INTRODUCTION TO ALGORITHMS

ALGORITHMS LEADING TODAY'S WORLD: A RESEARCH PAPER

CS 430

PRATISHTHA VERMA

A20357099

Fall 2016

TABLE OF CONTENTS

ABSTRACT

This research paper has been written with the intent to explore and dig deeper into the algorithms leading the world today. These algorithms have been analysed and an insight has been given as to how these algorithms work, there inputs and outputs and how these inputs are converted into these outputs. Hence, an attempt has been made to present a comprehensive structure of the algorithms.

PageRank is an algorithm used by Google which helps to rank various websites according to the keywords searched by the user. Facebook News Feed is an algorithm which not only helps to predict that whether the user would hit the "like" button for a particular post based on his/her past behaviour, but also as to how a user would interact with that post i.e. he/she can share it, delete it or save it and send it as a message to some friend. OKCupid Date Matching algorithm uses the responses of the questions answered by various users and tries to match such similar kind of users by terming them as "soulmates" by relying on the simple fact that higher is the "match percentage" of two individual's answers, more would be there real life compatibility. NSA Data Collection, Interpretation and Encryption are a series of techniques and algorithms utilized by the NSA to analyse the daily activities of millions of Americans by storing the entire records in a bid to predict any future terrorist activity so that any such probable unfortunate event can be foiled in correct time and culprits caught as well. Then, the algorithms utilized by Amazon and Netflix are analysed in detail which are called as "Product Recommendation Algorithms", which help to predict the kind of products that a user is highly likely to purchase the next time based on his/her previous purchase activity and list of searched products. Google AdWords is an online advertising service which helps the advertisers to compete and briefly display a small copy of the advertisements to web users based in part on cookies and keywords by utilizing the very famous Google Ad Rank algorithm. MP3 Compression are algorithms utilized to store and send digital audio files across the Internet by making sure that the space occupied by them is the least. Lastly, the machine learning based advanced algorithms have been discussed which help to predict the future market trends by predicting the stock prices.

INTRODUCTION

An algorithm is a procedure or a formula for solving a problem, based on a sequences of pre specified actions. For instance, a computer program can be regarded as an elaborate algorithm. In the field of mathematics and computer science, an algorithm usually is a small procedure that solves a recurrent problem.

These days, algorithms are being widely utilized throughout all the areas of Information & Technology. For instance, a search engine algorithm usually takes multiple search strings of keywords and operators as input which are entered by the user and searches it's associated relevant database for the probable web pages and returns corresponding web pages.

There are various types of algorithms existing today which are categorized according to their usage. For instance, various sorting algorithms exist today like Quick Sort, Bubble Sort, Selection Sort, etc. which helps us to find a finite amount of data and arrange it properly.

In this research paper we have taken one step further and tried to analyse various higher level algorithms used in the real world by some of the world's most famous and powerful organizations like Google, Facebook etc. which helps them to predict the user behaviour more closely and cater to their needs in a much better way. For instance, Google uses them so that they can display the links to various websites in a much better way while Facebook uses these algorithms to display the news feed to a user in such a way that there is the maximum probability of that user to interact with that post and also to filter out or display those posts at the end which might not be liked by the user.

There are multi-dimensional aspects to these algorithms because the parameters which are of interest to one user may be completely irrelevant to the other. For instance, an article about user x's favourite fiction book may not be of interest to user y, hence a very subtle mechanism plays an important role here which decides that which particular post is to be displayed to which user and how.

Google search

Algorithms basically are computer processes and formulas and instructions, etc. that help to answer a user's query. They rely on more than 200 "signs" or "clues" which help to make it possible to guess to what in particular a user is searching for. These include multiple signs like certain keywords on websites, or the type of the content, the date on which the content is posted, the region from where the user is accessing the webpage, and the most popular algorithm used till date called "PageRank". The better PageRank a webpage has, the higher is the probability of it being found in the google search. Also, each year Google change's its search algorithm around 500 to 600 times, to include new updates to fine tune the entire user search process. Most of these updates are minor, however Google does make major changes which affects the entire search process in a big way. For example, some of them are and Google Hummingbird, Google Mobile Friendly Update, Google Panda Update, Google Penguin Update, Google Pigeon Update, Google Payday Update, Google Pirate Update, Google EMD (Exact Match Domain) Update, etc.

Earlier, Google used to crawl/search the web which was a long process. It would run its algorithm for about 30 days, and then take about a week to index the webpage it found in order to display it to the user. Then, the actual pushing of this data to the search engine would take about another week, making this entire process very cumbersome! Sometimes, it would also find a data centre that had old data and sometimes with a new one. To make this more efficient, after Google crawled for 30 days, it'd re crawl pages with a high PageRank such as the New York Times homepage, to see if anything new or important had been published. But for the most part, this was not a great process, since search results would quickly be out of date considering the 30-day crawl time.

However, these days the scenario is completely changed. Google uses PageRank as a primary source or determinant in its class of ranking algorithms. The better PageRank a webpage has, the higher is the probability of it being discovered and displayed by Google early in the crawling process. For example, crawling in strict PageRank order, Google would find the CNNs and The New York Times of the world, as well as other very high PageRank websites, first. After that, Google used an update called Update Fritz, which significantly changed the search process. Google broke the web into various segments and Google crawled that part of the web, refreshing it every night, or at any given point, Google's main base index would only be so out of date, because then it'd loop back around and refresh it with the newly crawled pages. This was a much more efficient way to crawl since, rather than waiting for everything to finish, Google was incrementally updating its index each and every day to pin point and give near to exact results to the user.

After the web crawling is done, it indexes the web page it finds. For example, Donald Trump are two words that appear next to each other. We have to find the documents in which these words appear in. *Donald appears in documents 1, and 2, and 89, and 555, and 789. And Trump might appear in documents number 2, and 8, and 73, and 555, and 1,000. So that instead of having the documents in word order, we have the words, which are present in document order. Then the other scenario is considered to understand the procedure of how Google ranks webpages in search results. For example,* if web page 1 has 'Donald' but not 'Trump', and web page 2 has 'Trump' but not 'Donald', those two are out of the running. If web page 10 has both 'Donald' and 'Trump', it's a possibility. Furthermore, web pages 89 and 73 would be out because they also don't have the right combination of words. If web page 1023 has both 'Donald' and 'Trump,' it is considered by the PageRank algorithm.

Google tries to find all the pages it believes contains the words searched by the user, either on the page itself, or in backlinks of the webpage, or in anchor text pointing to the webpage. Once Google has completed 'Document Selection' i.e. it decides the documents which it has to consider for the page ranking procedure, it tries to these webpages. Hence, in order to rank any page Google considers not only the PageRank value but also around 200 other important factors. Ranking any page is not easy as it sounds and is a highly complex procedure. For instance, one webpage may have a good reputation because it also has a high PageRank value, but it also may only have the word 'Trump' in it once, and it may have 'Donald' somewhere else on the page. On the other hand, there might be a page that has the words 'Donald' and 'Trump' right next to each other (so it has proximity), and the page also has a good reputation with a lot of web links pointing to it.

Hence, Google tries to balance both the relevancy and authority to extract reputable pages that are also about what the user or the searcher is looking for. But it's not as simple as that, considering Google is taking those 200+ different ranking factors to provide exact to approximate results.

PageRank

PageRank is the most popular search algorithm used by Google to rank websites in their search engine results. It was named after Larry Page. It counts the number and the quality of webpage links to a page to find the rough estimate of how significant a website is. It is assumed here that more important websites are likely to receive more links from other websites and will have a higher PageRank value and greater is the likelihood of it being discovered. It is a link analysis algorithm, and it gives a numerical weighting to each element of a hyperlinked set of documents. This concept may be applied to any collection of entities with reciprocal quotations and references. The numerical weight which is assigned to an element *E* is referred to as the *PageRank of E*.

