then be analyzed with existing theoretical tools. Once the model has been successfully analyzed and shown that it meets the desired goal, we can be certain that it has met the stated hypothesis. The new model will also need to ensure the security and privacy of users' data. It will need to handle Big data, clean the dataset, create user profiles, process and analyze the data received, identify information of security potential and generate dynamic challenging questions. The model will need to show that it achieves the goal in authenticating users while being able to handle users' data and profiles, and being simple and user friendly.

1.3 Contribution

This thesis proposes a new solution to the issues brought about by leveraging the characteristics of Big data and the use of learning techniques. The solution focuses on developing a secure, user friendly system by analyzing users' information and classifying them into behavioral profiles with the use of learning algorithms. The new solution is also brought about by the integration of learning techniques to be used in identifying information of security potential, that can be used in generating the challenging questions. The proposed authentication model will be developed, so that for each of these generated profiles, a real-time (i.e., just in time) set of challenging questions will be prepared. These questions cover all actions that explicitly represent an instantaneous specific user's behavior. the novelty of this approach is that these questions will be chosen in a way that the security and usability requirements are maintained. "Security" because each question is issued only once to protect the users' responses from being compromised. This overcomes the pitfalls of knowledge-based authentication (KBA) methods, described in the literature section. Note that, users' profiles are generated from endless source of users' information thereby the uniqueness of the questions is possible. "Usability" because Just-in-time capability guarantee the data being fresh to help the legitimate user to easily remember and successfully complete the challenge.

1.4 Thesis Outline

The remainder of the thesis is organized as follows:

- Chapter 2 provides a literature review about current authentication systems being used with their strengths and weaknesses analyzed. This chapter also provides a background of the related work done on the 4th factor of authentication and systems that dubbed this factor. An analysis and investigation was done to show why we needed to find a new novel solution for user authentication.

  Chapter 2 also introduces the framework for next generation user authentication that is useful in understanding our proposed work, and explains briefly how our proposed work relates to its second component.

- Chapter 3 describes the first solution proposed, the IDA model. It first presents the workflow of the proposed model and then describes the data types, behavioral characteristics that define how the system will behave. Furthermore, chapter 3 discusses about the adaptable mechanisms and machine learning techniques that will be used to build up knowledge and classify profiles. A general discussion is then presented with a use case study to demonstrate how the system might work if implemented.

- Chapter 4 starts by presenting related work on recommender systems and how the investigation of RS led us to seek collaborative filtering as part of the second solution proposed as hybrid based filtering approach for user authentication. The second solution is then discussed in more details, with the three main components of the proposed system explained.

- Chapter 5 shows how the proposed solution can be put to use by demonstrating different use case scenarios that could take place in the real world. It also presents a case study where an experiment has been done to show an implementation where the proposed approach can also be used.

- Chapter 6 concludes this thesis and offers future research directions and suggestions.

Chapter 2

2. Background

Digital authentication has always been used to establish confidence in user identities electronically presented to an information system. To address this matter and users' concerns about their privacy, we need to properly understand how the current systems work and overcome their weaknesses. This will help develop acceptable systems that will meet the users' and clients' needs in digital authentication. Overcoming these weaknesses will create a more secure system and will help protect confidential information.

The objective of this chapter is to provide background information about the current systems in use with their weaknesses analyzed, which made us seek this topic to find a new novel solution.

2.1 Literature Review

Authentication has always been a key factor in any software system. The different authentication models that are used are classified based on three factors, "Something you know", "Something you have", and "Something you are" [21], as presented in figure 2.1. Password authentication method is the most common method of something you know. It is the simplest and easiest mechanism that can be installed over an insecure network. Yet, it is also known to be the weakest type of security measure for authentication. “Something you have” model is based on something the user has, such as smartcards and security tokens. Users have to present what they possess to an information system to prove his/her identity. However, cards can be easily duplicated, and security tokens can be stolen [47]. “Something you are” is based on biometric traits verification that vary between fingerprints, retina scan, palm prints, facial recognition, speech recognition, to walking posture [22]. The problem that faces biometric authentication systems is the high cost of equipment needed, and that human traits are impossible to be replaced if compromised. And in the newest smartphones, fingerprint security is considered one of the weakest security measure for a mobile phone, where researchers were able to bypass systems and fake users’ identities [25][50] [51].

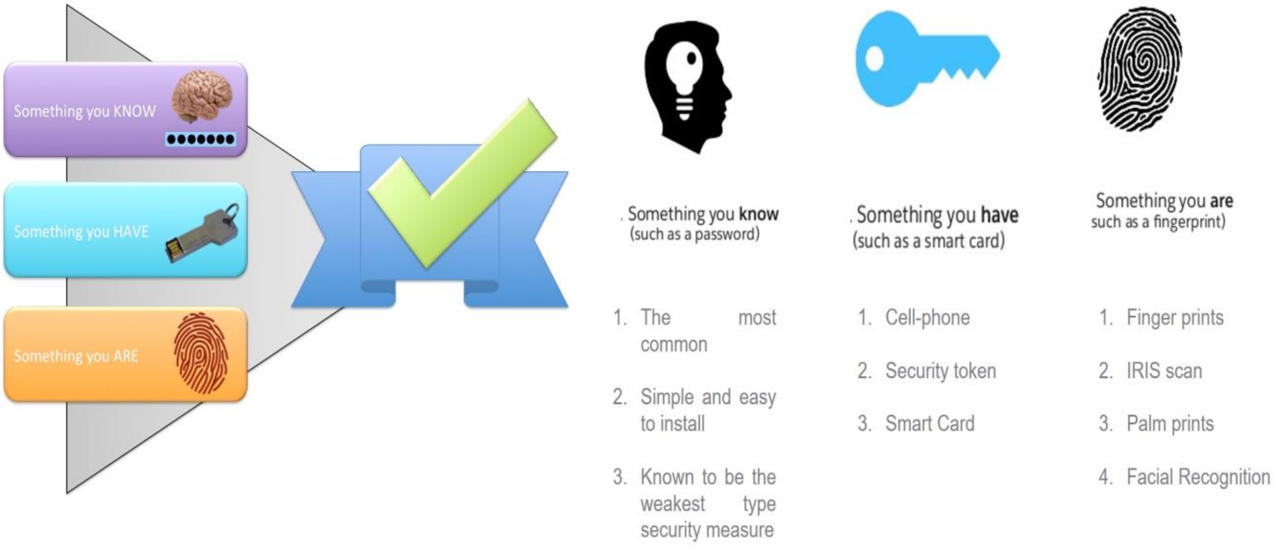


Figure 2.1 The three authentication factors

All three models are susceptible to breaches. In May, e-Bay went down in a blaze of embarrassment, where hackers had managed to steal personal records of 233 million users, with usernames, passwords, phone numbers and physical addresses compromised [27]. In 2008 a denial of service (DoS) attack cost Amazon around $3.6 million when their servers were down for 2 hours [23]. With the new fingerprint technology used in mobile phones, a demonstration showed that the phone's biggest vulnerability is the fingerprint [25][50][51]. On the other hand, multi factor authentication seemed to provide stronger security, yet attackers found ways to bypass such systems. In 2013, $46.5 million were stolen from a bank that relied on two factor authentication [24].

A new factor of authentication "something you do" is introduced as an alternative to the first three factors [21]. In this regard, research is being carried on user dynamic biometric, where user's behavioral patterns are studied. Behaviors such as voice pattern recognition, handwriting, and typing rhythms. University of Regensburg collaborated with Germany based company Psylock GmbH [26] and released an authentication software that focuses on person's typing behavior. Thus, verifying a user based on their typing behavior on a computer keyboard. Valuable work has been done in designing, implementing, and evaluating the dynamic biometric-based authentication techniques [31][32][49] that promise to provide secure user authentication. Yet these methods have issues in either scalability or performance requirements. A second system that uses users' behaviors is multi factor authentication system-MACA. A user has to use username-password mechanism as a first step and then a user profile is built with information captured about him/her where it will be used to assess the difference in the information during a second log-in attempt. The system uses two factors that are not secure. A username-password approach is the weakest system as shown previously, and the second approach is susceptible to Denial of Service attack.

On the other hand, RSA presented risk-based authentication system; an implementation for “something you do” [33]. Risk based authentication system assesses risk score for users trying to access the system. Once a risk score exceeds a certain limit, a user may be required to provide a higher level of authentication. But risk-based authentication system hasn’t proven to be foolproof to Denial of Service (DoS) attack [28][48]. As a result, this shows the need to find and explore different models that overcome these security weaknesses.

Another implementation approach for “something you do” commonly referred to as Knowledge-based authentication (KBA) that is mostly used in financial institutions or websites [39]. KBA is an authentication scheme in which the user is asked to answer at least one “secret” question. This question is brought using either static or dynamic methods. In static methods, the question is chosen from a pre-defined set of question/answer pairs, example; “where did you spend your honeymoon?” or “Who was your favorite teacher?”. However, in dynamic methods, the question is generated from information within a person’s credit history or public records, example; “What is the credit limit of your credit card ending 1234?” or “What was your street address when you were 10 years old?” It is obvious that, if someone has shared that information on a social media site, the answer can be easily guessed.

Recently, a new user authentication framework has been proposed [10], where the author presented a new perspective in utilizing the authentication factor “something you do”. This framework leverages the characteristics of Big data to provide accurate patterns of users’ behaviors. Where Big data is defined by the 3Vs; volume, velocity and variety [40][41][43]. And Big data generally describes datasets which cannot be perceived, stored and processed by classical approaches and technologies within tolerable time [43].

Therefore, we analyzed the related work that will be discuss in more detail in section 2. The first direction sought resulted in proposing a model that studies the behaviors of users and classifies them into behavioral profiles. This will be discussed in more detail in chapter 3. While the second direction sought lead us to approach the problem in a slightly different way. The second solution proposed relies CF which is similar to the technique used in recommender systems where real life applications take advantage of machine learning; that is defined as the field of study that gives computers the ability to learn without being explicitly programmed". And now it is the combination of several disciplines such as statistics, information theory, theory of algorithms, probability and functional analysis [30][45]. This is discussed in more details in chapter 4.

The following section introduces and discusses the new framework for next generation user authentication where the concept behind our work was built on.

2.2 Big data-based Authentication Framework

Recently, a framework for next generation user authentication has been proposed [10]. This framework investigates the development of a new authentication system that utilizes Big data analytics based on "something you do" verification. Analytic tools will be used to study and analyze changes in user behavior and data. This analysis will allow to generate real-time information about users. Whenever identifiable information is captured, organizations will be able to authenticate users and provide the decisions as a service to different internet-based applications. Figure 2.2 shows the components of the framework for the next generation user authentication.

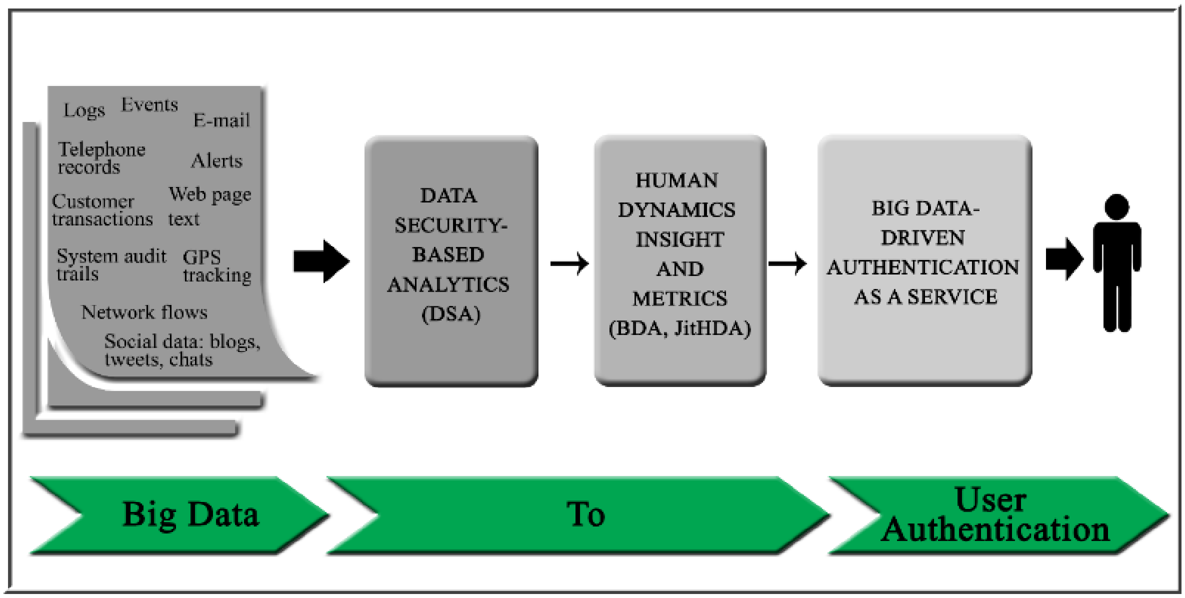

Logs
Events
E-mail
Telephone records
Alerts
Customer transactions
Web page text
System audit trails
GPS tracking
Network flows
Social data: blogs, tweets, chats
DATA SECURITY-BASED ANALYTICS (DSA)
HUMAN DYNAMICS INSIGHT AND METRICS (BDA, JitHDA)
BIG DATA-DRIVEN AUTHENTICATION AS A SERVICE
Big Data
To
User Authentication

Figure 2.2 The main components of Big data-based user authentication framework

2.3 Data Security-based Analytics (DSA)

Organizations capture and analyze big datasets in real-time, but to be able to do that, organization will have to use tools that manage Big data's 3Vs (velocity, volume and variety) [40][41][43]. Volume or the size of data, where the grand scale and rise of data outstrips traditional store and analysis techniques. Variety makes Big data really big, where data comes from variety of sources and has three types: structured, semi structured and unstructured. And velocity is required not only for Big data but also all processes. Big data should be used as it streams into organization in order to maximize its value [40][42][43]. As such, the first component of this framework; Data Security-based Analytics (DSA). DSA leverages large scale data processing engines such as Spark [36] and Hadoop [37], utilize Splunk and Hadoop power, and define criteria to provide real-time identification of data with security potentials.

2.4 Human Dynamics Insight and Metrics (BDA, JitHDA)

The outcome of DSA will be clustered into human dynamic based information, where profiles will be created based on actions of humans and their behaviors. This task is managed by the second component of the framework, the Big data-driven authentication tool (BDA). BDA tool assembles human dynamics into security user profiles. These profiles will be used as security profile input to a just in time human dynamic authentication engine (JitHDA). JitHDA is main tool in BDA component, where it will generate random set of challenging questions, with one question relate to actions performed by a user. the purpose of the randomness is to create a certain level of uncertainty. The randomness or uncertainty will prevent any attacker from acquiring knowledge about which question relates to a certain user, which in return will protect user privacy. This technique will allow BDA tool to perform a security just in time challenge to authenticate the user [10].

2.5 Big data-driven Authentication as a Service

The third component of the framework is a tool that promotes Software as a Service (SaaS) authentication tool. Since the framework relies on Big data, this will not only allow to authenticate organization’s own users but also to authenticate users for other applications on the cloud.

2.6 Summary

This chapter provides first a literature review and background about current authentication systems and their weaknesses. It also presents the new factor of authentication “Something you do”, that has been dubbed in recent work. We have investigated and analyzed the systems that relies on the new authentication factor and presented their weaknesses, to show why we need to investigate a new novel approach for user authentication. Moreover, we presented the related work of the new next generation framework, and showed why it is important in our proposed work.

Chapter 3

3. Innovative Data Authentication Model

Big data has revolutionized the way the world uses data. An important process used is to study human behavior and provide ads, recommendations based on what a user is looking for, searching for, and watching when using the digital world.

The traditional method of authentication in computing is the challenge-response mechanism. The investigation of Big data and behavior showed that we are able to take advantage of Big data characteristics, that are volume, velocity and variety to recognize patterns and identify behavioral characteristics. The behavioral characteristics can be used as a shared secret between two parties, so that one party asks a question as a challenge and the other party must reply with a correct answer as a response.

The first direction that have been addressed during our investigation of the big-data based authentication framework lead us to propose an innovative data authentication model (IDA). A model that leverages the characteristics of Big data and learning techniques such as, association learning and classification to create behavioral user profiles and identify information of security potentials.

3.1 IDA Model Workflow

The first proposed solution is the result of the investigation of the second component of the Big data-based authentication framework described in chapter 2 section 2.2. The proposed model aims to recognize users' information and classify them into behavioral profile. The classification done will be used in generating challenging questions to authenticate users.

There are two main activities involved to achieve this goal; (i) create user profiles to study human dynamics and behavior, analyze user behaviors and actions, and classify user behavior accordingly, (ii) generate questionnaire on the fly to authenticate users.