This results from a mathematical algorithm based on web graph, which describes the directed links between the pages of World Wide Web acting as nodes and hyperlinks as edges, taking into consideration authority hubs such as cnn.com. The rank value indicates as to how important a particular webpage is. A hyperlink to a page counts as a vote of support. The PageRank of a page depends on the number and PageRank metric of all pages that link to it i.e. the incoming links. A page that is linked to by many pages with high PageRank receives a high rank and is much more likely to come up in the web crawl.

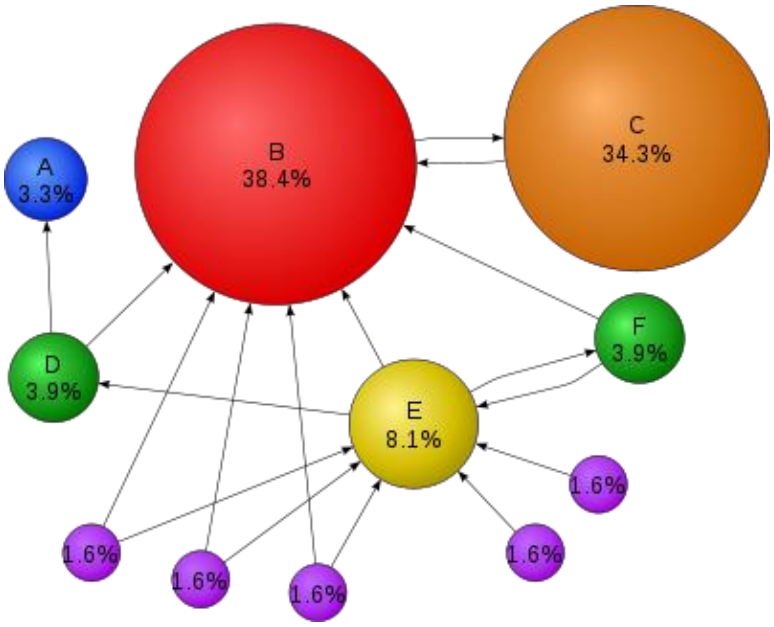

Figure 1 Source: https://en.wikipedia.org/wiki/PageRank

Page C has a higher PageRank than Page E. However, there are fewer incoming links to Page C. Also, one of the link to C comes from an important page B, having the highest PageRank value and hence Page C holds a high value. Considering the fact that web surfers who surf a random page have an 85% likelihood of choosing a random web link from the page they are currently surfing, and a 15% likelihood of jumping to a page chosen at random from the entire web, they will reach Page E 8.1% of the time. Hence, the probability of reaching Page E is much lower than Page C, even though the incoming links are higher in Page E than Page C. Without considering the damping factors, all web surfers would eventually end up on Pages A, B, or C, and all other pages would have PageRank zero. In the presence of damping, Page A effectively links to all pages in the web, even though it has no outgoing links of its own. Hence, the damping factor does play a critical role here.

To understand damping factor, consider an imaginary surfer who randomly clicks on web links and then suddenly stops clicking. However, the probability that this person would continue to click on web links is the damping factor d.

After various studies and multiple researches, it has been concluded that the value is 0.85.

Consider a scenario of the entire World Wide Web, in which only four pages exist A, B, C, D. Any link from a page to itself or any outgoing link is ignored. Hence, all the four pages have same PageRank value. As we take into account the probability distribution, 0.25 value is assigned to all the web pages.

Also, if the only links in the system were from pages **B**, **C**, and **D** to **A**, each link would transfer 0.25 PageRank to **A** upon the next step, for a total of 0.75 PageRank value.

PR(A)= PR(B) + PR(C) + PR(D)

Assume that now page B had link to pages C and A, page C had a link to page A, and page D is linked to all the three pages. On first iteration, page B would give its value 0.125 to A since it is linked to A. Similarly, 0.125 is transferred to page C. Since page C is only linked to page A, it would transfer all of its value to page A. Page D is linked to all the other pages, hence it transfers a value of 0.083 to all of them. Hence, page A will have a PageRank value of 0.458 now.

PR(A)= PR(B)/2 + PR(C)/1 + PR(D)/3

In other words, PageRank value of every webpage is equal to the document's own PageRank score divided by the number of outbound links L( ).

PR(A)= PR(B)/L(B) + PR(C)/L(C) + PR(D)/L(D)

Also, the PageRank value of any webpage u is expressed as;

$PR(u)=v€B_u \sum PR(v)/L(v)$

Facebook news feed

Facebook is an American for-profit corporation and online social media and social networking service based in Menlo Park, California, United States. The Facebook website was launched on February 4, 2004, by Mark Zuckerberg, along with fellow Harvard College students and roommates, Eduardo Saverin, Andrew McCollum, Dustin Moskovitz, and Chris Hughes.

It basically scans and collects all the data posted in the last week which can include either uploaded data or when any user 'likes' by hitting the button. This data is collected at one place for all the friends present in a user's friend list, every user followed by a particular user, and every Facebook page. Roughly, if a person has a few hundred friends in its list, which is a highly likely event, it would be close to 10k posts every week. Now this algorithm plays a major role so as to extract the posts which maybe of utmost interest to a user and has the maximum probability of getting a "like" by that particular user and ranks all of them in a particular order. Most users, considering the time spent by them on Facebook, are highly likely to see only the top few hundred posts.

The exact algorithm used by Facebook remains unknown till date. However, it is assumed that a minor sub algorithm is present within a major algorithm. This entire logic works on the fact that a "relevancy score" is assigned to each of the post present in the user's newsfeed which helps to sort all these posts. However, this score varies from user to user. For example, a particular post about an elementary school may be of interest for a user who had studied there. But the same post, may not be valued by some other user.

The algorithm not only helps to predict that whether the user would hit the like button based on the past behavior, but also as to how a user would interact with that post. For instance, whether the user would comment, share or hide that post, etc. Once, the "relevancy score" has been assigned which is specific to the user and that particular post, the sorting algorithm puts them into an order so that the user can see it. Hence, the topmost post has present in the newsfeed of a user has the highly probability of generating a reaction from the user i.e. in the form of likes, sharing, commenting, etc.

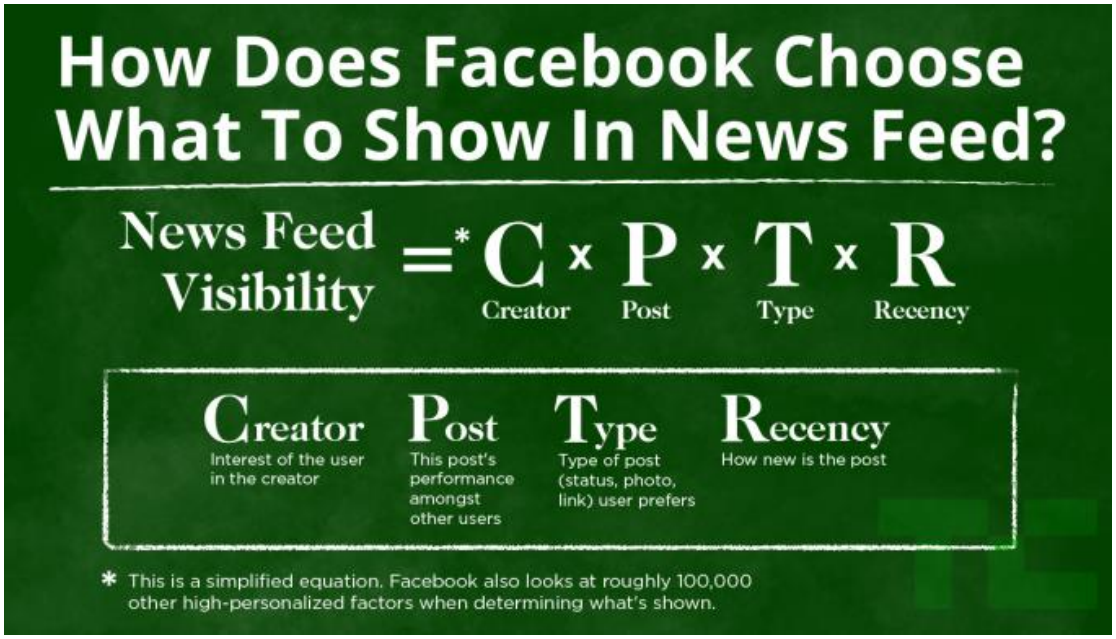

Figure 2 Source: https://techcrunch.com/2016/09/06/ultimate-guide-to-the-news-feed

This figure shows the 4 basic factors considered by Facebook, along with 200 others. “Creator” shows how interested is our user in the Facebook user who uploaded or shared the post. It may be possible that the Facebook user has been classified as a family member, or as someone for whom the user (which we are analysing) has classified to view all the posts first. These kind of changes, gives the user the power to control its news feed and supplies highly accurate and reliable data to the algorithm. “Post” shows roughly the time spent by the user on some ‘post x’, relative to other posts in its news feed. For example, we cannot take any decision on the fact that, if a user spends 2 seconds on post 1, and 10 seconds on post 10, then post 10 is more important than post 2. It may be possible that out of 10 seconds, 7 seconds went on to load the page due to slow internet speed, while the user spent only 3 seconds to actually read it, which roughly is the same time. Hence, a whole lot of factors are taken into account here. “Type” shows the category of the post a user generally prefers to read, for instance the user may be interested in reading other user’s statuses, or watching some videos shared by others. The user may not be interested to view the check-ins made by other users, and hence may never interact with these kind of posts i.e. no commenting, or liking, etc. “Recency” here shows as to how recent a particular post is, either it was shared 1 day before or few hours or few seconds, etc. Generally, more recent posts only are present in a user’s newsfeed to keep them updated.

In a nutshell, Facebook's newsfeed algorithm operates to make sure that the following factors are improved –

a) Higher Quality Posts

This parameter makes sure that the posts displayed to the user are of higher quality i.e. the probability of likes, comments, etc. for that post are greater relative to other posts. Hence, it consists of ranking algorithm, which calculates the score of the news feed story and accordingly displays it to the user.

b) More Relevant Ads

This is an attempt to display only those ads to the users which they might be interested to see and fewer display of those ads similar to the ones which people have already hidden. Every time a user logs in to Facebook, thousands of such ads are filtered to determine the only the best ones i.e. whether from local business or from some well-known clothing brand, etc. Hence, in order to show the most relevant ad, the opinions of both the people and marketers are taken into account. For instance, marketers tell the kind of people who they think might be interested in their ads. The people tell as to what are the kinds of ads they want or do not want to see. So, when a user interacts with certain ad by liking, etc. News Feed gets to know about the particular ad the user is interested to see. Similarly, is the case when the user hides some ad, so it is inferred that the user doesn't wish to see more of such ads. Hence, a continuous optimization continues to improve the ads displayed to the user.

d) Minimizing Hoaxes

Hoaxes are a form of News Feed spam which includes scams (For example, "Click here to win a lifetime supply of body lotions"), or deliberately false or misleading news stories ("Man sees dinosaur in a disco in Utah"). People often share these hoaxes and later decide to delete their original posts after they realize they have been tricked. These types of posts also tend to receive lots of comments from friends letting people know this is a hoax, and comments containing links to hoax-busting websites. As a result, the testing team at Facebook is two times more likely to delete these types of posts after receiving such a comment from a person present in his/her friend list.

Hence, minimization of hoaxes would mean that people would see fewer of such hoaxes after they have marked them to be a hoax. This particular step would mean that the small set of publishers who are frequently into the habit of posting such hoax content would find a significant decrease in the distribution of such posts.

e) Showing live videos when they're live

A user's News Feed is made up of posts from the friends and pages a person has been connected with. These posts can either be status updates, photos, videos and now the newest feature i.e. Facebook Live videos introduced in December in the United States. This feature has particularly become very popular and more and more people and Pages are creating and watching live videos.

Since, more and more people are watching these Live videos, they are being considered as a new content type which are very different from normal videos and an attempt is being made on how to rank them for people in News Feed. Also, an update has been made to News Feed so that Facebook Live videos are more likely to appear higher in News Feed of a user when these videos are actually live, compared to after they are no longer live. Also, it has been proven that statistically people spend more than 3x more time watching a Facebook Live video on average compared to a video that's no longer live since, the Facebook Live videos are more interesting in the moment than after the fact.

f) The See First Feature

This new feature allows a user to choose friends or pages whose stories they want to see first at the top of their News Feed. To implement this, ranking score is used to order stories based on how interesting Facebook believes a particular person or page is to a user, as ideally the user is the only one who knows what is most meaningful to them and whose stories they want to see first. This technique helps to give a ranking score to a news feed activity directly by the user itself and is considered as most trustworthy.

OkCupid date matching

OkCupid or OKC is an American-based international operating free online dating, and social networking website that consists of member created quizzes and multiple choice questions, to be answered by them. Based on the analyses of these responses, the two individual are paired up with each other, which the website calls as “soulmates”. Also, while answering any question, a user indicates his or her own answer, the answers he or she would accept from partners, and the level of importance he or she places on the question. The results of these questions can be made public. Here, multiple modes of communication are utilized which includes instant messaging, emailing, etc.

Here, any adult may join this website and may wish to either use this service for free or pay for it and become an A-List member. Such members do not see any kind of advertising, and have many more filtering options in the “A-List Matches” section of search results. Such members also have control over there profile display i.e. they can choose to display or not their profiles to other users.

This algorithm also assigns a numerical weight value to each and every question answered by the user, that corresponds to a certain “importance rating.” These answers are then compared to “potential matches” in same specified geographic areas. However, this formula is most likely to show “lowest possible match percentage”, a person could have with the compatible other. It also gives an “enemy percentage”, which may represent a raw percentage of answers which do not match i.e. incompatible answers

The OkCupid website relies on the simple fact that higher is the “match percentage” of two individual’s answers, more would be there real life compatibility, and successful relationship is highly likely. However, this supposition is greatly flawed. According to Kevin Lewis, a professor at University of California, San Diego the real life scenario is completely different as relationships do not follow any such statistic and are very subjective too. This algorithm also assumes that people know what they want and hence according to their answers decisions are taken. But in totality, very few people actually exist who know what they want. Highly popular notion of “opposites attract each other” changes the premise here. People do not know what they want and it is very difficult to capture that in words, until & unless they meet such a person and then decide. But this algorithm is more likely to give a compatibility match for two people whose answers match. Thus, the single way concept utilized is flawed here and hence, the algorithm is somewhat crude here. Certain experiments were also conducted on the user here, to come up with an algorithm as accurate as possible:

Experiment 1: The Blind Date Test

Here, a day was decided on which the profile photos were removed for around 7 hours. This was the "Love Is Blind Day" on OkCupid—January 15, 2013.

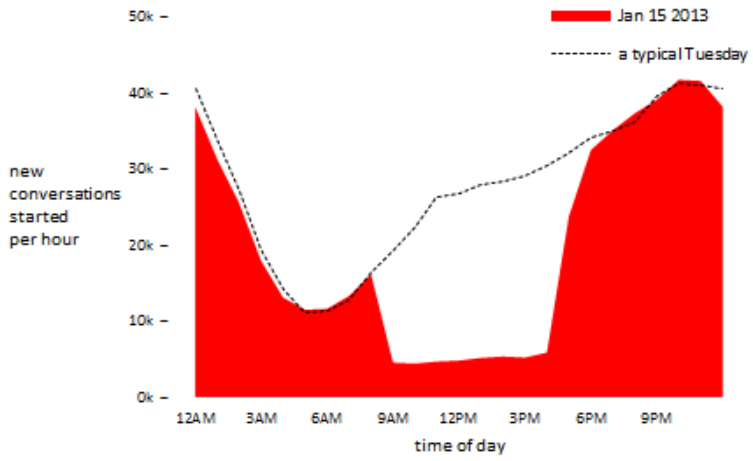

Figure 3 Source: https://blog.okcupid.com/index.php/we-experiment-on-human-beings/

Comparing these statistics to the time after 7 hours, it was found that,

- ➔ People responded to first messages i.e. 44% more often
- ➔ Conversations went deeper
- ➔ Contact details exchanged more quickly
- ➔ Hence, OkCupid worked much better

Hence, when the photos were restored, the conversations were not deeper and some people left online chatting midway too. It was concluded that looks were very important to people, especially to a higher percentage of women, and basically people are as shallow as technology allows them to be.