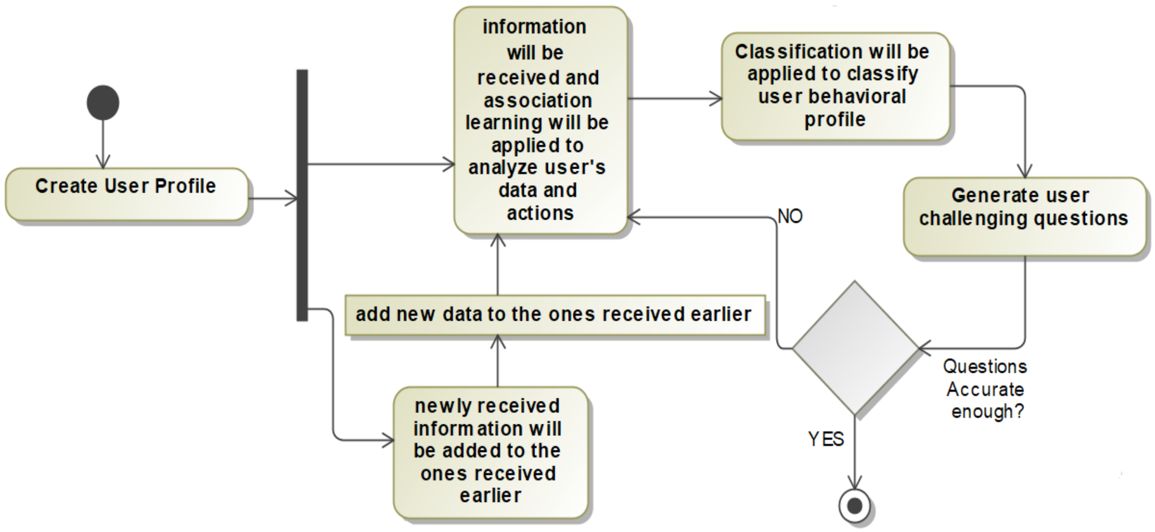

Create User Profile
information will be received and association learning will be applied to analyze user's data and actions
Classification will be applied to classify user behavioral profile
Generate user challenging questions
add new data to the ones received earlier
newly received information will be added to the ones received earlier
NO
Questions Accurate enough?
YES

Figure 3.1 IDA model workflow

Figure 3.1 shows the activities in IDA, where information will be received from DSA, so that a user profile will be created that contains data of security identification potentials. After profiles are created, a learning algorithm will be applied to start with the process of data analyzing and classification. The first algorithm is association learning, it will be used to analyze users' information and recognize patterns about users' behaviors, dynamics and actions, and to build up knowledge to help categorize users' personalities into different personality types. Personality type is the collection of characteristics and traits that are determined by specific pattern of human behavior, and it is distinguished by the behavioral tendencies people have and actions they make [46]. Research has shown that personalities vary between 16 types from adventurer, entrepreneur to strategic, extrovert and others [29][46]. Once actions and behaviors are analyzed, classification algorithm will be used to categorize user profiles into behavioral profiles that identify behavioral characteristics.

The next step that follows is generating user unique, non-repeated challenging questions to authenticate users. Questions consists of real time events and situations that reflect recent behaviors of users. The questions generated should be accurate to proceed to the authentication process, if not a revised version will override the older one by iterating through the whole process again; analyzing information, classifying profiles and then generating challenging questionnaire. A certain threshold for questions and information accuracy is set to detect whether the information analyzation process should be repeated or not. The threshold is set according to the percentage of occurrence of a certain behavior(s) over time that can be used in security identification, such as an irregular behavior made during a certain day of the week, or a comparison of two orders that arrived on different dates when purchased at the same time. The advantage of the unique, non-repeated questions is that they are not susceptible to social engineering [38], where this method is used to bypass security systems using static general questions. For example, questions used in email verification process and account login as additional type of security measure. Social engineering is a common way intruders use to retain information to bypass security systems and customers' accounts that use a static security measure. It is the art of exploiting the weakest link of information security systems; the people who are using them [38]. Users are manipulated through various methods to release information that results in intruders performing unauthorized actions.

Since data is changing and dynamic; building knowledge through association learning, user profiling and classification, and generating challenging questions, makes the relation between the three tasks interrelated.

The following section describe the detailed components of the above model. In section 3.1 creating and customizing behavioral types will be discussed, explaining the behavioral classification that will be used, and section 3.2 discusses the adaptable mechanism with machine learning that will aid in user profiling to help in generating security questions in real time.

3.2 Behavioral types and Big data analytics

Many systems utilize user profiles for different purposes. For example, one primary and main uses of user profiles are in recommendation systems [35], that rely on the searching habits of the users. Organizations like Facebook use user profiles to find potential friends based on relationships and groups joined, while LinkedIn takes advantage of skills and professional information stated by users' profiles to recommend potential employees/employers. Behavioral and personality profiles are studied to reach an accurate assessment of persons' behavioral characteristics. The profiles are studied since they reflect an individual's attitude toward the external world and the direction of general interest. Most of the known personality types are categorized according to Carl Jung's research [29][46]. C. Jung first developed the theory that individuals each have psychological type based on the behavioral tendencies they have. This type of study increases the ability to classify user profiles based on relevant information about people, related and recognized patterns in users' actions and behavior.

The following subsection shows and describes in more detail the type of data that system will analyze in order to move to the next process of creating behavioral profiles.

3.2.1 Data types

User profiling has been used in many systems from social networks, to recommendation systems, which shows that is effective and useful analyzing users' data. However, to create the behavioral profiles, different types of data should be analyzed. For our proposed model, behavioral profiles depend on variety of information that identify each individual. For instance, IP address(s), device profile, email address(s), location(s) and check-ins, calendar, web surfing history and mobile activity, purchase activities, movie(s)/TV show(s) ratings, medical history, languages used, activity based on time stamp and time zone, news feeds and follow-ups; that vary from sports, technology, politics, business, to arts, style, food, fashion, etc.

- ***IP address***: on the Internet, one geolocation approach is to identify the subject party's IP address, then determine what country; including down to the city and post/ZIP code level

- ***Location***: Sharing location is highly important. It reveals different routes a person takes on daily basis, or various places and/or events a user attends. This shows the importance of using location in security identification.

- ***Web history:*** Refers to the list of web pages a user has visited recently—and associated data such as page title, keywords searched, and time of visit—which is recorded by web browser software as standard for a certain period of time. Which all can be used to generate security questions to authenticate users.

- ***Purchase activities:*** shows what stores a user go to, what brands a user likes, and marketing ads that interest him/her.

- ***TV shows/Movie/ online streaming***: indicates what type of movies and TV shows that a user likes, what is his/her taste in genre, such as action, adventure, drama, comedy, etc. And what topics did the online streaming search include, which can be used in the authentication process.

- ***Medical History:*** may include information about any sort of medication a user is taking, disease a user is dealing with. This can be used in the authentication process whenever questions are generated.

- ***News feed:*** varies from sports, politics, technology, to science, news, arts, or any sort of news feed user follows or subscribes to. With which data can be used to recognize what a user is interested in.

3.2.2 Behavioral Types and Classification

Research has been done by scientists to allow them to accurately categorize personality, character, and behavioral types of people into different categories [29][46]. Research and studies have been done by C. Jung who first developed the theory that indicates that individuals each have psychological type based on the behavioral tendencies they have. This type of study increases the ability to categorize user profiles based on relevant information about people, related and recognized patterns in users' actions and behavior.

For our proposed work, we have specified different behavioral types; based on the different personality types studied in [46], to match our needs in building behavioral profiles. These types will help in building up knowledge about each user using the adaptable mechanisms of machine learning; that will be discussed in section 3.3, to help identify irregular behavior of users. The following categories describe each type of behavioral category that will be used in the first proposed approach.

- ***Extravert/social:*** individuals with such personality type are social, open, outgoing, has numerous contacts with others, and prefer communicating in groups. They are often seen in events, conferences, parties, workshops, etc.

- ***Introvert:*** individuals energized by solitary, focused on their jobs, prefer one on one conversations instead of group communication. Also individuals with such behavior prefer quietness to finish their work. study, research, etc.
- ***Athletic/Sporty:*** refers to individuals who tend to have behaviors that relate to sports, such as attending sport games, follow sport feeds, have gym or any sport subscriptions.
- ***Adventurer:*** refers to those who tend to have adventurous behavior, such as going on trips, fishing, camping, buying tools that relate to such things. It would also include individuals who tend to try new things, such as new cuisine for example.
- ***Smoker/non smoker***
- ***Workaholic/not:*** refers to individuals that spend too much time working. It can be determined by how much extra time they spend at work weekly, and if they still follow up on work while they are off shift or on a vacation.
- ***Shopaholic/not:*** includes individuals that are considered to be addicted to shopping. It can be shown through the number and type of purchases made within a specified period of time.
- ***expert/fan:*** depends on interests and the time spent in doing a task or activity. It can be revealed based on when an individual started a particular behavior/task/action, such as work, hobby, cuisine, attending motor shows, etc.
- ***Student-researcher-worker:*** it depends on the type of occupation an individual has. For instance, attending School X, researcher at school X, lab Y, or any sort of facility. Worker depending on job, might be blue collar, white collar, knowledge worker/intellectual.

The following section will present the adaptable mechanism that will be used. As part of adaptable mechanism, machine learning will be discussed in more details to show how it will be used to achieve our goal in building up knowledge and classify user profiles into behavioral profiles to identifying information of security potential.

3.2.3 Adaptable Mechanism and Machine Learning

Data mining is the process of processing huge amount of data to discover insights and build knowledge from big datasets [52]. The goal of data mining is to extract information from a data set and transform it into understandable structure for further use. Machine learning is the field of study where computers are able to learn without being programmed [45]. And it is a combination of several disciplines such as statistics, information theory, theory of algorithms, probability and functional analysis [45]. An example often cited is the algorithmic classification of email into spam and non-spam messages without user intervention. In the context of user profiling, machine learning can be used for learning user behavior by identifying patterns. Machine learning is used to explore and analyze dataset to produce useful, reliable results through learning from hidden relationships. Different types of algorithms have been proposed for machine learning and Big data analytics. Such as, clustering, association learning, and classification [36].

The following subsections will discuss in more details the learning techniques that will be used in the proposed solution.

3.2.3.1 Association Learning

Through different machine learning techniques, useful data can be aggregated to produce accurate challenging questions for authentication. One of the learning techniques that will be used is Association learning. Association learning is a method for discovering interesting relations between variables in big datasets. For example, the rule “dough, cheese” found in the sales data of a supermarket would build the knowledge that if a customer buys cheese and dough together, s/he is most likely to also buy pizza sauce.

Different association learning algorithms has been proposed to recognize frequent patterns in different data sets. For example, using FP-Tree algorithm [37] [53]. The frequent-pattern tree (FP-tree) is a compact structure that stores quantitative information about frequent patterns in a database. FP-tree algorithm starts by one scan of database to identify the set of frequent items. The FP-tree as the tree structure is defined as follows:

1 Iterate through user profile and specify the minimum occurrence of set of actions

2 Check for specified patterns that relate to behavioral actions

3 Count and find the frequency of occurrence of the specified patterns

4 Prioritize the patterns and actions based on occurrence

The FP growth algorithm is used to mine the data analyzed and find a complete set of frequent patterns to identify the behavioral actions of a user.

The output of association learning, will be the input of data classification learning method. Data classification plays a role in categorizing users' behavior actions into behavioral profiles. Section 3.2.3.2 will present more details about classification.

3.2.3.2 Data Classification

Data classification is the problem of identifying to which of categories an observation belongs. For example, the system might assign a user one or a combination of classifications such as worker, adventurer, athletic, etc.

Different classification algorithms (C4.5, KNN, neural networks, ID3, etc.) can be applied to serve our goal in categorizing data, where the classification process will take place as follows:

Iterate through user profiles and set a minimum group priority

2 Check the output of the association learning for the patterns identified

3 Link/group the patterns into categories

4 Count the frequency of each occurrence in each group

5 Prioritize the groups and actions based on occurrence

6 Compare group priority to the minimum threshold set and order them accordingly

7 Classify profiles according to group priorities

The process of classification is used to help determine behavioral profiles in order to help the system identify information of security potentials. As a result, the irregular behaviors that have been flagged will be used to generate unique challenging questions as a challenge for a user to pass.

The following section presents a general discussion of our proposed work. The discussion will walk through the general process of IDA model, and will show how it answers the hypothesis that has been presented.

3.3 General Discussion

To understand the general work flow of the proposed work it is worthwhile to review what the outcome we are looking for. Data is fast changing, large in volume and diverse, thus resulting in a dynamic model. Fast changing and diverse data allows updating and generating challenging dynamic questions to challenge users. The questions that will be generated will be simple, direct and straight forward, rather than the static questions that are used in current authentication systems. For example, sample of questions that are generated for a social, extravert worker classified profile, would be as follows:

1. What are the names of restaurants you went to twice in the last two months on “Richmond” street?
2. Name two to four cities where the organization you work for opened offices in during the last month?
3. Name a café/restaurant you made a purchase from two days ago, that is located within two blocks from your work location.
4. What website you used to place an online order using a visa credit card on the day Blue Jays game took place on July 18th?

Every Behavioral profile will have a set of unique non-repeated challenging questions that change with time. In order to authenticate a user, s/he should be able to identify the actions and behaviors s/he made during a specific period of time, and answer the question(s) accordingly. So, an intruder who is trying to access an account will have to answer dynamic challenging questions that are generated on regular basis, through which only the user knows their answers.

The below table shows how dynamic data can be used to link user information and user classification and authenticate users.

Table 3.1 IDA behavioral analysis

	IP address	**Location**	**Web browsing history**	**Purchase activity**	**Medical history**	**TV Shows/ online streaming**	**News feed**
Extravert	x	x	x	x			x
Introvert							
Athletic/Sporty		x				x	x
Adventurer			x	x		x	
Smoker/non					x		
workaholic		x					
shopaholic							
Expert/fan			x	x			x
Student/researcher/worker	x		x				
Patient diagnosed with illness			x		x		

The analysis of the table shows how different profiling information is integrated with profile classification. For example, five profile attributes reveal that this user is extravert, into sports or sporty and adventurer. IP address and location shows that s/he attends different events in different places. Assume that, an IP address is recorded during mobile phone activity and its location is identified through the signal captured by phone carrier’s towers on the last weekend of June, revealed that a user called Mike attended London festival from noon till 4 PM, and he spent the evening at a well-known restaurant in downtown, then afterwards the night was spent at a mud spa “Novo Spa” that is known to be a place where people have some leisure time after a long weekend. Secondly, the user’s news feed and location attributes show that Mike is into sports. For example, Mike has attended different sport events during the past 2 months when the “Toronto Raptors” had their game against Cleveland; where his location was stored by cell towers. On the other hand, the news follow-up included a track of various sport magazines, newsletters, and basketball games. Moreover, web browsing history, purchase activity, TV shows and news feed showed that the user is interested in different groups that arrange outdoor activities, like camping, fishing, food tasting, painting and photography workshops, etc. As an example, the user watched a series of YouTube videos about wild life, setting up tents, protection against bears, and videos about water painting, portraits and landscape pictures, and so on. Studying and analyzing a table that links user data to user classification will contribute in the challenging question generation process. For example, from the below table different information of security potential can be used in asking about different places the user attended, what type of sports the user likes, what kind of purchases the user made (cloths, books, tickets for different events, equipment, etc.). All of which can be used in the authentication mechanism.

3.4 Use Case study

An innovative data authentication model (IDA) has been proposed as an outcome of the investigation about BDA tool in order to provide a concrete implementation of representing user dynamic behaviors.

The following scenario illustrates how IDA works, starting from creating user profiles, to studying and classifying human dynamics and generating questionnaire with security potentials to authenticate users. Mike Ross leaves home daily from 8 am from area with postal code “NxY xCy” and returns back at 5:30 pm as recorded by cell towers (through signals from his phone). His route is to a street known to be a banking street in downtown. His purchases and electronic receipts; that contain the time and date of purchases, show that every two to three months, a purchase is made from Hugo Boss store; in Toronto Eaton Center, and two online orders were placed last week. That happens to be confirmed by the day, time and location stored by the telecom company Mike has his sim card from, and the usage of a master card on an online shopping website. Mobile phones enable their users to turn on location and GPS services, to check in places individuals visit and got to, and share it on social media, so can cell phone carries, through cell towers. This information can certainly provide us with useful data about them. Through association learning, discovering relations between variables, for example, having a location record of a banking street on week days from 8 to 5:30 pm, and having receipts from Hugo Boss store, indicates that Mike works at a bank. And through classification algorithm, information analyzed will reveal that Mike buys suits and he is a professional employee. Analyzing the information given through different machine learning techniques; buying from Hugo Boss every period of time, attending a route to work that is known as a banking street, all provide a conclusive result that Mike is a professional employee in a bank. Having information from various sources will provide useful information about users, which will allow for the creation of a dynamic user profile. Once user profile is classified or categorized into certain classification, the system will have enough information to generate a set of questions with which some of security potentials to use in the authentication process. A set of questions that could be generated are as follows:

- Name the store you went to on the beginning of this month to buy a suit.
- Name the route you took this Tuesday instead of the regular route you take on weekdays to work.
- Which package was received using express shipment last week?
- How many bus transfers you made to reach the mall that contains the store that you bought your suits from?