Experiment 2: The Profile Picture Test

Here, two parameters were chosen to rate a user profile, "looks" and "personality". Each dot here is a person. The two scores are within a half point of each other for 92% of the sample after just 25 votes. Here, according to both these parameters, people consider "looks" and "personality" as same. This means that the face value is same, and a person's thinking doesn't matter.

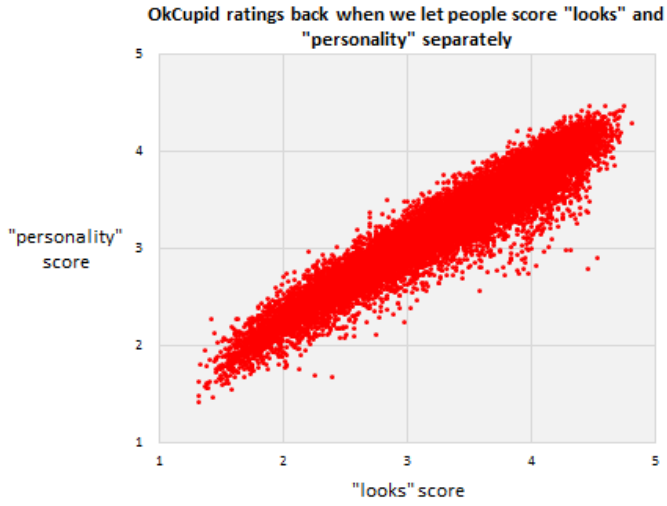

Figure 4 Source: https://blog.okcupid.com/index.php/we-experiment-on-human-beings/

This scenario is slightly modified, and the next experiment was conducted. Here a small sample of users were taken and profile description was hidden. This generated two independent sets of scores for each profile. Score 1 was for profile picture and description. Score 2 was for only "profile picture." Every user is a dot here.

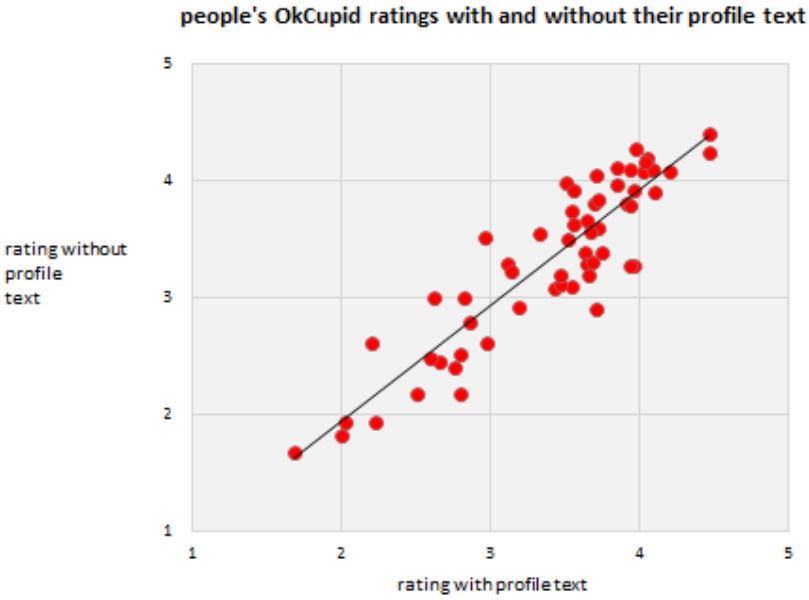

Figure 5 Source: https://blog.okcupid.com/index.php/we-experiment-on-human-beings/

Hence, it was concluded that again profile picture counts for everything while the profile text is worth nothing.

Breaking down human attraction into algorithmic components is very difficult. There are however, 3 pillars of the OkCupid Algorithm, described here:

Pillar One

This pillar consists of the fact that what a person thinks, what he/she wants their life partner to think and how important are answers to those questions answered by the other person. Hence more important is a particular answer, higher is the weight value assigned.

However, this premise is likely to fail because of "opposites attract" notion and the limitations of an individual to correctly express what he/she wants or what he/she doesn't want until they see it. Hence, certain components may cause temporary attraction between two individuals, but none of them are good predictors of true compatibility.

Pillar Two

This pillar doesn't consist much of anything, but tries to express the fact that how self-obsessed an individual can be on such a website to find a person who is similar to them in all aspects.

For instance, if person "a" is a lawyer who likes movies, then person "b" who is also a lawyer but doesn't like movies, doesn't matter in this match making scenario.

Pillar Three

This pillar consists of the following statistic,

Irrelevant = 0
A little important = 1
Somewhat important = 10
Very important = 50
Mandatory = 250

Here, a "little important" is scored higher than "irrelevant". "Mandatory" is 25 times more than "somewhat important". But this grading scale is a bit anomalous in the fact that, 'somewhat important" is 10 times greater than "little important", but "very important", which in a way indicates a much better and a stronger expression of emotions is incremented only 5 times. Hence, the basis on which these people matching decisions are taken are very ambiguous in OkCupid algorithm.

NSA DATA COLLECTION, INTERPRETATION & ENCRYPTION

Earlier, the domestic law enforcement agencies collected the data only after some suspect had been identified. This resulted in lost intelligence and missed opportunities, as some other details would have been found and given a deeper insight about the crime to be conducted and highly important information revealing the suspect but also his/her accomplices involved in it. This can also help the authorities to prevent the criminal act from taking place. This act of data collection which is a kind of proactive measure, can result in identification of new targets as well. These measures were basically taken up after 9/11 to make sure that no such activity happens again in the United States.

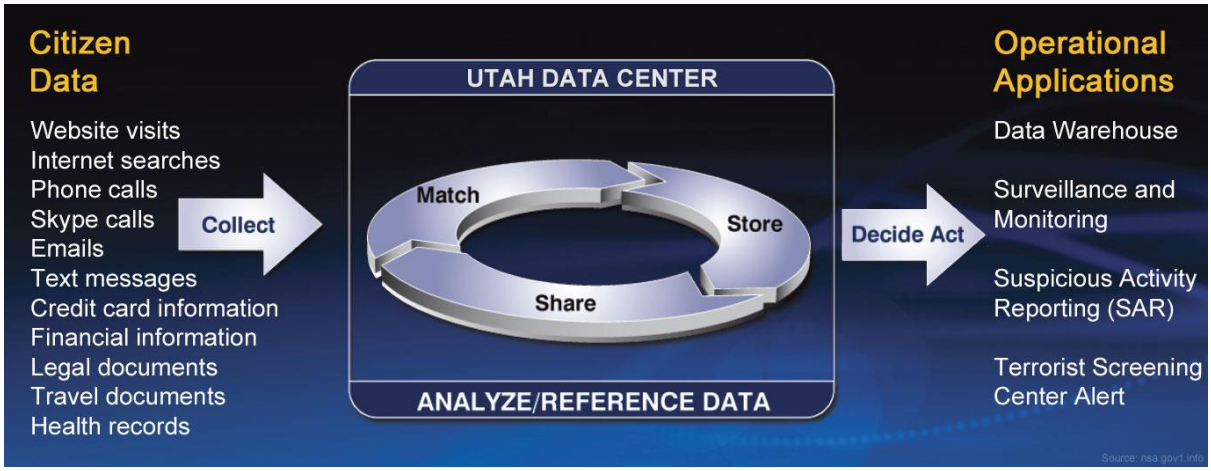

Figure 6 Source: https://nsa.gov1.info/surveillance/index.html

According to the secret interpretations of the PATRIOT ACT, top-secret Fourth Amendment exceptions allowed by the Foreign Intelligence Surveillance Court, and broad cooperation at the local, state, and federal level, a national data warehouse containing information about each and every person in the United States of America can be built.