The challenging questions generated are dynamic due to the fact that information are changing on hourly, daily, weekly basis. Answering a question(s) that relate to the actions Mike does ensures the authenticity of the person Mike claiming to be.

In contrast to a lot of two factor authentication systems tend to use static questions for extra security in case of repetitive failure in username/password authentication. The type of questions used (1) are easy to answer and exploit the system if an intruder has a certain level of knowledge in social engineering, (2) and are limited to information related to birthday, favorite pet, favorite restaurant, best friend name etc, as in:

- What is the name of your best friend?
- What is the name of your first pet?
- When is your birthday?
- What is the name of your high school?

What differentiates IDA's challenging questions and having a set of general questions is the dynamic data that allow to analyze fresh information from actions and behaviors made in an hourly, daily, weekly basis, etc. where users will have to answer unique, non-repeated questionnaire that tackle actions that took place in a particular day and time, and that reflect an unusual behavior that occurred and happened for a period of time and required an extra attention from the user. The challenging question used will not be used in any other session when trying log in or gain access to a system, due to the fact that data is changing in an hourly basis, behaviors and actions will be processed on a regular basis to ensure the uniqueness of questions asked, where any intruder trying to gain access to an account won't be able to do so, even when trying to log in again with the right answer, the question will be a totally different one and also won't be used again in any next session (in every session a new challenging question will be asked, and it will be asked once). In case the user is not able to answer the question asked, s/he can change the question just as simple as in forgetting the password mechanism where a user can change the general question used to be answered to the next one s/he remembers.

3.5 Summary

In this chapter, IDA model is proposed as an implementation of the BDA tool, the second component of the next generation user authentication framework. Building and disclosing accurate user profiles can be highly effective in providing a new user authentication tool. Using Big data techniques; data analytics and machine learning, that play an important role in: (1) building and creating user profiles and building up knowledge, (2) linking profiles to different categories. Where association learning algorithm is used to discover relations and patterns in datasets and build up knowledge, and classification algorithm is used in the process of identifying to which set of categories an observation belongs. The third task in the IDA model is to generate on the fly unique challenging questions based on the dynamic data analyzed and the classification made. Thus, real time authentication process; comprised of studying user behavioral profiles and generating challenging questions, is the main goal of the IDA model, that leverages the characteristics of Big data to provide a new alternative of "something you do" authentication factor.

During our continuous investigation, an enhanced solution has been proposed. The enhanced solution will take advantage of recommender systems, and learning techniques behind them to predict user behavior and identify irregular behavior accordingly.

Chapter 4 introduces recommender systems, types and related work. And discusses in more details a hybrid-based filtering approach for user authentication that is the result of the earlier research and investigation of IDA and Big data, and the leaning techniques behind recommender systems.

Chapter 4

4. Hybrid-based Filtering Approach for User Authentication

Traditional methods of authentication rely on a challenge-response mechanism. The investigation of IDA model showed that we are able to take advantage of Big data to recognize patterns and identify behavioral characteristics. The behavioral characteristics can be used as a shared secret between two parties, so that one party asks a question as a challenge and the other party must reply with a correct answer as a response.

The second direction that have been taken during our investigation of the IDA model lead us to propose a hybrid-based filtering approach for user authentication. A model that utilizes the characteristics of CF to create user profiles and identify information of security potentials. Collaborative filtering showed to be very practical in many real-world applications. This will be presented in more details in sections 4.1.

4.1 Related work

Recommendation systems (RS) are systems that offer or predict users' preferences [54]. Recommender systems are used and utilized for different purposes and in different areas. Areas such as movie recommendation, article recommendation, products, books, news, financial services, music, etc. the goal of recommender systems is to recommend a new service(s) to users after considering several attributes. These attributes rely on previous references or users with similar interests.

RS utilize learning algorithms to study users' information and recommend a service. The following sections presents and discusses some of the most known recommender systems.

4.1.1 Web Page Ranking

Web page ranking recommendation system is a system that aims at providing the best websites that could be relevant for the search query provided by a user. the most known webpage ranking recommendation system is Google's search engine. Google search uses "PageRank" algorithm [13] as one of the algorithms and factors to rank websites in their search engine. PageRank works by counting the number of links and checking the quality of links to a page to determine a rough estimate of how important a website is. Where important websites are likely to receive more links from other websites. Basically, it measures the importance of website pages, based on links from other websites. This algorithm is aided by different algorithms (perform link analysis) such as google panda, and google penguin [14].

4.1.2 Amazon Product Recommendation

Amazon’s algorithm takes into account, (1) regular actions and behaviors, (2) current virtual shopping cart, and (3) different users with similar past. Note that, regular actions and behaviors: consists of purchases and ratings of products. Amazon’s approach matches user’s purchased and rated item(s) to similar items, where customers tend to buy or purchase items together, when compared with customers that bought item A and purchased along with it item B [15]. It also takes into account the current shopping cart a user has, what items s/he is interested in, and what are possible items that would interest her/him that are related to what is in the cart [16].

4.1.3 Social Recommendation

Social networks use collaborative filtering, the same algorithm used by amazon, but instead of items, it is used to filter and recommend friends, groups and other sort of connections that a user may be interested in [17]. Networks such as Facebook and LinkedIn examine the connections between a user and their friends, their behavior (such as tags, likes, comments, groups and pages they follow), checking certain profiles more than once, all are monitored by an automated system which then studies the data and provide recommendation accordingly.

4.1.4 News Content Recommendation

Different recommendation approaches are used in this type of systems such content recommendation, collaborative filtering, and hybrid based filtering [18]. Some of the known news content recommendation services are offered by Outbrain and Taboola (the world’s two largest, content discovery and recommendation platforms), and Yahoo [18]. The three main approaches used in news content recommendation are, content-based recommendations, collaborative filtering, and hybrid filtering.

4.1.4.1 Content Recommendation

Content recommender systems recognize articles based tags associated with them [18]. A system would recognize that a user read about sports for example, and then recommends other sports articles. It is similar to a librarian that knows what a user is reading and where his/her interests are and points him/her toward sections with similar topics. Some advanced content based filtering use natural language processing to better understand what an article is about. For example, the New York times uses a natural language processing technique called latent dirichlet allocation that allows it to know what an article is about by counting the number of times a certain word appears, and compares it to other articles. It then categorizes an article into different topics depending on the words used. For example, 50% science and technology, 20% environment.

4.1.4.2 Collaborative Filtering (CF)

Systems that use CF recommend content based on a user past reading habits and compares his/her history with people with similar history, or with similar interests. As in the given example above about the librarian, a librarian who knows the user’s interests are similar to another student or customer, can recommend that user to read a book that the other customer liked. Typically, it uses the user history to recommend articles to another one with same preference of reading history**.**

4.1.4.3 Hybrid Filtering

Hybrid filtering combines both content-based and collaborative filtering to give a recommendation based on both content of an article, and the person who is reading. For example, Outbrain combines contextual algorithms, behavioral algorithms, and personal algorithms to serve a good recommendation. Contextual algorithms analyze the context the user is reading now and searches for similar content and topics. Behavioral algorithms that learn behaviors of groups of users. For example, saving records of the most visited documents in a site, the most rated documents, the ones most shared, etc. adding to that the content people read and liked in the earlier times. Personal algorithms learn the look up the history of a user, or group of people, and gives recommendations that is related to their interests.

Chapter 4 presents the second solution where the proposed model investigates the utilization of some learning algorithms used in RS to create and classify user profiles, analyze their information, build up knowledge to compare current actions with predicted actions, and generate challenging questions, to authenticate users in real time and on demand. In other words, if RS have the assumption that people who buy X also buy Y, if a user u was among those people, yet u did not buy Y, RS will recommend u to buy it. However, in our model if we have the same assumption, we will build a knowledge that it is normal for a user u to buy Y, yet if u bought Z instead, the model will interpret it as a unique action and the system will generate a dynamic question to challenge the user to verify his/her identity.

4.2 Hybrid-based Filtering Approach for User Authentication

The proposed solution is the result of the investigation of the second component of the Big data-based authentication framework described in chapter 2 section 2.2, and the IDA model discussed in chapter 3 section 3.1. The proposed model aims to recognize users' information and create user profiles. This is followed by applying learning techniques such as collaborative filtering to identify irregular behavior to generate challenging questions to authenticate users.

For this purpose, our approach has two main activities – Figure 4.1 - that are involved to achieve this goal.

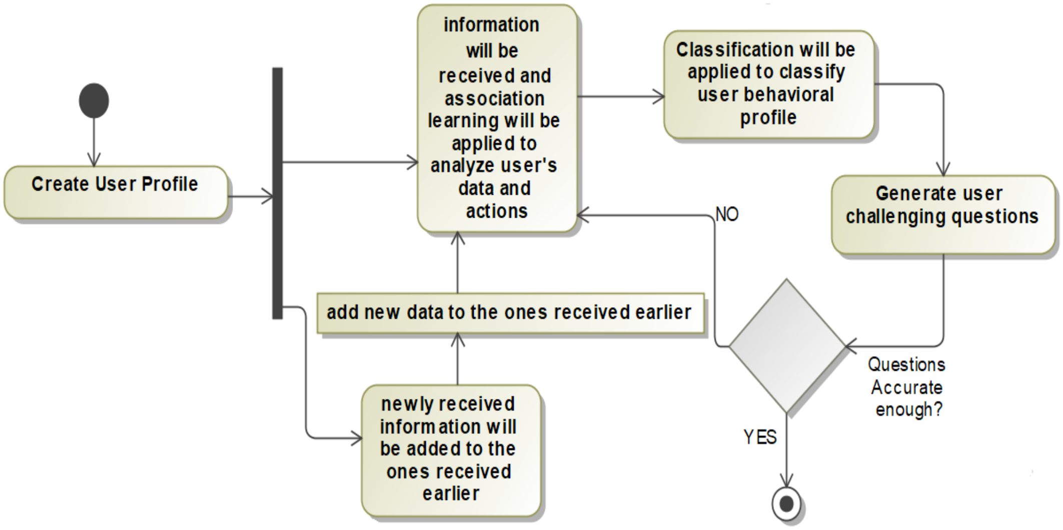

Create User Profile
information will be received and association learning will be applied to analyze user's data and actions
Classification will be applied to classify user behavioral profile
Generate user challenging questions
add new data to the ones received earlier
newly received information will be added to the ones received earlier
NO
Questions Accurate enough?
YES

Figure 4.1 Activity workflow

These activities are modified version of the one proposed in chapter 3. The first activity is receiving information from various sources to create users' profiles to study human dynamics and behavior, and analyze their actions and behaviors, then classify behavior accordingly. The second activity is to identify information of security potentials and generate unique challenging questions on demand to authenticate users.

The two activities described in section 4.1 are decomposed into three main components, where the first activity of the model is decomposed into the learning component, and the second activity is decomposed into the listening, and challenging components. The model triggers on the event of user's data received and the corresponding user's profiles created/modified - Figure 4.2. The three components integrate to achieve one common purpose that is to predict the user's u behavior based on the similarities, check his/her neighboring users v behaviors, based on time performed and whether they were occurred previously or not. Hence the model would be able to identify whether these actions lead to regular or irregular behavior so that the challenging questions can be accurately created on demand. The detail explanation of the model components is presented in the following sections.

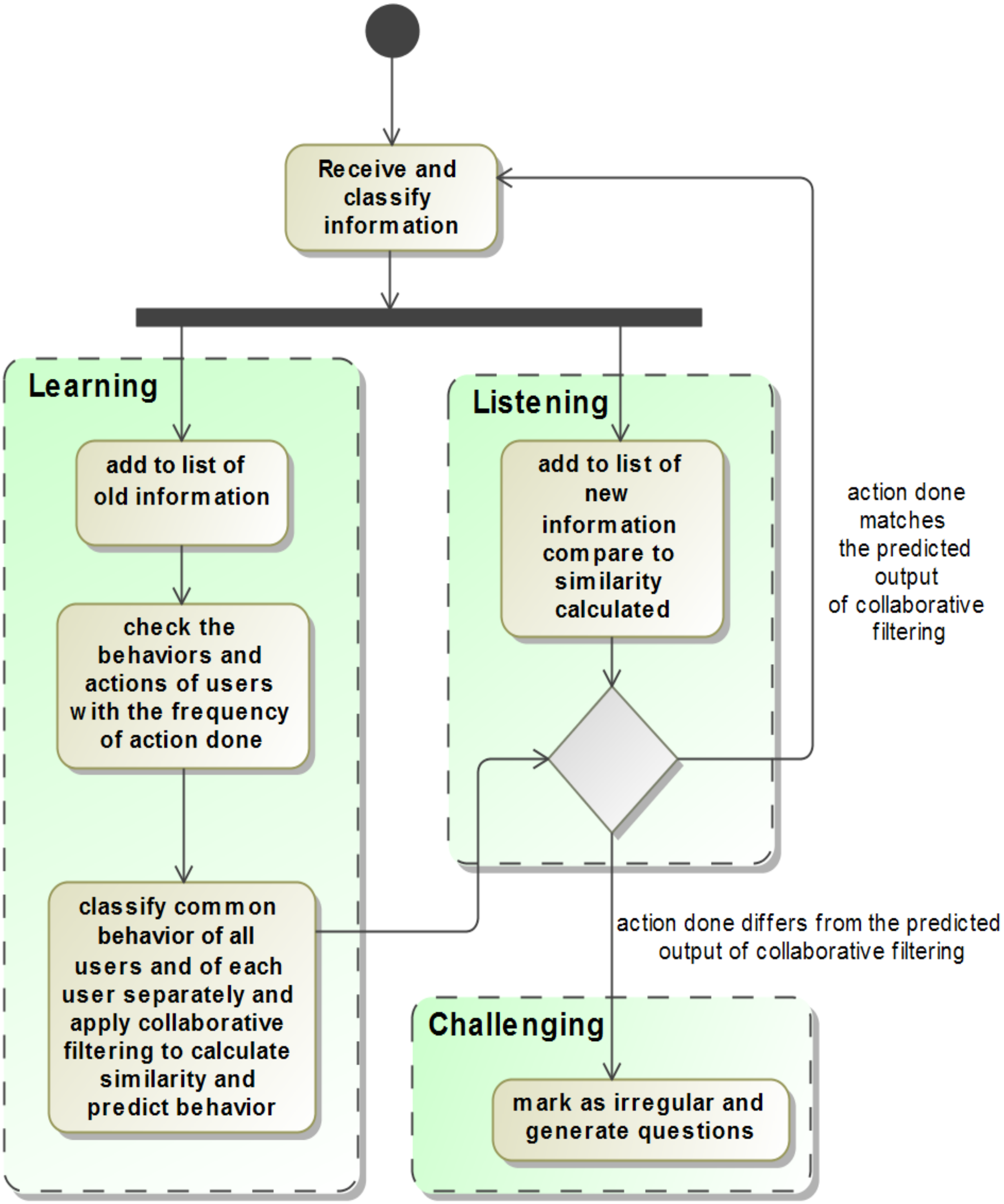

Receive and classify information
Learning
add to list of old information
check the behaviors and actions of users with the frequency of action done
classify common behavior of all users and of each user separately and apply collaborative filtering to calculate similarity and predict behavior
Listening
add to list of new information compare to similarity calculated
action done matches the predicted output of collaborative filtering
action done differs from the predicted output of collaborative filtering
Challenging
mark as irregular and generate questions

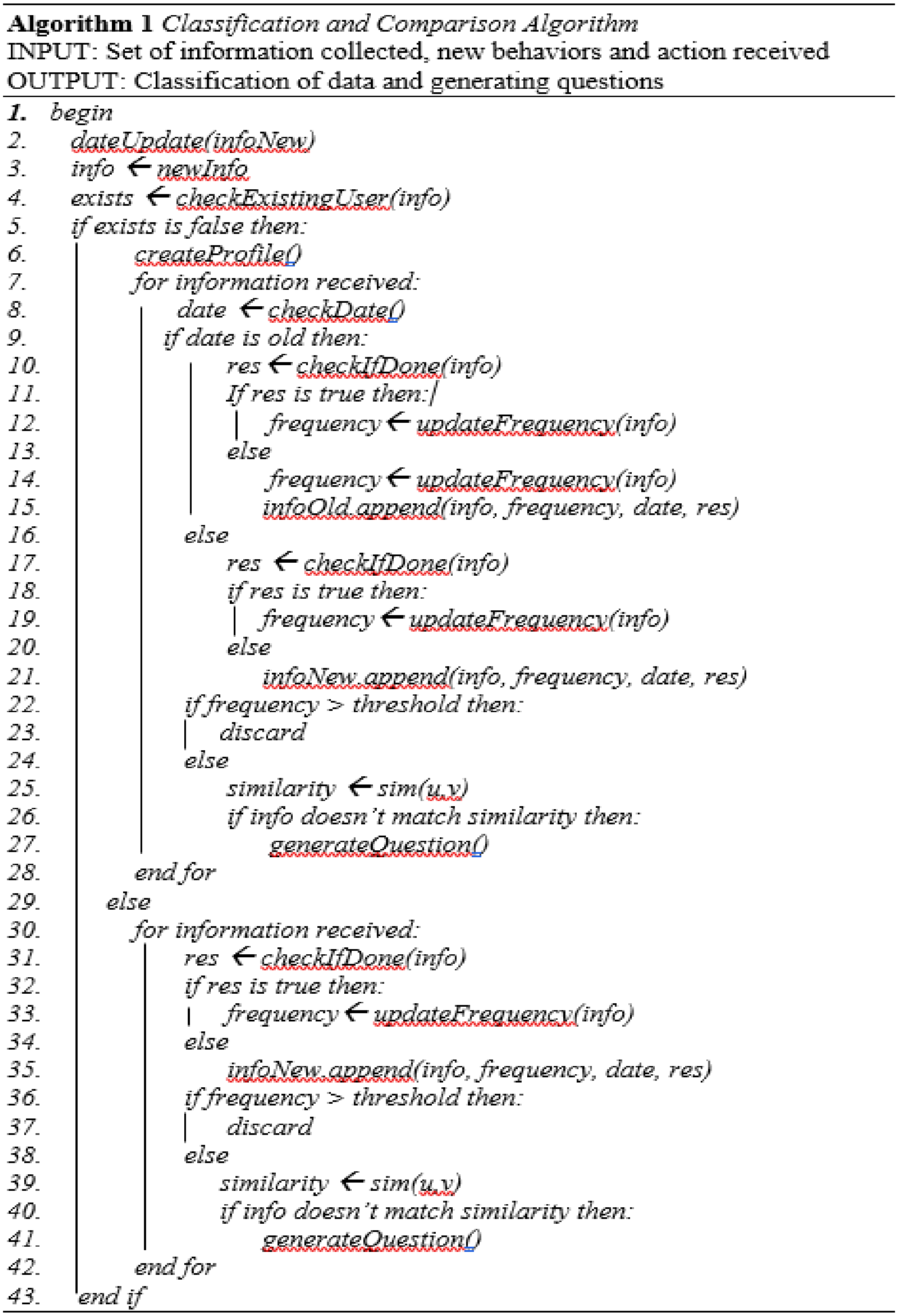
Figure 4.2 Authentication Model Work-flow

4.2.1 The learning Component

The learning component starts on the receiving of the user's data for the first time, to classify information into two categories, old and new based on the time stamps of the actions. Every action has a time stamp associated with it, so a time range will be set in order to classify information into new if the time stamp lies in the time range T set, otherwise will be classified as old. "T" will be automatically updated on a regular basis, based on the real date and time, in order to accurately classify data into the correct category. Based on algorithm 1, the proposed model will start by classifying data based on timestamps, the information will then be checked if done before and keep track of how many times it was done.

This process is similar to Netflix and Gmail's systems, that takes into consideration different attributes to learn about every user. where Netflix's system works on providing the best recommendation regarding TV shows and/or movies for each and every user to watch. The main problem with Netflix to solve is to find shows/movies that are compelling to view. To do so, Netflix's system and algorithms take into consideration various attributes. Some of these attributes are, personalized video ranker (PVR), top N video ranker, trending now, continue watching, page generation (row selection and ranking). As well as, evidence selection that define Netflix experience, and search, where searching is the process when a user looks for something (20% of recommendation is based and influenced by what a user looks for). Combining all these features together the algorithms used make up the complete Netflix recommender system. Similarly, the proposed model will take advantage of various attributes such as, actions' time stamps, actions related to navigation and commute (including location at different times during a day, week, month), purchase activities (including way of paying), online search history, online streaming activity and preference, electronic devices usage, news feed follow ups. So will keep track of the number of times a certain action(s) has been made, and classify based on the frequency of actions. The more frequent the more common a behavior will be.

The third step in the learning component is to apply collaborative filtering technique to calculate similarity between a user u and neighboring users v, and predict how a user will behave compared to users with similar behavior. The objective of using collaborative filtering is to predict how likely a user will perform a certain action in the future based on the similarity with neighboring users.



Algorithm 1 *Classification and Comparison Algorithm*
INPUT: Set of information collected, new behaviors and action received
OUTPUT: Classification of data and generating questions

```
begin
   dateUpdate(infoNew)
   info ← newInfo
   exists ← checkExistingUser(info)
   if exists is false then:
         createProfile()
         for information received:
            date ← checkDate()
            if date is old then:
               res ← checkIfDone(info)
               If res is true then:
                  frequency ← updateFrequency(info)
               else
                  frequency ← updateFrequency(info)
                  infoOld.append(info, frequency, date, res)
            else
               res ← checkIfDone(info)
               if res is true then:
                  frequency ← updateFrequency(info)
               else
                  infoNew.append(info, frequency, date, res)
            if frequency > threshold then:
               discard
            else
               similarity ← sim(u,v)
               if info doesn't match similarity then:
                  generateQuestion()
         end for
      else
         for information received:
            res ← checkIfDone(info)
            if res is true then:
               frequency ← updateFrequency(info)
            else
               infoNew.append(info, frequency, date, res)
            if frequency > threshold then:
               discard
            else
               similarity ← sim(u,v)
               if info doesn't match similarity then:
                  generateQuestion()
         end for
   end if
```

Figure 4.3 Classification and Comparison Algorithm

4.2.2 The Listening and Challenging Components

The Listening component comes to play by checking irregular behavior either by analyzing and comparing a user's own behavior (check scenario II), or by calculating similarity and predicting an action (scenario I and III). When the user's action is marked as irregular by the listening component, the challenging components start to generate unique challenging question to authenticate the user u.

In order to predict future behavior, the model compares actions of a user u, actions and behaviors for all users' v combined. Similar to Gmail's inbox, the model compares newly received information with predicted information to determine whether the new actions analyzed are irregular or not, and will keep learning and update the set of behaviors classified by calculating similarities between users and actions. To calculate the similarity between a user "u" and group of users "v" (where v will be assumed a user that consists of group of users) we will use Pearson correlation [43][44] equations (1) and (2); please check appendix for the derivations. Pearson correlation is the covariance of two variables divided by the product of their standard deviations [44].

$$sim\ (X,Y) = \frac{\Sigma_{i=1}^{m}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma_{i=1}^{m}(X_i - \bar{X})^2}\sqrt{\Sigma_{i=1}^{m}(Y_i - \bar{Y})^2}} \qquad (1)$$

Simplifying equation (1) to equation (2)

$$sim(x,y) = \frac{\mathrm{n}\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{\sqrt{n\Sigma X_i^2 - (\Sigma X_i)^2}\sqrt{n\Sigma Y_i^2 - (\Sigma Y_i)^2}} \qquad (2)$$

Where N is number of users or items being considered based on the matrix built, x is the score of 1st attribute, y is the score of 2nd attribute, where actions will be represented by numeric values, to calculate the similarity which is defined by the Pearson correlation (1) - (2). The output value of the Pearson correlation lies between -1 and 1 (the closest the value to 1 the more similar two users are).

4.3 Summary

This chapter proposed a comprehensive framework and describes a hybrid model for user authentication and the components of the new model architecture. It also introduces and overview of the function of each component the rely on Big data’s characteristics and the adaptable mechanisms that take advantage of the learning techniques. In addition, it shows how to take advantage of recommender and prediction systems and how they will to generate unique dynamic challenging questions.

Chapter 5

5. Evaluation

This chapter discusses the methodology that has been investigated with regard to recommender systems and learning techniques to recognize behavior characteristics in users' data, and presents different use case scenarios and a case study to illustrate the experiment that has been conducted. The experiment done used MovieLens public dataset. And the implementation done was customized to handle the data that we have.

5.1 Methodology

The authentication model that has been proposed is an outcome of the investigation about recommendation systems in order to provide a concrete implementation of user dynamic behaviors. The following scenarios illustrate how the proposed model should work, from receiving information, and analyzing them, to generating challenging questions. Due to the fact that we weren't able to get data from different sources to evaluate the accuracy of the proposed model, we have provided different use case scenarios to illustrate how the application would work. Moreover, in section 5.3 we demonstrated a use case study based on movie ratings data set “MovieLens”, where a system prototype was implemented to demonstrate a real case scenario of how the system might behave.

5.2 Use Case Scenarios

Assume the following example where Jonathan Ross is an employee in one of the banks, and works in the downtown office. After receiving Jonathan’s data, the following process will take place. First a profile for Jonathan will be created. The model will start by looking for actions made by Jonathan, such as, commuting, making phone calls, purchasing activity, online searching and streaming behavior, news feed follow-ups and social interaction, electronic devices used. Using analytic techniques, the model can recognize patterns with related actions.

The following sample use case scenarios illustrate how the proposed model works, from receiving user’s information, classifying them to generating challenging questions.

5.2.1 Scenario I

For the following scenario considers that Jonathan’s profile (user 3 in table 5.1) contains a list of phone calls that will be compared to two family members since they are all registered under the same account. The below example is given to relate age and type of calls made. In the given example, the conclusion was that the older the person is, the number of their international calls is lower, and the younger the individual is, the number of their international calls is higher.

The given example below explains in more details, how this correlation is made based on the data given. The results of the similarities are based on the pearson correlation presented in chapter 4 section 4.2.2.

Table 5.1 User Phone Records sample

P	**Age-x**	**Ncall-y**	**Incall-z**	**xy**	**xz**	**x^2**	**y^2**	**z^2**
1	40	15	40	600	1600	1600	225	1600
2	18	1	90	18	1620	324	1	8100
3	24	2	110	48	2640	576	4	12100
total	82	18	240	666	5860	2500	230	21800

Using the proposed model, the system will determine whether a profile exits or not, if user already exists in the system, it will check if the action is previously done or not, update list of actions and the frequency if it has already been done, check time of action and classify it under the right category of actions. Assume the system generated table 5.1, where p represents persons (1, 2 and 3), their ages, NCall is national calls, and Incall is international calls. Calculating similarity between attribute x and y, and x and z, sim(x, y) = 0.979, and sim(x, z) = -0.8535. This shows that x and y are correlated while x and z are not. Whenever x increases y should increase, on the contrary, whenever x increases z should decrease. A prediction should reflect an increased number of national calls done by Jonathan, if observed otherwise in his behavior, this means the new phone calls made, doesn’t match the similarity calculated and prediction made, a question will be generated asking about the incident recorded to authenticate Jonathan whenever needed. Note that, only Jonathan knows the answer for the question generated that might be, for example, as follows:

Name the country you made an international call to on 13th of Jan 2017.

Also, note that this question will be used once during any authentication process, due to the fact that the model is receiving vast amount of information to process and analyze, information will be always changing.

5.2.2 Scenario II

Consider that for the same user, Jonathan, has a purchase activity as shown in table 5.2. (N represents no, and Y represents yes). After checking and analyzing the data the model verifies that most of Jonathan's purchase activity is as follows, for example he uses visa for non-online shopping and MasterCard for online shipping.

Table 5.2 Purchase activity records sample

Card	store	Online purchase	Date
visa	walmart	N	Dec 1 2016
Visa	Walmart	N	Dec 1 2016
Visa	Columbia	N	Dec 2 2016
Mastercard	Netflix	Y	Dec 2 2016
Visa	Starbucks	N	Dec 3 2016
Mastercard	Amazon	Y	Dec 3 2016
Visa	Walmart	N	Dec 5 2016
Visa	starbucks	N	Dec 8 2016
Visa	Sobeys	N	Dec 9 2016
Mastercard	amazon	Y	Dec 11 2016
Visa	starbucks	N	Dec 17 2016
Mastercard	Bestbuy	Y	Dec 19 2016
Mastercard	Amazon	Y	Dec 22 2016
Mastercard	Walmart	N	Dec 23 2016
Visa	Walmart	N	Dec 24 2016
Visa	Sobeys	N	Dec 24 2016
Visa	Sobeys	N	Dec 24 2016

By analyzing and comparing Jonathan's behavior alone, the model identifies that on Dec 23 2016 Jonathan used his MasterCard once for a non-online shopping. This type of information is considered of security potential, and a question will be generated regarding this matter.

What type of card did you use for paying on Dec 23rd in Walmart? or

What is the date and place for using MasterCard for non-online shopping?

Only Jonathan knows the answer for these questions, and they will only be used once for authentication. The model will keep track of such behavior in case it is repeated in the future, so as the frequency increases, such behavior won’t be considered of security potential anymore.

5.2.3 Scenario III

In this use case scenario assume that Jonathan’s streaming habits and locations are as follows. Jonathan usually watches movies with genre action, adventure, comedy, his data reveal his TV shows and movies ratings. Collaborative filtering in this case compares Jonathan’s preference to the common preference of all users, and outputs certain movies and TV shows that Jonathan that might watch in the future. If it happens that any certain point Jonathan watched a horror show, which contradicts what was predicted by the proposed model based on the similarity with neighboring users, it marks such behavior as irregular. A question asked might be as follows:

What movie did you watch on Jan 10 2017?

Note that, the questions generated will be used only once, so if an intruder; happens to pretend to be Jonathan, will not be able to get access to a Jonathan’s account, even if he/she knows the answer to a question asked earlier, the intruder will have to answer a new generated question.

5.2.4 Scenario IV

In this use case scenario, assume that Alice uses an online movie service where she watches movies with adventure and animation genres as presented in table 5.3. The system will then build up knowledge about what Alice will most likely watch. The knowledge built will be labeled as regular behavior as presented in table 5.4.

The system has recorded that Alice was watching the following movies during the week of April 3, 2017:

Table 5.3 Alice's records

Movie ID	Title	Genres
64695	Sword of the Stranger (2007)	Adventure, Animation
95473	Dragon Ball Z: The Return of Cooler (1992)	Adventure, Animation
95475	Dragon Ball Z: Cooler's Revenge (1991)	Adventure, Animation
99766	Superman: The Return of Black Adam (2010)	Adventure, Animation
102154	Superman Unbound (2013)	Adventure, Animation

The system will build up knowledge that Alice most likely will watch movies like:

Table 5.4 Built up knowledge on what Alice most likely will watch

Movie ID	Title	Genres
95771	Dragon Ball Z: Broly Second Coming (1994)	Adventure, Animation
95780	Dragon Ball Z: Bio-Broly (1994)	Adventure, Animation
95782	Dragon Ball Z: Fusion Reborn (1995)	Adventure, Animation
95963	Dragon Ball Z: Wrath of the Dragon (1995)	Adventure, Animation
95965	Dragon Ball Z: The Father of Goku (1990)	Adventure, Animation
103233	LEGO Batman: DC Heroes Unite (2013)	Adventure, Animation

109776	Dick Figures: The Movie (2013)	Adventu
112175	How to Train Your Dragon 2 (2014)	Adven

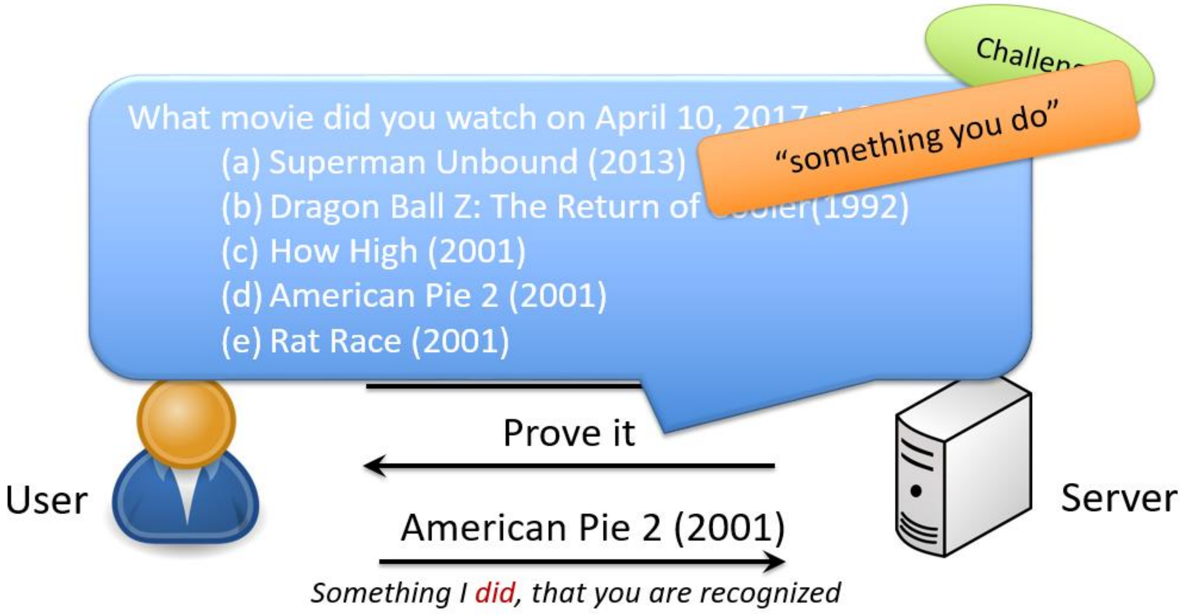
Regular behavior

If Alice watches a movie different from what has been predicted to be regular, will mark that as irregular and the data analyzed will be used to generate one time challenging question to authenticate Alice, as presented in table 5.5 and figure 5.1

Table 5.5 Irregular behavior recorded

Movie ID	Title	Genres
4718	American Pie 2 (2001)	Comedy

Irregular behavior



What movie did you watch on April 10, 2017

(a) Superman Unbound (2013)

(b) Dragon Ball Z: The Return of Cooler(1992)

(c) How High (2001)

(d) American Pie 2 (2001)

(e) Rat Race (2001)

Figure 5.1 Sample authentication process

Collaborative filtering approach was applied on a public data set to simulate the behavior of the proposed model as a proof of concept; create user profiles, analyzing information based on users' actions, predict behavior and check for irregular behavior. this will be presented in the next section, where the experiment has been done.