Everyday people across the United States leave a digital trail of data about their daily routine, for instance which road intersections they have driven through, what roads they have crossed, list of friends and/family members contacted everyday via text or call etc. Since, before any crime is conducted it is very difficult to predict which piece of data may be important and give some clue about any suspicious activity, hence every such data is collected with equal importance.

The data which is collected by the National Security Agency includes internet searches, websites visited, emails sent and received, social media activity, blogging activity, videos watched and/or uploaded online, photos viewed and/or uploaded online, mobile phones apps downloaded, phone call records, text messages sent and received, skype video calls, online purchases and auction transactions, credit card/debit card transactions, financial information, legal documents, travel documents, health records, cable television shows watched and recorded, commuter toll record, electronic bus and subway passes/Smart Passes, facial recognition data from surveillance cameras, educational record, arrest records, driver license information, etc.

There exist multiple techniques to extract data from the users. Some of them are:

a) Google Cloud Exploitation

Here, the NSA "MUSCULAR" program allows them to easily conduct large scale gathering of data outside of the jurisdiction of the Foreign Intelligence Surveillance Court by secretly tapping into the communication links between Google's data centres present outside the United States. The Special Source Operations (SSO) also devised a special way to give complete access of the entire data stored by google for its users.

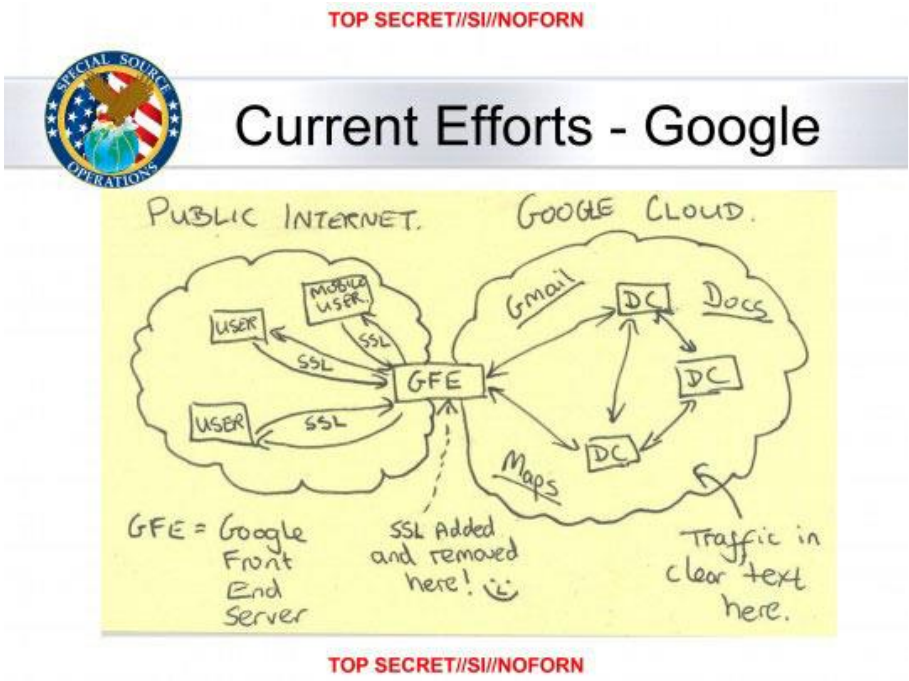

Figure 7 Source: https://nsa.gov1.info/surveillance/index.html

b) NSA Prism Program- Best Source of Raw Intelligence

Here, the NSA's partners at the FBI Data Intercept Technology Unit extracts information from the servers of nine major American internet companies like Microsoft, Yahoo, Google, Facebook, PalTalk, AOL, Skype, YouTube and Apple. This very important partnership has given direct access to audio, video, photographs, and multiple other important documents, which help to closely track targeted individuals over a period of time and closely monitor their activities.

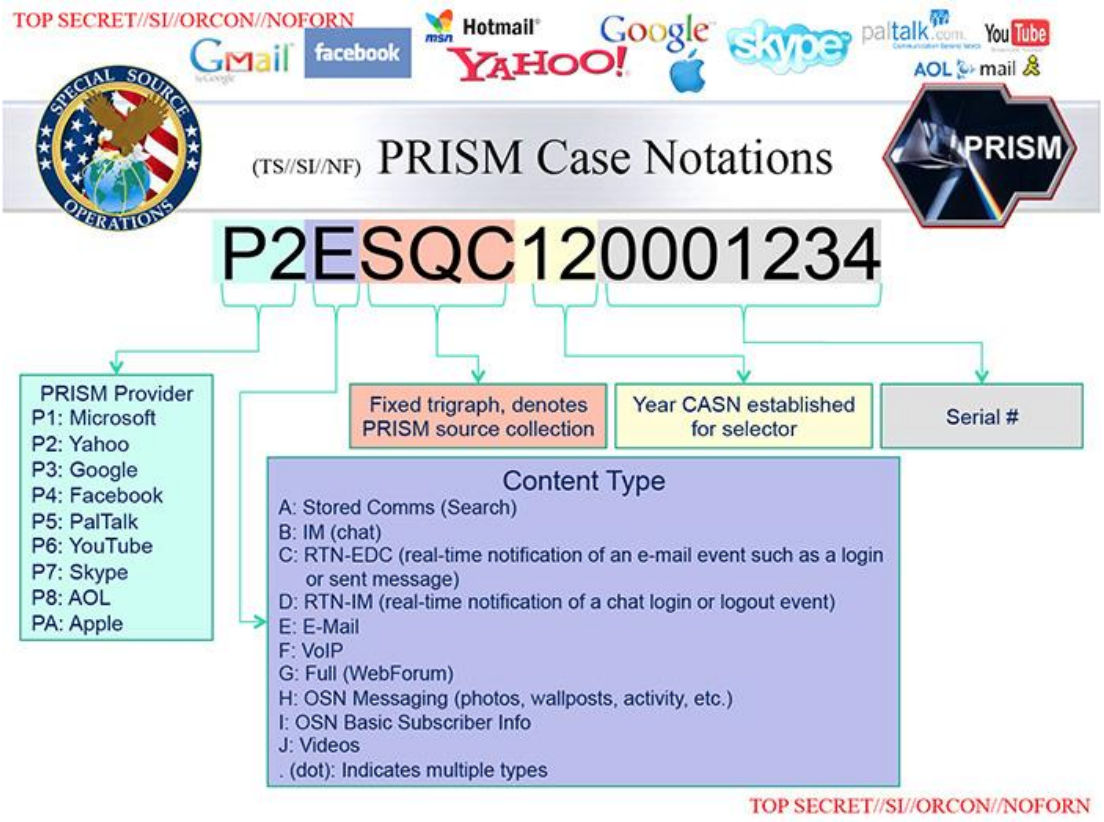

Figure 8 Source: https://nsa.gov1.info/dni/prism.html

This image explains the case file in a PRISM program in which, a tapped or a recorded email message is present, provided by yahoo with its respective PRISM source collection number, selector number and its respective serial number.

Here, encrypting a message involves scrambling it through a combination of randomly generated key and some complex mathematical concepts. The NSA and the UK counterpart of NSA which is GHCQ, consider this as one of the biggest threats in their ability to view and analyse vast amount of data.