5.3 Case Study

The evaluation was performed by means of a case study. The case study uses a public data set: MovieLens. This dataset consists of 20 million ratings, 465,000 text tags applied to 27000 movies, and 138000 users. the system uses the attributes defined in the data source, such as user id, movie id, ratings, time stamps, genres, description about movies.

To conduct the evaluation, the functional paradigm of Scala Language and GraphLab libraries were used in the implementation process. And the language that has been used for the implementation is python. Appendix B shows the libraries, methods hardware specifications that have been used.

Since the data we have is archived and old, we considered a time in 2004 as present time. The time is recorded by day, month, year, hours, minutes, and seconds. So, the data being imported and used by the system is considered as a present streaming data based on timestamp. For example, assume the present time is the range between 28-04-2004 13:10:57 and 29-04-2004 00:00:00. Every action of watching a movie before that range of date and time is considered old, and during is present. So, we assume the data is imported in real time because we used it as a stream based on the time stamp.

Table 5.6 shows the number of neighbors used for comparison for a user and number of iterations done.

Table 5.6 Parameters used to train the CF models

Parameter	**Value**
Number of neighbors	64
Number of Iterations	10

5.3.1 Case Study: Loading and Preparing Data

For the CF prediction engine, the case study uses two datasets; movie item data that contains 27,279 movies, and user action data that consist of 20 million ratings. The item data dataset describes movies with their titles, description, genres, etc.

Table 5.7 Item Dataset Sample

MovieID	**Title**	**Genres**
1	Toy Story	Adventure, Animation, Children, Comedy, Fantasy
2	Braveheart	Drama, Action, War
3	Star Wars	Fantasy, Sci-Fi, War, Action

While the action data dataset contains the users’ behavior with movie items, and includes, the user id, movie id, rating and timestamps.

Table 5.8 Action Dataset Sample

UserID	MovieID	Rating	TimeStamp
1	2	3.5	1112486027
1	5	5	1112484686
1	30	4	1112484817

Once data is loaded, we transform it in various ways. For example, we extract year and title from original title column, and combine them together, and then parse genres column by splitting on commas.

5.3.2 Case Study: Model Definition

The model starts by splitting the data into training set and validation set, and then cleans-up the data. Splitting the data into training set validation set and test set; 50% - 25% - 25% respectively, is to ensure that we have enough observations in the training data. And cleaning the data is to reduce errors and improve the data quality.

The model then applies collaborative filtering to create user profiles and aggregate information then computes the similarity between items using the observations of users who have interacted with similar items. Given a similarity between item i and j, sim(i, j), it scores an item j for user u using a weighted average of the user’s previous observations Iu. Where Iu is the interactions of user u with items n. The result of running CF and classification and comparison algorithm (algorithm 1) is the irregular behavior that will be used in generating the challenging questions. This is presented in figure 5.2 and outlined by algorithm 2.

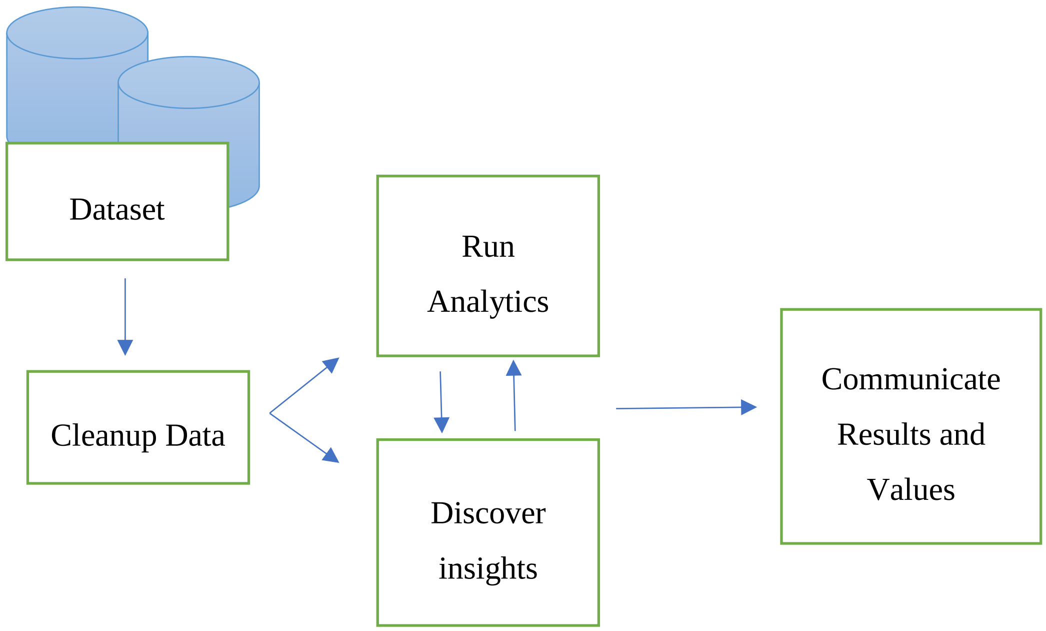


Figure 5.2 Model Chain Construction

Algorithm 2: Loading and Preparing Data
INPUT: Set of information collected from sources
OUTPUT: Dataset of aggregated information from different sources

```
begin
    // read data from directory
    reading data ← sframe.read_csv(path.join(data_dir, ‘movies.csv’))
    // aggregate data from different sources
    sframe.read_csv(path.join(data_dir, ‘ratings.csv’))
    // extract important information

    //extract (year, title and genre)
    extract year
    extract title
    extract genre
    extract date and time
    get the metadata ready:
    urls = Sframe.read_csv(path.join(data_dir, ‘movie_urls.csv’))
    items = items.join(urls, on=’movieId’)
    users = Sframe.read_csv(path.join(data_dir, ‘user_names.csv’))
end
```

5.3.2.1 Training Model

In the training model, the data is split to training set and validation or testing set, where collaborative filtering will be applied to calculate the similarity between a user u and neighboring users to build up knowledge on what would be considered as regular behavior. This is shown in figure 5.3 and outlined by algorithm 3.

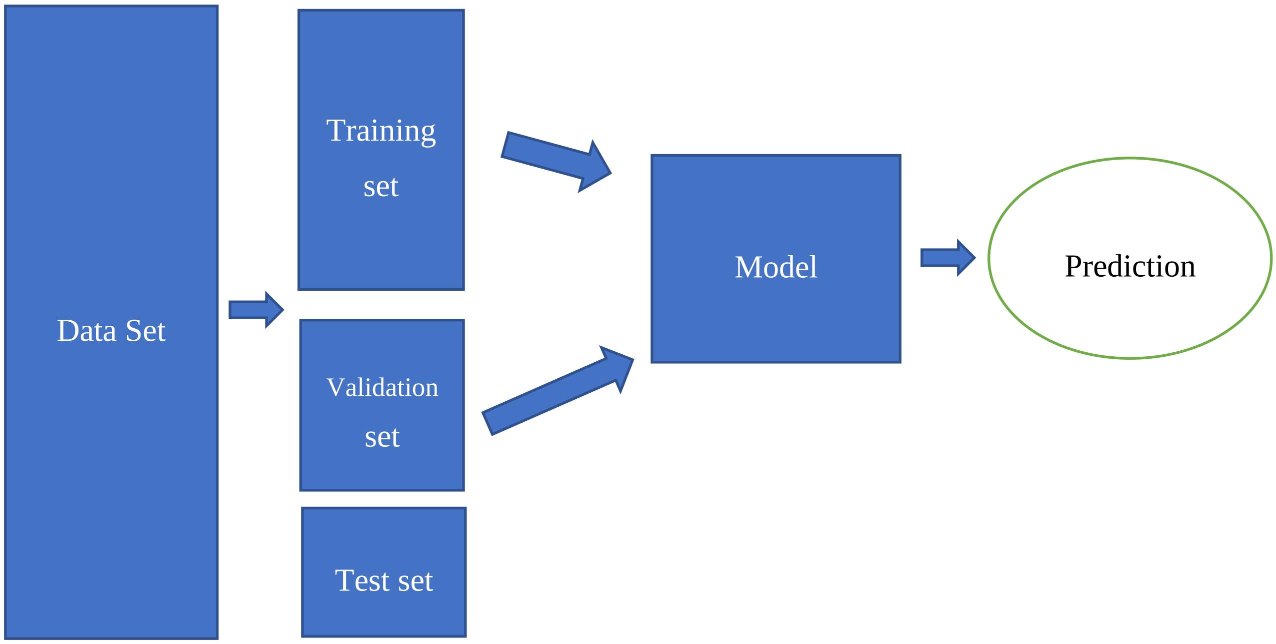


Figure 5.3 Prediction Model Work flow

Algorithm 3: Training and validating Data
INPUT: Dataset of aggregated information
OUTPUT: training data using item similarity

```
begin
    // splitting data into training set and validation set, ratio is 50 for training, 25 for validation and 25 for testing
    train_data = graphlab.Sframe(rating_base)
    test_data = graphlab.Sframe(ratings_test)
    training_data, validation_data ← gl.recommender.util.split_by_item(actions, 'UId', 'MId')
    // data is split using the graphlab built in function.
    // in our case data is split by item
    model = itemSimilarity.recommender.create(training_data, 'UId', 'MId')
end
```

To evaluate the experiment and check the difference between popularity similarity that is based on classification and user-user filtering and item-item based filtering, precision and recall was used.

5.3.2.2 Precision and Recall

In information retrieval, the precision is the fraction of retrieved instances that are relevant, in other words it is how many selected items are relevant. While recall is defined as the fraction of relevant instances that are retrieved, in other words it is how many relevant items are selected.

A perfect precision score of 1 means that every result retrieved by a search was relevant but says nothing about whether all relevant documents were retrieved, on the contrary a recall score of 1 means that all relevant documents were retrieved by the search/prediction/recommendation, but says nothing about how many irrelevant documents was also retrieved.

The relationship of precision and recall is inversely related. So, for our testing environment we have tested how item-item similarity and popularity similarity and compared them to show which model is better in building our prediction system.

There are three choices of similarity metrics to use: ‘jaccard’, ‘cosine’ and ‘pearson’. Jaccard similarity is used to measure the similarity between two set of elements. In the context of recommendation, the Jaccard similarity (JS) between two items is computed. Jaccard is a good choice when one only has implicit feedbacks of items (e.g., people rated them or not), or when one does not care about how many stars items received.

If one needs to compare the ratings of items, Cosine and Pearson similarity are recommended.

A problem with Cosine similarity is that it does not consider the differences in the mean and variance of the ratings made to items i and j.

Another popular measure that compares ratings where the effects of means and variance have been removed is Pearson Correlation similarity presented by equation (1) and (2) in section 4.2.2.

This is outlined by algorithm 4; where algorithm 1 that is represented by authenticate() method is called to build profiles, compare actions and build up knowledge to generate questions.

Algorithm 4: Training Recommender and prediction model and irregular behavior identifier
INPUT: Data aggregated in the training section
OUTPUT: Challenging Questions

```
begin
    // train data aggregated
    model ← recommender.create(training_data, 'UId', 'MId')
    checking ratings, timestamps genres of watched movies
    // algorithm 1 runs to check irregular behavior
    authenticate () ← compares prediction with behavior
    // the output of algorithm 1 is compared with what has been predicted and generates questions.
end
```

The performance of user similarity compared to item similarity is outlined by algorithm 5.

Algorithm 5: Model Performance
INPUT: Item Similarity and User Similarity
OUTPUT: performance comparison

```
begin
    // takes in user_similarity model and compares it to item_similarity
    item_similarity = itemSimModel.recommend(items=range(1,10), k= 64)
    user_similarity = userSimModel.recommend(users=range(1,10), k= 64)
    performance ← graphlab.compare(validation_data, (user_similarity, item_similarity))
    // displays graphical representation of the performance of the two models.
    graphlab.show_comparison(performance, (user_similarity, item_similarity))
```

```
6. end
```

Table 5.9 is for the user similarity model tested, and table 5.10 is for item similarity model.

Table 5.9 Precision and Recall for user similarity model

cutoff	Mean precision	Mean recall
1	0	0
2	0.00053	0.00010
3	0.00035	0.00010
4	0.00026	0.00010
5	0.00021	0.00010
6	0.00017	0.00010
7	0.00015	0.00010
8	0.00013	0.00010
9	0.00011	0.00010
10	0.00021	0.00021

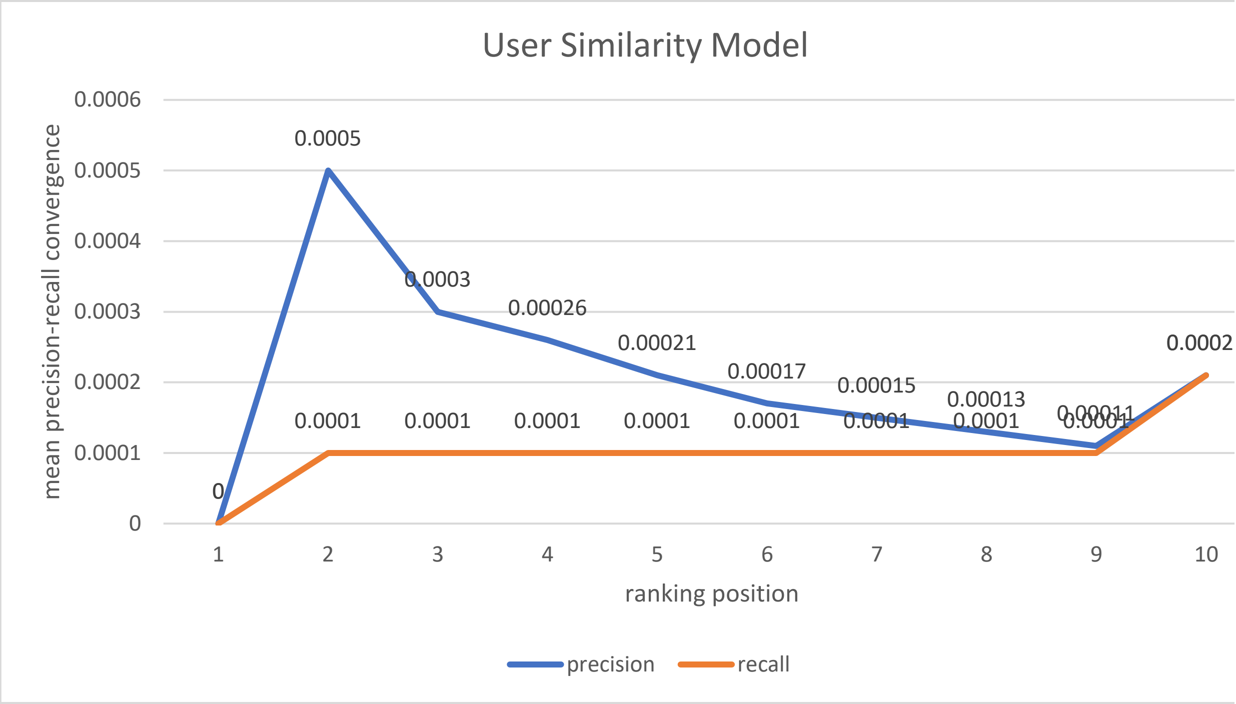


Figure 5.4 Precision and Recall for popularity model

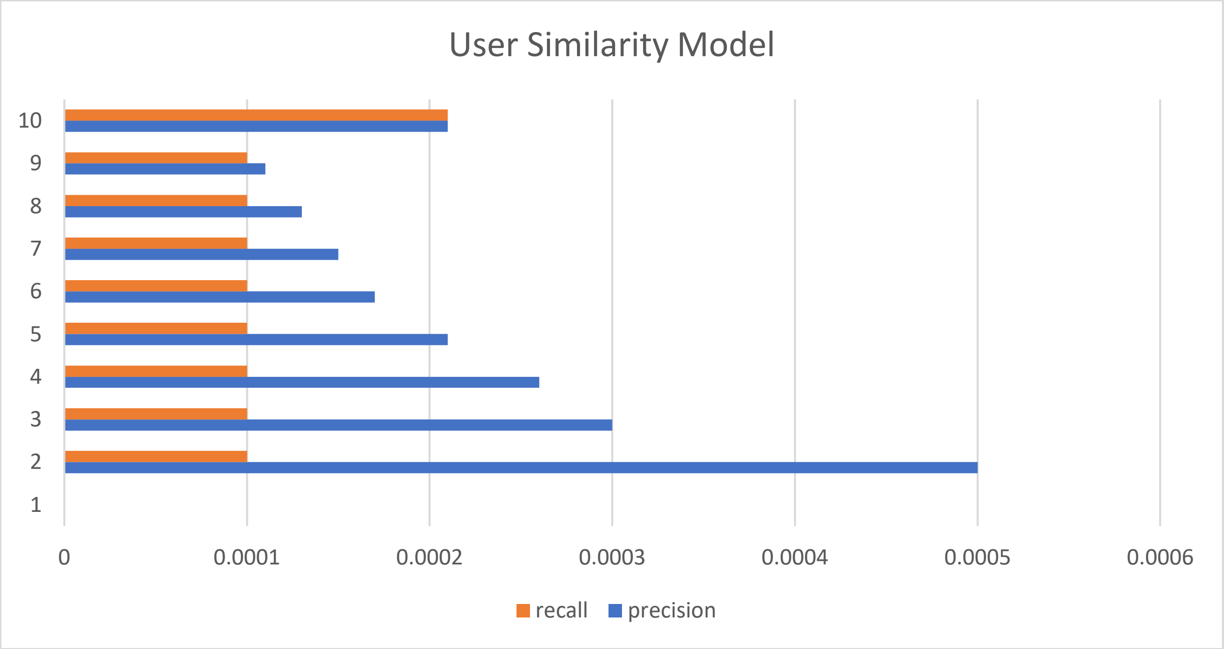


Figure 5.5 Precision and Recall for popularity model

As we can see, the performance of the system is not that great, as the score is about 0.002; figure 5.4, compared to the results of the item similarity model which is almost 0.06 – table 5.10 and figure 5.6.

Table 5.10 Precision and recall of Item Similarity Model

cutoff	Mean precision	Mean recall
1	0.00848	0.00084
2	0.00742	0.00148
3	0.00777	0.00233
4	0.00715	0.00286
5	0.00615	0.00307
6	0.00618	0.00371
7	0.00605	0.00424
8	0.00596	0.00477
9	0.00612	0.005514
10	0.00583	0.005832

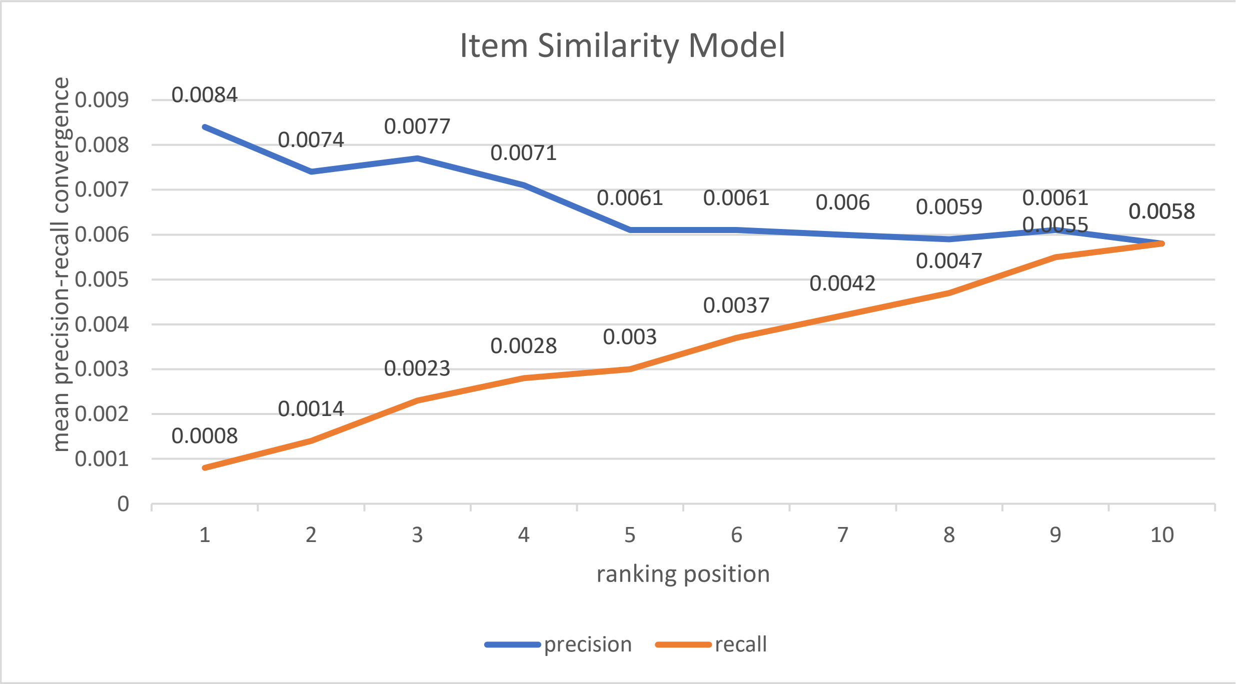


Figure 5.6 Precision and recall of Item Similarity Model

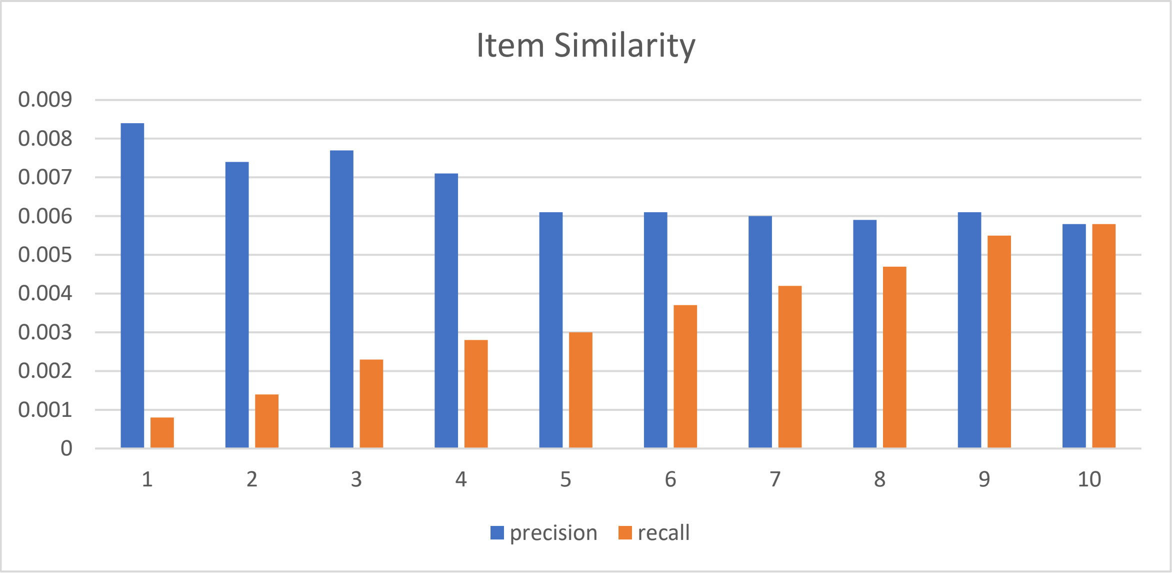


Figure 5.7 Precision and recall of Item Similarity Model

Comparing each of recall and precision of both models independently, we notice a big difference between the two models.

Table 5.11 Recall comparison between the two models

cutoff	recall user similarity	recall item similarity
1	0	0.00084
2	0.000106	0.00148
3	0.000106	0.00233
4	0.000106	0.00286
5	0.000106	0.00307
6	0.000106	0.00371
7	0.000106	0.00424
8	0.000106	0.00477
9	0.000106	0.00551
10	0.000212	0.00583

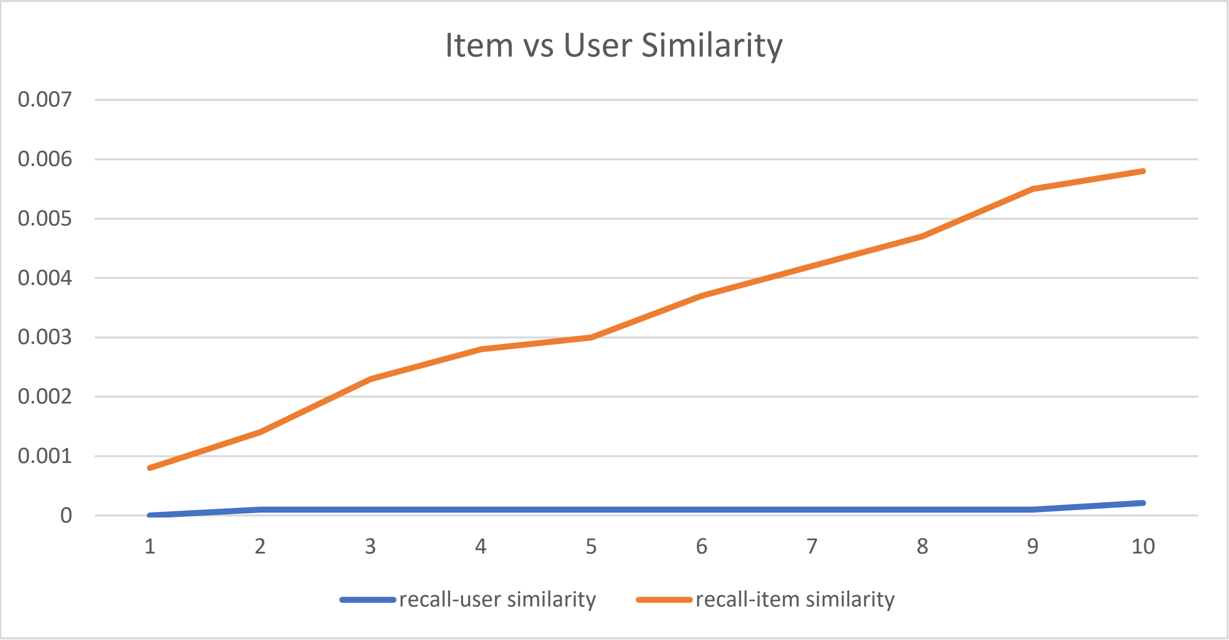


Figure 5.8 Recall comparison between Item Similarity and popularity similarity

Table 5.11 and Figure 5.8, table 5.12 and Figure 5.9 show that the item similarity result in more relevant predictions compared to the user similarity model.

Table 5.12 Precision comparison between the two models

users' similarity	precision user similarity	precision item similarity
1	0	0.00848
2	0.00053	0.00742
3	0.00035	0.00777
4	0.00026	0.00715
5	0.00021	0.00615
6	0.00017	0.00618
7	0.00015	0.00605
8	0.00013	0.00596
9	0.00011	0.00612
10	0.00021	0.00583

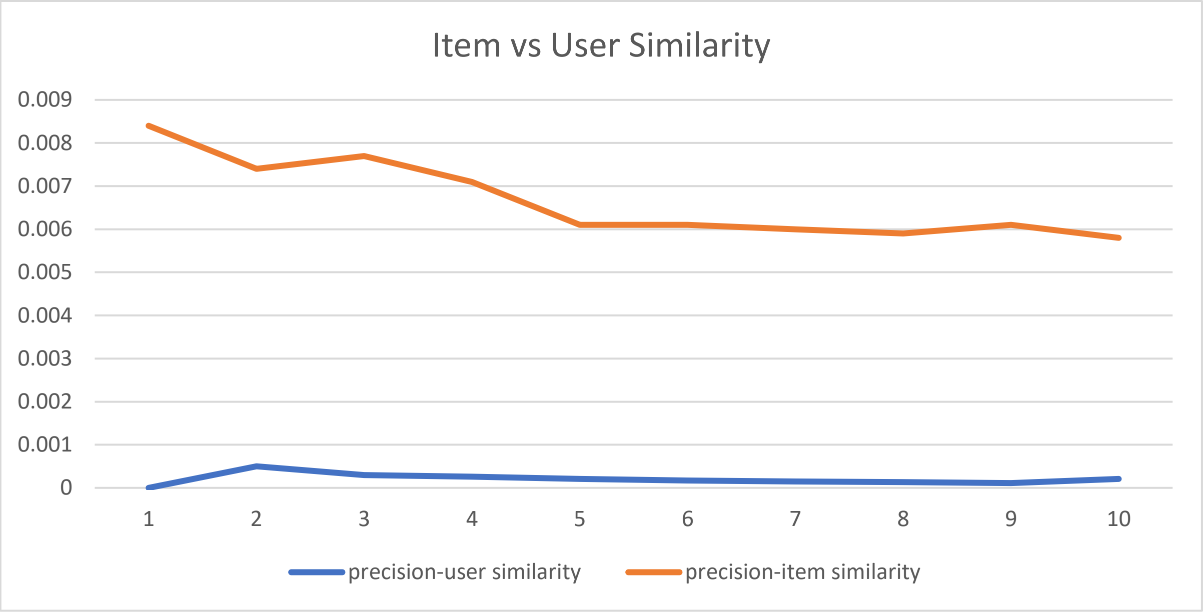


Figure 5.9 Precision comparison between the two models

The following section demonstrates the experiment environment that have been tested.

5.3.2.3 Test-bed

The following experiment has been conducted in a local server with the following characteristics:

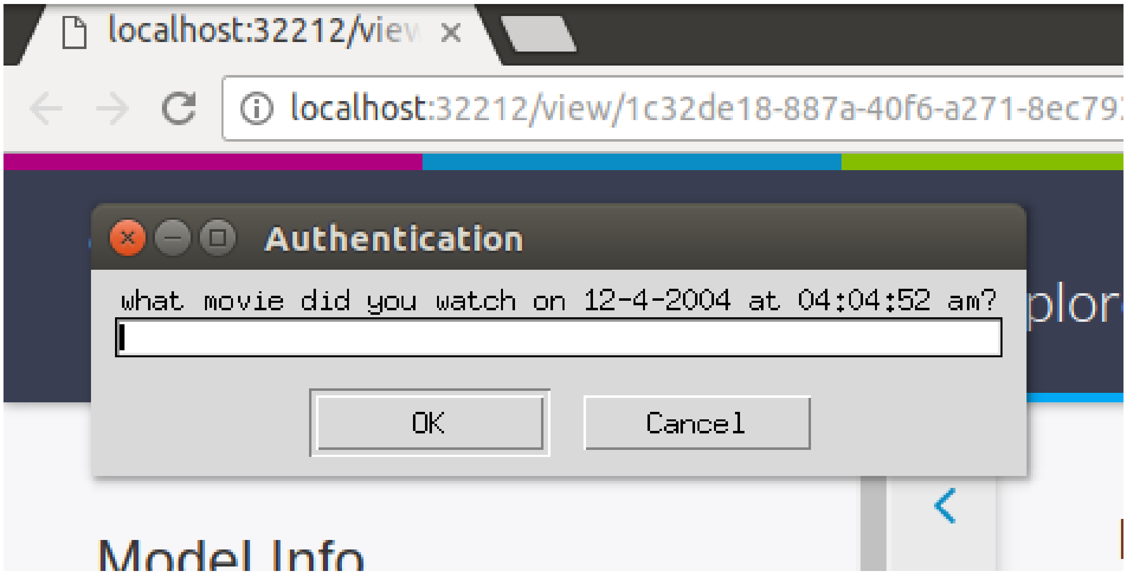
1. 16 GB of Random Access Memory (RAM)
2. Intel I7-6700HQ (4 cores, 8 threads)

Due to the unavailability of desired data to work with, we managed to modify our testing environment to match the data we have, which is a movie database, with ratings of users.

The experiment conducted uses an authentication system that rely on collaborative filtering to aggregate users' information, check irregular behavior, predict actions and generate dynamic challenging questions. Note that the data we have worked with is complete, so the test case study conducted works based on the data we have.

We have specified a certain date to act as current time for the system to use as current date and time in order to check when a user has watched a movie that doesn’t match what have been predicted or recommended, and takes into account the rating and genres of the movies that had been watched. So basically, it checks the date of movies watched with the genres and ratings given to each movie and compares them with what have been recommended, and identifies irregularity accordingly.