Many of the Internet companies have given assurance to its millions of users that the communications carried out between them is fairly secure. For instance, the WhatsApp service introducing the encrypted key feature, etc. However, according to the infamous Edward Snowden leaked documents, many of the US and UK intelligence agencies have broken down and successfully snooped on much of the online encryption.

Much of this snooping was not done by traditional code-cracking, but by working to covertly undermine the international standard on which encryption relies. In order to access even more data, the intelligence agencies have compromised the computers of millions of internet users. However, all these steps were only undertaken with the intention to prevent any other major terrorist activity or potential cyber-attack on any of these countries. In totality, the efforts are really commendable.

Edward Snowden here has endorsed a combination of TOR and PGP (Pretty Good Privacy). The TOR is a network that helps to protect privacy and the physical location of the user by giving anonymity, with multiple volunteers providing communications. The PGP can be effectively and efficiently used to encrypt the data.

I. TOR Network

It is a protective layer present between the user and the internet. It provides an anonymous path between the user and the websites visited by them. The various steps are:

a) User's Computer

TOR program runs on the User's computer. It encrypts and sends all the data to the TOR network.

b) Into the Network

The encrypted information, still considered as unbreakable, is sent into the TOR network.

c) Untraceable

Information travels through the TOR network, taking random paths and simultaneously making the source and destination as untraceable.

d) Document Decryption

Exit node decrypts the untraceable information and sends it to the destination.

e) In the clear

From the exit node, unencrypted but anonymous data flows in and out of the internet.

f) The Internet

Websites view the user currently visiting them but from a random location, not even identifiable to the user as well.

II. Pretty Good Privacy

It is an effective and efficient data encryption technology commonly used to encrypt emails.

A) Document Sending

PGP depends on users, and consists of two keys i.e. public and private. Both these keys have to be used together to perform any kind of encryption and/or decryption.

B) Document Encryption

Sender here uses a random session key to encrypt the file. Message is signed with the sender's private key and encrypted with the receiver's public key.

C) Encrypting File

The encrypted file and the sender's key are sent to the receiver. Here, PGP unlike TOR does not hide the user, but makes the file encryption strategy much stronger and effective.

D) Document Decryption

Receiver here verifies the signature and the sender's public key and performs the decryption with their private key. Encryption is performed using the receiver's public key.

E) Message Delivered

Receiver can then easily replicate this process using the sender's public and private key.

PRODUCT RECOMMENDATION ALGORITHMS IN AMAZON & NETFLIX

Amazon is an American electronic commerce and cloud computing company with headquarters in Seattle, Washington. It is the largest Internet based retailer in the world with respect to total sales and market capitalization. It started as an online bookstore and then expanded its business to sell DVDs, food, toys, jewellery, etc. It is also the world's largest provider of cloud infrastructure services i.e. IaaS. Under Amazon Basics, it also sells certain low-end products like USB cables, etc.

Amazon did introduce one of the biggest innovations in online shopping i.e. using the product recommendation algorithms. Whenever a user is logged in the website, he/she can see suggestions about the products they can purchase. These suggestions are based on the previously purchased products by the same customer or the products which the customer must have searched on their website but not purchased. For instance, an avid fiction book reader would see multiple recommendations for fiction books whereas a mother with young kids would see a mention of toys and children's books. This homepage personalization strategy adopted by amazon has led to a substantial increase in the total sales and the overall likability of the website.

Amazon has millions of customers present all over the world. New customers have very less information about their preferences while regular customers know the entire dynamics of the website. Hence, the data on which these algorithms operate are constantly updated and changed every minute. The second most important factor is the speed with which the product suggestions are given to the customer. They must get displayed within seconds and with accuracy as well. These algorithms have worked by finding a set of customers who have bought or rated the same items as some other customers. Hence, it attempts to logically group similar customers i.e. create a cluster of customers.

Collaborative Filtering Algorithms

These algorithms represent each customer as a vector of all the products on sale. Each entry in the vector is positive if the customer has bought or rated the item. The entry is negative if the customer gave a negative review about it or the entry is empty if the customer made his/her opinion unknown.

Generally, the entries remain empty for most of the customers registered on the website. Some variant factors in the popularity of the items are used to indicate the significance of the items that are either less popular or familiar. Recommendations are then created by the algorithm by finding a similarity value between the current customer and all the other registered customers on the website.

The most obvious way to find the "angle" value between the vectors is by calculating the cosine value using the dot product divided by the product of the vector lengths. Hence, larger is the cosine value, smaller is the angle and hence more similar are the customers.

However, all these calculations are computationally expensive due to the presence of millions of customers and billions of such expensive calculations have to be carried out. There are minor ways to reduce the number of calculations by sampling the customer base or ignoring the unpopular items, however due to the humongous amount of data involved the calculations are still very expensive and sometimes require greater amount of precision and accuracy in the result generation.

Customer Clusters Algorithm

This algorithm involves the use of cluster models. The goal is to prepare the customer base by dividing it into multiple different clusters and then assigning the currently logged in customer to one of the clusters. The current customer is assigned to only that cluster in which similar customers are present. Once the cluster has been identified, the product recommendations comes from the purchases, ratings etc. done by the customers present in that cluster only. The most difficult part is the creation of clusters in this algorithm.

Clustering of customer data is done by creating a random number of empty clusters and assigning any randomly selected customer to each of these clusters based on the similarity quotient. Since all these clusters are created without any link to any other cluster, certain other sub algorithms etc. must be used to combine or divide clusters as they are being constructed.

The concept of cluster models being used to generate product recommendations is very computationally less intensive as compared to other techniques. This is because, since multiple customers who have given similar ratings, etc. to a product are grouped together, any other recommendations displayed are always the same for at least that cluster of customers. Most of the intensive work and calculation is generally carried out during the cluster creation itself. However, this method tends to produce low quality recommendations as the purchases/ratings are averaged out within a cluster itself. The number of clusters can still be increased though and the final matching done to be more refined, but the computation time would exponentially increase.

Item to Item Collaborative Filtering

This algorithm is a modified version of the collaborative filtering algorithm. Instead of matching customers, it tries to match the match the items purchased. For instance, for each item X in the catalog, find all the customers C who purchased the product X. For only this particular group of customers, find all items Y purchased by C and record the names of those customers who bought both X and Y. Then for all such pair of products i.e. X and Y, calculate the similarity between X and Y in the same manner as collaborative filtering algorithm. Although the calculations done here are computationally expensive, it can be carried out beforehand and hence, once the similarity values between every pair of items have been determined the recommendation list follows easily.

Netflix Recommendation Algorithm

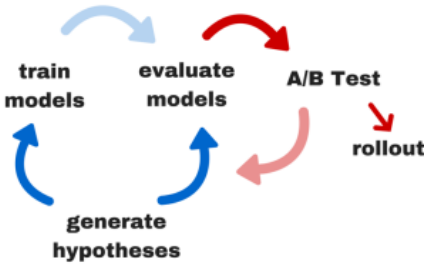

Figure 9 Source: https://www.rtinsights.com/netflix-recommendations-machine-learning-algorithms

Earlier Netflix relied heavily on the ratings given by customer when they shipped DVDs by mail. However, these days Netflix does rely on multiple parameter based huge data set. This includes what each and every registered member watches, the time they watch, the length of time they spend on Netflix, the popularity of videos in the catalogue, and various other important parameters.

This multi parameter data goes into several different algorithms powered by statistical and various machine learning techniques. Here both the supervised techniques i.e. classification and regression & unsupervised techniques like dimensionality reduction through clustering or compression are utilized.

A personalized video ranker algorithm or PVR selects the order of videos in multiple category rows, movie genre wise and then uses an arbitrary subset of the Netflix catalogue. Hence, closer a particular movie title is to this subset, the more likely it is to get played.