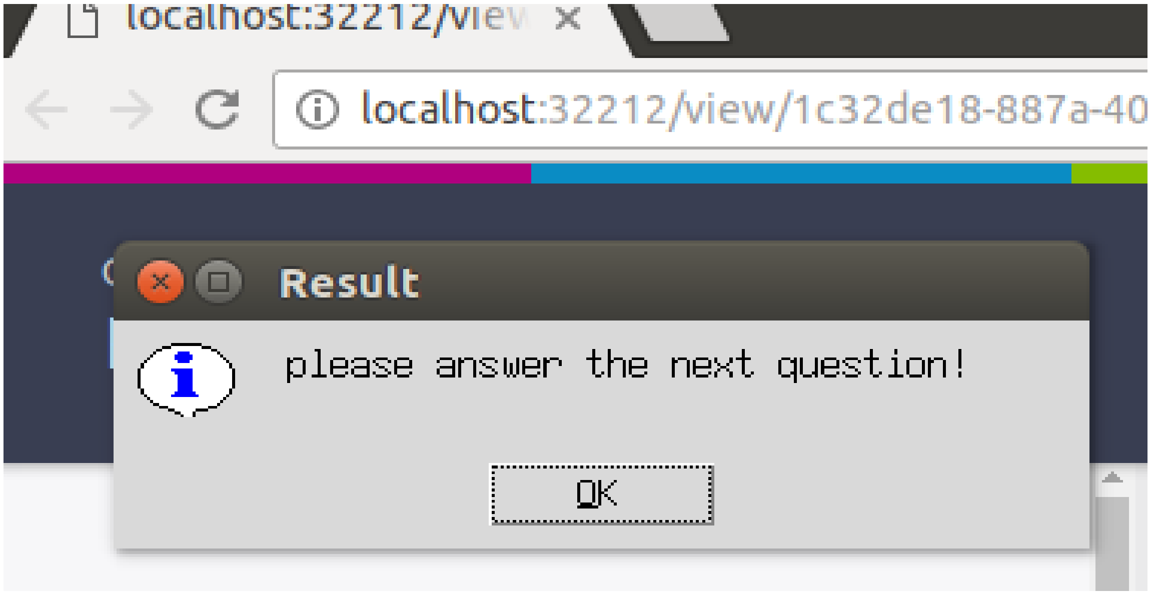
The following scenario is a test case that gives an example with user with id 7000. The system detected an irregular movie that have been watched and only user with id 7000 knows the answer. For easier demonstration, we had to check user 7000’s data to be able to answer the question for the sake of demonstration. But in real life, only the user who is trying to get access to the system will be able to answer the generated questions that are based on day to day behaviors, such behaviors will be collected from different sources, such as phone carriers, ISPs, banks, etc.



what movie did you watch on 12-4-2004 at 04:04:52 am?

Figure 5.10 User accessibility authentication

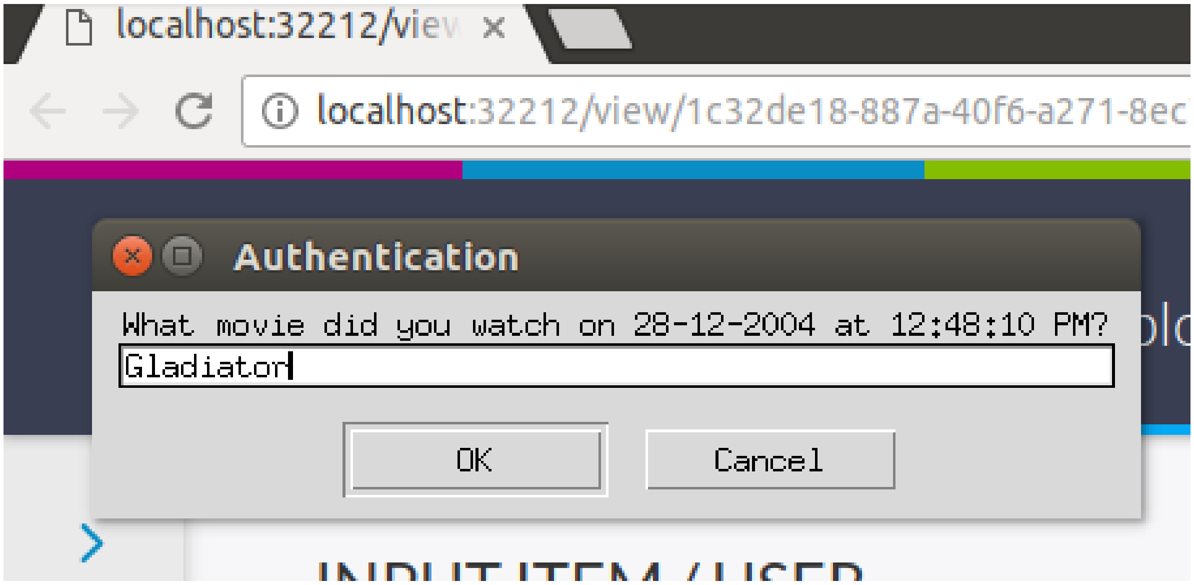
If the answer given is wrong, the system will ask the user to generate a 2nd question and so forth, as provided in figures 5.11, 5.12, 5.13.



please answer the next question!

Figure 5.11 Wrong Answer submission example

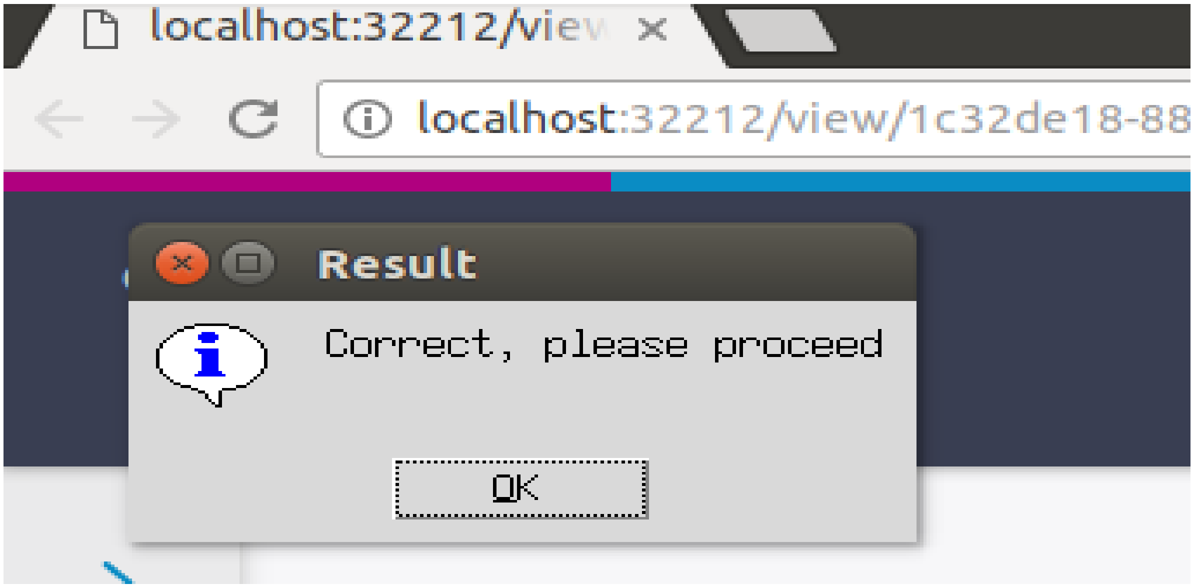
On the other hand, if the user answers correctly, it will simply indicate for the user to continue her/his work.



What movie did you watch on 28-12-2004 at 12:48:10 PM?

Gladiator

Figure 5.12 Answering a different question that is generated based on the user's data

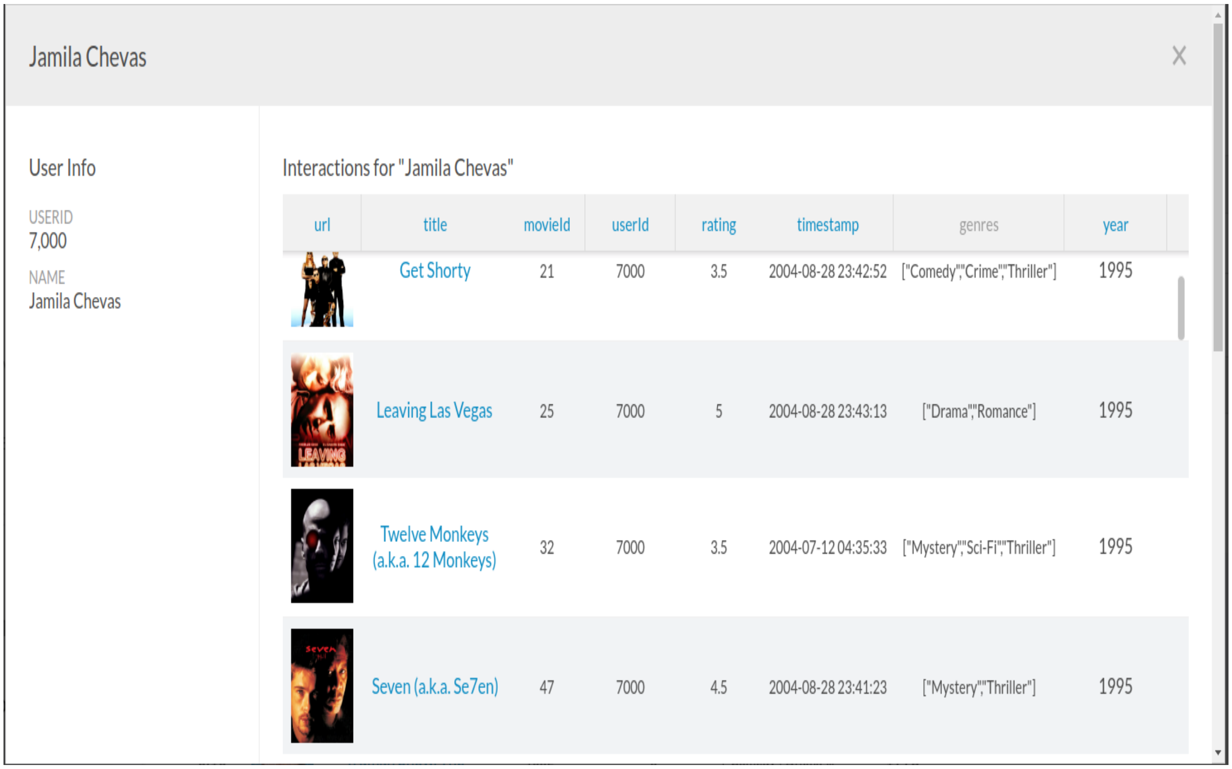


Result

Correct, please proceed

Figure 5.13 Correct answer submission

For our experiment, so after authentication the user will be able to check what he/she watched and what are the recommended movies. The system will keep track of what the user watches and compares the choice with what is recommended as considered regular to watch – Figure 5.14, 5.15 - and will generate questions accordingly.



Jamila Chevas

User Info

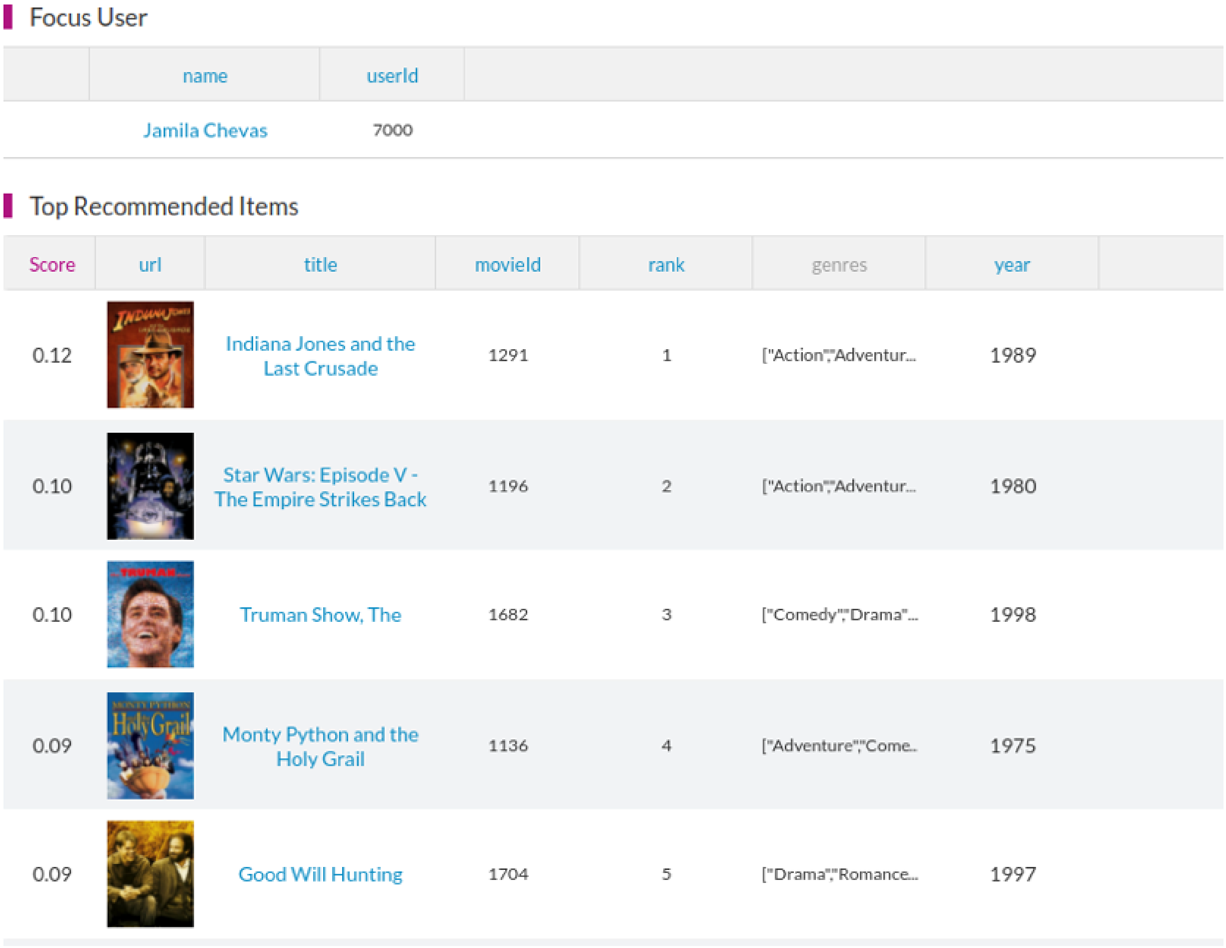
USERID
7,000

NAME
Jamila Chevas

Interactions for "Jamila Chevas"

url	title	movieId	userId	rating	timestamp	genres	year
	Get Shorty	21	7000	3.5	2004-08-28 23:42:52	["Comedy","Crime","Thriller"]	1995
	Leaving Las Vegas	25	7000	5	2004-08-28 23:43:13	["Drama","Romance"]	1995
	Twelve Monkeys (a.k.a. 12 Monkeys)	32	7000	3.5	2004-07-12 04:35:33	["Mystery","Sci-Fi","Thriller"]	1995
	Seven (a.k.a. Se7en)	47	7000	4.5	2004-08-28 23:41:23	["Mystery","Thriller"]	1995

Figure 5.14 what user 7000 watched



Focus User

	name	userId
	Jamila Chevas	7000

Top Recommended Items

Score	url	title	movieId	rank	genres	year
0.12		Indiana Jones and the Last Crusade	1291	1	["Action","Adventur...	1989
0.10		Star Wars: Episode V - The Empire Strikes Back	1196	2	["Action","Adventur...	1980
0.10		Truman Show, The	1682	3	["Comedy","Drama"...	1998
0.09		Monty Python and the Holy Grail	1136	4	["Adventure","Come...	1975
0.09		Good Will Hunting	1704	5	["Drama","Romance...	1997

Figure 5.15 What is predicted for user 7000 to watch

5.4 Summary

This chapter proposed a comprehensive framework for hybrid user authentication that considers behavioral data to identify information of security potential. The proposed approach allows the use of robust methodology that eliminates replay attacks against authentication systems by generating unique dynamic questions based on the users' actions.

Chapter 6

6. Conclusion and Future Work

This chapter presents a concluding summary based on the contribution of the proposed work for a hybrid based system for user authentication that is based on Big data and learning techniques. In addition, a description of possible future research on the proposed approach will be presented.

6.1 Conclusion

The proposed model aims to revolutionize the authentication process by adopting "something-you do" approach. This approach provides all the necessary means in order to take advantage of "something-you do" factor and provide a more convenient and more secure authentication process. The proposed authentication model attempts to generate user profiles, and generate set of challenging questions in real time. These questions cover all actions that represent the instantaneous specific user's behavior. What is special about this approach is that the questions will be chosen in a way that the security and usability requirements are maintained. Each of the questions used will be issued only once; which ensures their uniqueness, to protect users' response from being compromised. This overcomes the pitfalls of KBA methods described in the literature review in chapter 2. And the endless sources of information will guarantee the data being fresh to help legitimate the user to easily remember and successfully complete the challenge.