Top N Ranker is another famous algorithm utilized by Netflix to recommend movie titles. Also, a “Trending Now” row structure is utilized which uses short term trends, for instance an interest in action movies, etc. These trends are powerful and important predictors of videos and movies, etc. that is more likely to be watched by Netflix’s registered users. However, Netflix does make sure not to over personalize the recommendations. This is because, then the viewer would only get to view the movie recommendations based on the shows previously watched by them. Hence, the option to explore different choices which the user might like are fairly reduced, making the recommendation scenario dull or uninteresting. Hence, efforts are made to keep a balance in all these recommendations.

GOOGLE ADWORDS

Google AdWords is an online advertising service which helps the advertisers to compete and briefly display a small copy of the advertisements to web users based in part on cookies and keywords which are already determined by the advertisers. Google's Web pages and the web pages from partner websites are designed to allow Google to select and display the advertising copy as selected. Advertisers are asked to pay only when the user divert their browsing to find more information about the ad and the partner website also receives a copy of the income generated.

An algorithm used by Google to place advertisements on the search results page, whenever any user enters a search string is termed as "Ad Rank algorithm". It is based on "Quality Score". This score basically means that how well a particular group of ads, keywords and landing page etc. relate to what a user is searching for and the possibility of the user to click on that particular ad. There are multiple parameters connected to this advertisement scenario on Google.

Ad Rank is a value that is used to determine the ad position i.e. the location on a page where the ads would be displayed and whether they would be displayed at all or not. The main components of Ad Rank are the bids and quality of ads and the website. Google also incorporates the expected impact from the extensions and other ad formats while computing Ad Rank. Ad Rank is determined by using bid amount, Quality Score components i.e. expected clickthrough rate, ad relevance, etc. Hence, even if a particular user's competitor has a higher bid amount, the earlier user would still be able to win even at a lower price by using highly relevant keywords and ads. Ad Rank is calculated every time the ad is eligible to appear and competes in the auction, so the ad position can fluctuate based on the competition at that moment.

One of the major parameter of this entire Google AdWords algorithm is the Quality Score. It helps to give a more generalized view of the quality of ads. It is present in the range of 1-10, which is reported for each keyword and also the target pages triggered by them. The factors which determine the quality score are:

i) Expected clickthrough rate (CTR)

It is a keyword status that measures that how likely it is that a particular ad would get clicked by user when shown for that keyword, irrespective of the ad's position, extensions, and other ad formats that may also affect the visibility of the ad and also indicates that whether a particular keyword is likely to generate clicks on the ads. CTR AdWords which provides for a keyword in the account is an estimate based on the assumption that the search term would match that particular keyword or not.

At auction time, AdWords gives a more accurate expected CTR based on the search items, type of device, and other auction time factors, etc. There exist 3 possible status values, which are “above average”, “below average” and “average”.

An “average” or “above average” status means that there are no major problems with this keyword’s CTR in comparison with all other millions of keywords across “AdWords”. A “below average” status means that the customer who is advertising on Google should consider modifying the ad text so that it is more closely related to the top keywords. It can also be used to help identify keywords that might not be relevant enough to perform well.

ii) Ad Relevance

This status describes as to how well the user entered keyword matches the message in advertisements. There are 3 possible statuses that are possible:

a) Above Average / Average: This means that no major problems exist with the particular keyword’s ad relevance as compared to all other keywords across AdWords.

b) Below Average: This means that the ad or keyword is not specific enough or that a particular ad group may not be able to cover all the topics.

It is possible that a keyword has a high Quality Score and low ad relevance or vice-versa, since Google AdWords looks at a number of different quality factors when finding out the Quality Score. Even if this score is high, the individual factors can be studied closely to find out any further areas for improvement.

iii) Landing Page Experience

This describes that whether the landing page is likely to give a good experience to those users who click on that particular displayed ad and view the retailer’s website. This can be used to check that whether the user’s click is getting converted to some sales or signups on that website, etc. It has to be made sure that the landing page is clear and useful to all the customers and is clearly related to what a particular user is searching for. This can also have one of the three statuses listed below:

a) Average/ Above Average: Here, no major problem exists with the keyword’s landing page experience as compared to all the other keywords listed across AdWords.

b) Below Average: It means that some effort is required from the retailer’s side to improve the quality of their web pages, since the number of clicks should get converted into purchased goods or at least signups by the user.

Hence, the quality score is an aggregated estimate of how well a particular keyword has performed overall in the previous ad actions. Based on all these activities, each keyword gets a Quality Score on a scale from 1 to 10, where 1 is the lowest and 10 is the highest.

Null Quality Score means that aren't enough clicks done by the user to find out a keyword's quality score. However, it is not used at auction time to determine the Ad Rank Value.

Hence, the Ad Rank value is calculated in the instant some user does a search that causes or triggers an ad to compete in the ad auction. Here multiple real time signals are taken into account such as the query and user context e.g. type of device, language preference, etc. in order to find out more precise values of CTR, etc.

Working of Google AdWords

Every time a user searches something on Google, an AdWords auction is immediately created. All the advertisers who already have a keyword which matches with the search query would be competing in this auction. All the ads are placed according to their Ad Rank values i.e. higher the values, topmost is the position they will be places at. The formula utilized to calculate Ad Rank is:

Ad Rank = Quality Score * Bid

In order to beat the Ad Rank of the competitor, an advertiser actually has to pay the lowest amount, called the discounter. The formula used to calculate the amount is:

$$= Ad Rank to beat/ Quality Score + $0.01

Consider that a certain keyword has a quality score of 8 i.e. QS=8. Since the advertiser wants to be shown in the top position, the bid amount set by them is $100 per click.

Ad Rank= QS 8 * $100 Bid= 800

To beat this, the competitor would naturally want to have an Ad Rank value greater than 800. Assuming that they have a QS = 10 and a bid amount of only $10 per click. Hence, they have a lower bid amount but a very high quality score value. Thus, the Ad Rank would be;

Ad Rank = QS 10 * $10 Bid= 100

Actual amount payable by would be:

$$= Ad Rank to beat / Quality Score + $0.01

= 100 / 8 + 0.01 = $12.51

Hence even though, the bid was $100 for every click, the overall amount payable was only $12.51. In way, the advertiser whose ad is at the top has actually paid, an amount lesser than the ones below it.

Consider another instance, in which a $100 bid is set for the top most position by competitor C1, while the next competitor C2 has a QS 5 and a $10 bid. Competitor C3 has a QS 7 and a $7 bid while competitor C4 has a QS 5 on a $9 bid. The calculations for the top 3 positions are shown below:

Position 1:

Ad Rank to beat = 50 = (QS 5 * $10 Bid)

$6.26 = 50 / 8 + $0.01 (50 Ad Rank to beat / QS 8 + $0.01)

Position 2:

Ad Rank to beat = 49 = (QS 7 * $7 Bid)

$9.81 = 49 / 5 + $0.01 (49 Ad Rank to beat / C2 with QS 5 + $0.01)

Position 3:

Ad Rank to beat = 45 = (QS 5 * $9 Bid)

$6.44 = 45 / 7 + $0.01 (45 Ad Rank to beat / C3 with QS 7 + $0.01)

Hence here, the QS 8 competitor present at position 1 has actually paid lesser amount than the ones present below it. It can be concluded that a competitor with Quality Score 10 dominates the entire auction process. It also implies that not only the least amount would have to be paid for the topmost position, and also its popularity is the highest, this scenario also forces every other competitor to pay more. Hence, this algorithm entirely depends on the Quality Score i.e. better it is, higher are the chances of getting the topmost position irrespective of the bid amount.

MP3 COMPRESSION

MP3 is an audio coding format used for storing audio in the digital format, using the lossy data compression techniques which occupies 10 percent of less space than needed by the original uncompressed data. It is the most widely used standard for digital audio compression in order to play the music on most digital audio players like iPod, etc. and almost all the computing devices like laptops, cell phones etc.