In addition to the advantages that this approach provides, it faces challenges that might hinder its progress. These challenges are the privacy concerns of users with regard to data that is being used, and the legality of using data from different sources. Also, the idea of a centralized system that contains all the information should be addressed and studied extensively in order to prevent any data breaches. Therefore, the privacy concerns should be addressed in the future, to make sure users' data and information are not compromised, and the study of aggregating data from different sources is done in an optimized manner.

In this thesis, a group of methodologies that tackle the current authentication processes and its drawbacks, and how to overcome them with a new approach that relies on Big data has been presented. The investigation of the current authentication systems resulting in proposing two solutions. The first solution proposed was the IDA model and adaptable mechanisms. And the further investigation of IDA model led to propose the hybrid based authentication model as a second solution.

6.2 Future Work

In this thesis, a methodology for taking advantage of Big data and data analytics has been presented. The proposed methodology has been used to present a case study and a framework to simulate a real-life scenario. In addition, different use case scenarios have been presented and discussed that also show how the application might work in a real-life situation.

As of future work, many aspects of the system must be addressed, such as privacy concerns, availability of data, optimized approach to analyze data and build predictions.

In this study, the proposed framework was designed with the assumption that the system runs in a secure environment and the consent of users were taken to use their data to create user profiles. Future work should address and explore privacy and security issues associated with hybrid based authentication system.

Future work will also explore the use of optimization techniques in data mining and pattern recognition to enhance the efficiency and the performance of the system.

The model implemented to run the evaluation was developed as a proof of concept. A future study must address how to handle bigger datasets and the process of importing data from different sources.

As such more research can be conducted to study the behavior and performance of the system in retrieving data and answering queries. Where efficiency and performance are key factors to any system.

Given that the study was conducted in small environment, a more complex application can be designed and deployed as a software as a service application (SaaS). This will allow to use the service on demand and in different applications and be part of a standard protocol used by systems on the web.

Bibliography

[1] Ross J. Anderson. 2008. Security Engineering: A Guide to Building Dependable Distributed Systems (2 Edition). Wiley Publishing.

[2] F. Monrose and A. D. Rubin, “Keystroke dynamics as a biometric for authentication,” Future Generation Computer Systems, vol. 16, no. 4, pp. 351–359, 2000.

[3] Orcan Alpar, “Intelligent biometric pattern password authentication systems for touchscreens”, Expert Systems with Applications, Volume 42, Issues 17–18, October 2015, Pages 6286-6294, ISSN 0957-4174.

[4] Pin Shen Teh, Andrew Beng Jin Teoh, Connie Tee, Thian Song Ong, “Keystroke dynamics in password authentication enhancement”, Expert Systems with Applications, Volume 37, Issue 12, December 2010, Pages 8618-8627.

[5] Francesco Bergadano, Daniele Gunetti, and Claudia Picardi. 2002. “User authentication through keystroke dynamics”. ACM Trans. Inf. Syst. Secur. 5, 4 (November 2002), 367-397.

[6] B. Schneier, Biometrics: Uses and Abuses, Communications of the ACM, vol. 42, no. 8, pp. 136, Aug. 1999.

[7] A. Jain, L. Hong, and S. Pankanti, Biometric Identification, Communications of the ACM, vol. 43, no. 2, pp. 91–98, 2000.

[8] M. Kuhn, “Security—biometric identification”, 2003, at http://www.cl.cam.ac.uk/Teaching/2003/Security/guestslides-biometric-4up.pdf

[9] Ashfield, J.M. and Shroyer, D.C. and McConnell, E.C. (2014), “Dynamic authentication engine”. US Patent 8,745,698.

[10] Abdelkader Ouda, “A Framework for next generation user authentication, ” The IEEE International Conference on Big data and Smart City, Muscat, Oman, 2016.

[11] A. Ibrahim, A. Ouda, “Innovative Data Authentication Model”, November 2016. 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).

[12] Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor, Recommender Systems Handbook, Springer-Verlag New York, Inc., New York, NY, 2011, pp. 1-35.

[13] L. Page, S. Brin, “The PageRank Citation Ranking: Bringing Order to the Web”, Jan 1999, Stanford Digital Library http://ilpubs.stanford.edu:8090/422/

[14] P. Van Der Graaf, “Panda DNA: Algorithm Tests on the Google Panda Update”, November 2011.

[15] G. Lindem, B. Smith, J. York, “Amazon.com Recommendations item-to-item Collaborative Filtering”, IEEE Internet Computing Vol 7, issue 1, pp 76-80, January 2003.

[16] J. Schafer, J. Konstan, J. Riedl, “Recommender Systems in E-Commerce”, ACM conference on Electronic commerce, pp 158-166, 1999.

[17] V. Natarajan, “How does Facebook’s friend recommendation system work?”, July 2015. Retrieved from https://www.quora.com/How-does-Facebooks-friend-recommendation-system-work

[18] P. Brusilovsky, A. Kobsa, W. Nejdl, “The Adaptive Web”, Library of Congress Controll, Information Systems and Applications, p.325, March 2007

[19] W. Burr, D. Dodson, E. Newton, R. Perlner, W. Polk, S. Gupta, E. Nabbus (Sept 2013), “Electronic Authentication Guideline”, Retrieved from http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63-2.pdf

[20] R. Shirey (May 2000), “Internet Security Glossary”, Request for Comments (RFC) 2828, Retrieved from https://www.ietf.org/rfc/rfc2828.txt

[21] M. Stamp, “Information Security, Principles and Practices “, 2nd Edition, April 2011.

[22] Iridian Technologies, Iris recognition: science behind the technology, Retrieved from http://www.iridiantech.com/basics.php?page=1

[23] F. Merino (2010), “How much would Amazon lose in a DDoS attack?”, Financial trends and news, Retrieved from http://vator.tv/news/2010-12-15-how-much-would-amazon-lose-in-a-ddos-attack.

[24] R. Burgess (2012). Trojan bypass two-facto authentication, steals $46.5 million, TechSpot Newsletter. Retrieved from http://www.techspot.com/news/51037-trojan-bypass-two-factor-authentication-steals-465-million.html.

[25] R. Brandom (May 2016), “Your Phone’s biggest vulnerability is your fingerprint”. Retrieved from The Verge, can we still use fingerprint logins in the age of mass biometric databases? http://www.theverge.com/2016/5/2/11540962/iphone-samsung-fingerprint-duplicate-hack-security

[26] Psylock, (February 8, 2011), Retrieved from http://www.psylock.com

[27] J. McGregor (July 2014), “The Top 5 Most Brutal Cyber Attacks of 2014 So Far”. Retrieved from http://www.forbes.com/sites/jaymcgregor/2014/07/28/the-top-5-most-brutal-cyber-attacks-of-2014-so-far/#6737d3d521a6

[28] M. Darwish, A. Ouda, L. Capretz, “A cloud-based secure authentication (CSA) protocol suite for defense against Denial of Service (DoS) attacks”. Journal of Information Security and Applications, Elsevier (2015), http://dx.doi.org/10.1016/j.jisa.2014. 12.001. Volume 20: 90-98.

[29] Personality Type (June 2016), Retrieved from Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Personality_type#Carl_Jung

[30] K G. Nisha, k. Sreekumar, “A Review and Analysis of Machine Learning and Statistical Approaches for Prediction”, IEEE Inventive Communication and Computational Technologies 2017.

[31] F. Monrose, A. D. Rubin, “Keystroke dynamics as a biometric for authentication,” Future Generation Computer Systems, vol. 16, no. 4, pp. 351-359, 2000

[32] M. Kuhn, Security-biometric identification, Retrieved from http://www.cl.cam.ac.uk/Teaching/2003/Security/guestslides-biometirc-4up.pdf

[33] White paper by RSA, (2013). RSA Risk-Based Authentication. Retrieved http://wbobjects.cdw.com/webobjects/media/pdf/rsa/H11465_RBA_WP_0113.pdf?cm_sp=RSAShowcase-_-Cat1Link4-_-SecurID+White+Paper .

[34] J. Han, J. Pei, Y. Yin, “Mining Frequent Patterns without Candidate Generation”, ACM SIGMOD Record, vol. 29, issue 4, pp. 1-12, June 2000

[35] O. Hasan, B. Habegger, L. Brunie, N. Bennani, E. Daminani, “A Discussion of Privacy Challenges in User Profiling with Big data Techniques: The EEXCESS Use Case”, 2013 IEEE International Congress on Big data, pp 25-30, July 2013

[36] Apache Spark, a general engine for large scale data processing, Retrieved from http://spark.apache.org

[37] Apache, Hadoop, an open-source software projectnf0r reliable, scalable, distributed computing. Retrieved from http://hadoop.apache.org

[38] M. Bezuidenhout, F. Mouton, “Social Engineering Attack Detection Model: SEADM”, IEEE Conference on Information Security for South Africa, pp 1-8, August 2010

[39] Ashfield, J.M. and Shroyer, D.C. and McConnell, E.C. (2014), “Dynamic authentication engine”. US Patent 8,745,698.

[40] S. Sagiroglu, D. Sinanc, “Big data: a Review”, IEEE Conference on Collaboration Technologies and Systems (CTS) 2013, July 2013.

[41] Z. Asad, M. A.R. Chaudhry, “A Two-Way Street: Green Big data Processing for a Greener Smart Grid”, IEEE Systems Journal,vol. 11, issue 2, Feb 2016

[42] C. Eaton, D. Deroos, T. Deutsch, G. Lapis and P.C. Zikopoulos, “Understanding Big data: Analytics for Enterprise Class Hadoop and Streaming Data”, Mc Graw-Hill Companies, 978-0-07-179053-6, 2012

[43] E. Kijsipongse, U. Suriya, C. Ngamphiw, S. Tongsima et al., “Efficient large pearson correlation matrix computing using hybrid mpi/cuda,” in Computer Science and Software Engineering (JCSSE), 2011 Eighth International Joint Conference on. IEEE, 2011, pp. 237–241.

[44] K. Pearson, F.Galton, “Typical Laws of Heredity”, Nature, pp. 492-533, April 1877.

[45] A.M. Medina, “Machine learning and Optimization”, 2014, https://www.cims.nyu.edu/~munoz/files/ml_optimization.pdf

[46] R. Prim, C. Rodrigues, L. Carvallo, “ Cattell’s Personality Factor Questionnaire (CPFQ): Development and Preliminary Study”, jan-apr. 2014, Vol. 24, No. 57, 29-37. doi:10.1590/1982-43272457201405

[47] B.Chen, W.Kuo, L. Wuu, "Robust smart-card-based remote user password authentication scheme", International Journal of Communication systems, Feb 2014, Vol. 27, issue 2 pp. 377-389

[48] R. Amin, GP. Biswas, "An improved RSA based User Authentication and session key agreement protocol usable in TMIS", Journal of Medical Science, Aug 2015, pp 39-79

[49] I. Tsimperidis, V.Katos, "Keystroke forensics: are you typing on a desktop or a laptop?", ACM Proceedings of the 6th Conference in Informatics, pp 89-94, Sept 2013

[50] A. Roy, N. Memon, A. Ross, "MasterPrint: Exploring the Vulnerability of Partial Fingerprint-Based Authentication Systems", IEEE Transactions on Information Forensics and Security 2017, Vol 12, issue 9, April 2017

[51] V. Goel, "That Fingerprint Sensor on Your Phone is Not as Safe as You Think", april 2017, retrieved from New York Times, https://www.nytimes.com/2017/04/10/technology/fingerprint-security-smartphones-apple-google-samsung.html

[52] A.Yassine, S. Singh, A.Alamri, "Mining Human Activity Patterns From Smart Home Big data for Health Care Applications", IEEE Access Journal, June 2017, Vol 5, pp 13131-13141

[53] J. Han, J. Pei, Y. Yin, "Mining Frequent Patterns without Candidate Generation", Conference on the management of Data, ACM Press New York, 2000

[54] J. Bobadilla, F. Ortega, A. Hernando, A. Gutierrez," Recommender Systems Survey", Science Direct ELSEVIR, March 2013

Appendix A

$$sim\ (X,Y) = \frac{\Sigma_{i=1}^{m}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma_{i=1}^{m}(X_i - \bar{X})^2}\sqrt{\Sigma_{i=1}^{m}(Y_i - \bar{Y})^2}}\ (1)$$

m represents number of samples

Xi and Yi are the samples indexed with i

And $\bar{X} = \frac{1}{n}\ \Sigma_{i=1}^{n} X_i$

So Sim (X,Y) = $\frac{\mathrm{n}\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{\sqrt{n\Sigma X_i^2 - (\Sigma X_i)^2}\sqrt{n\Sigma Y_i^2 - (\Sigma Y_i)^2}}\ (2)$

The covariance can be calculated as:

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y}\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

and after rearranging (2) we get sim (X,Y) = $\frac{\Sigma X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\Sigma X_i^2 - n\bar{X}^2\right)}\sqrt{\left(\Sigma Y_i^2 - n\bar{Y}^2\right)}}$

Appendix B

The experiment where the case study was run on a private server with 4 Cores Intel core i7-6700HQ CPU @ 3.50 GHz and 16 GB RAM running Ubuntu 14.04.2 LTS and Ubuntu 16.0.4.

List of libraries used:

Graphlab as gl	GraphLab Create enables developers and data scientists to apply machine learning to build state of the art data products
Graphlab.toolkits.recommender.util import precision_recall_by_user	Compute precision and recall at a given cutoff for each user. In information retrieval terms, precision represents the ratio of relevant, retrieved items to the number of relevant items. Recall represents the ratio of relevant, retrieved items to the number of relevant items.
Tkinter	Tkinter is Python's de-facto standard GUI (Graphical User Interface) package
tkSimpleDialog	This module handles dialog boxes
tkMessageBox	The tkMessageBox module is used to display message boxes in your applications. This module provides a number of functions that you can use to display an appropriate message
pandas as pd	pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both

	easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. pandas is well suited for many different kinds of data: Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet Ordered and unordered (not necessarily fixed-frequency) time series data. Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels Any other form of observational / statistical data sets. The data actually need not be labeled at all to be placed into a pandas data structure The two primary data structures of pandas, Series (1-dimensional) and DataFrame (2-dimensional), handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering. For R users, DataFrame provides everything that R's data.frameprovides and much more. pandas is built on top of NumPy and is intended to integrate well within a scientific computing environment with many other 3rd party libraries.
Os import path	It was used to join paths, where it joins one or more path components intelligently. The return value is the concatenation of _path_ and any members of _*paths_ with exactly one directory separator (os.sep) following each non-empty part except the last, meaning that the result will only end in a separator if the last part is empty. If a component is an absolute path, all previous components are thrown away and joining

	continues from the absolute path component. It was used to read data and join data from different files.

List of Methods used

itemSimilarity.recommender.create()	Creates a recommender that uses item-item similarities based on users in common.
itemSimilarityRecommender-ItemSimilarityRecommender()	A method that ranks an item according to its similarity to other items observed for the user in question using pearson similarity
userSimilarityRecommender-UserSimilarityRecommender()	A method that ranks an item according to similarity between users observed for the user in question using pearson similarity
ItemSimilarity-Recommender.evaluate_precision_recall()	Computes a model's precision and recall scores for the dataset where item similarity has been used.
UserSimilarity-Recommender.evaluate_precision_recall()	Computes a model's precision and recall scores for the dataset where user similarity has been used.
gl.SFrame.read_CSV(path.join(data_dir, ' '))	Constructs an SFrame from a CSV file or a path to multiple CSVs.

userSimModel.recommend(users=range(x,y)), k=…)	Gets recommendations for k users and prints them, users range between x and y where user id of k users is specified.
itemSimModel.recommend(users=rang(x,y), k=…)	Gets recommendations for k items and prints them, item range between x and y where item id of k items is specified.
Item.join():	Joins components of a path when constructing a URL in python.
authenticate()	The method where algorithm 1 has been implemented to check learn and listen for regular and irregular behavior to classify them into two categories where irregular behavior is used in generating challenging questions
graphlab.compare()	Compares the performance of models on the data set used. The two models are user similarity and item similarity.

Curriculum Vitae

Name:	Anas Ibrahim
Post-secondary Education and Degrees:	Beirut Arab University Beirut, Lebanon 2010 – 2015 B.E. The University of Western Ontario London, Ontario, Canada 2015 - 2017 M.E.Sc
Related Work Experience	Network Engineer SAMA S.A.L Offshore. 2011 – 2015 Teaching Assistant. The University of Western Ontario 2015 – 2017 Research Assistant. The University of Western Ontario 2015 – 2017

Publications:

- A. Ibrahim, A. Ouda, "Innovative Data Authentication Model", November 2016. 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).

- A. Ibrahim, A. Ouda, "A Hybrid-based Filtering Approach for User Authentication", May 2017. 2017 IEEE 30th Annual Canadian Conference on Electrical and Computer Engineering (CCECE).