The audio files on a compact disc is converted from an analog source, for instance the master tape. The analogue wave here can't be recorded digitally, hence a digital audio processor is utilized to sample the analogue audio wave 44,100 times a second. Hence at every tick, the digital audio processor works out the amplitude of the original very difficult audio wave.

The second big task is the recording of the sound waves. The entire process is given by the below diagrams. Consider that we are sampling the waveform given below:

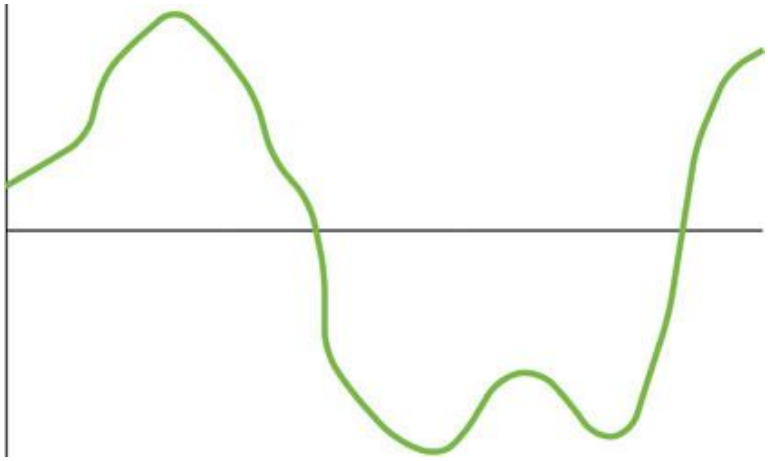

Figure 10 Source: http://www.techradar.com/news/audio/how-mp3-compression-works-916093

If the sampling of the above waveform is done at a lower rate, some of the peaks and troughs are missed, and the resulting waveform is very different and unclear. Hence, it is very necessary to sample it correctly.

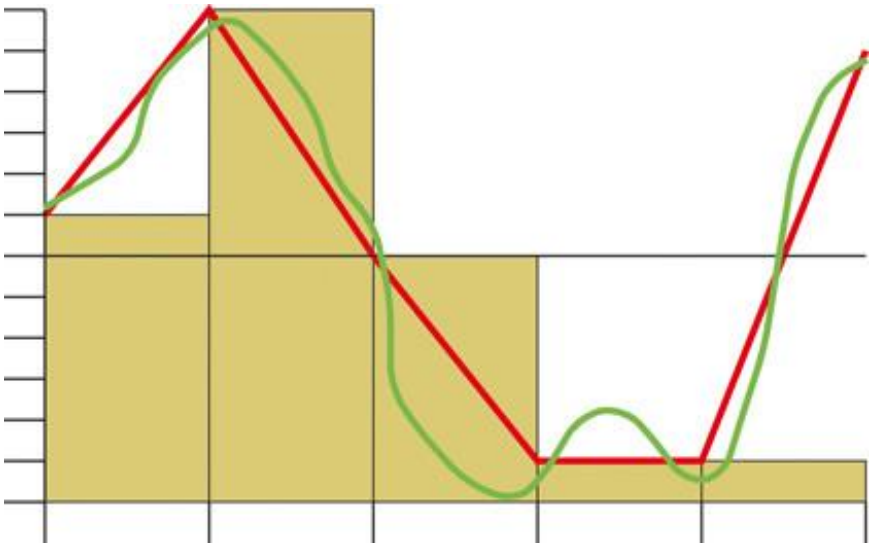

Figure 11 Source: http://www.techradar.com/news/audio/how-mp3-compression-works-916093

Here the above waveform is depicted in “red” and it looks different from the original waveform. This basically shows that the sampling needs to be carried out much more often. Since the human ear is able to hear only up to 20 KHz at the maximum, the sampling is thus needed to be done twice so to properly capture the highs and lows of the audio wave at that particular frequency. With the earlier fudge factor added, the rate was only 44,100 Hz.

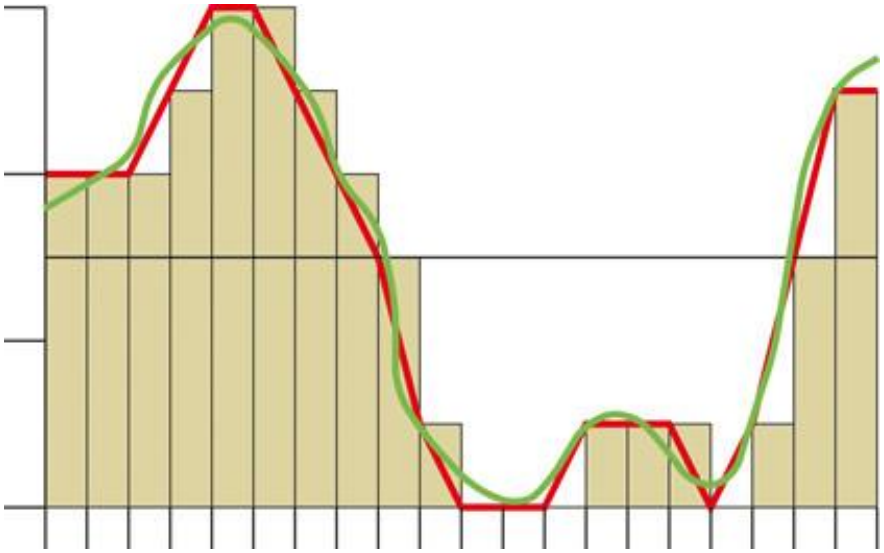

Figure 12 Source: http://www.techradar.com/news/audio/how-mp3-compression-works-916093

The above depicts a completely different scenario in which the number of possible values for the amplitude are fairly small. So, from the original amplitude the processor must be able to choose the closest values which can be recorded. Hence, even due to high sample rate, the measurements of the amplitude are pretty difficult.

The MP3 compression is based upon the psycho-acoustic model of compression of human sensitivity towards frequencies. Humans can hear frequencies in the range between 20 Hz to 20 KHz and is most sensitive between 2 KHz to 4 KHz. Thus, MP3 carries out "destructive compression" in which all the frequencies in the non-audible range are rejected while others are selected. However once deleted, these frequencies can never be removed. Whenever any file is encoded into MP3, different compression levels can be chosen. Also size of file is directly proportional to quality. Thus, an MP3 file created with 128 Kbit compression will have a much greater quality than that of a 56 Kbit compression.

The MP3 format utilizes an "hybrid transform" technique to convert a time domain signal into a frequency domain signal. This model is same for all the three layers of audio defined by MPEG, but the codec complexity multiplies with each layer. Here, the codec divides the data into frames and each frame has a total of 384 samples, 12 samples from each of the 32 filtered sub-bands.

The various steps in the MP3 algorithm are:

a) Convolution filters are used to divide the audio signal into frequency sub-bands that approximate to around 32 critical bands.

b) According to the psycho-acoustic model, the amount of masking is determined for each band respectively caused by the nearby band.

c) If the power is within acceptable limits, then only the number of bits which are needed to represent the coefficient such that the noise is well below the masking effect is thus found.

d) The bit stream is thus formatted.

Here, a high quality critical band filter is utilized for MP3. Also, the pyscho-acoustic model includes temporal masking effect, taking into account stereo redundancy and usage of a Huffman coder.

ALGORITHMS TO PREDICT MARKET FLUCTUATION USED IN FINANCIAL INSTITUTES

In the financial markets these days, genetic algorithms are most commonly used to find the best combination values of parameters in a trading rule, and they can be built into artificial neural network models in order to choose stocks and identify those trades prevailing in the market. Genetic algorithms (GAs) are basically problem solving methods (or heuristics) that mimic the process of natural evolution of the human race. Unlike artificial neural networks (ANNs), which are designed to function like neurons in the brain, these algorithms utilize the concepts of natural selection to find out the best solution for a problem. As a result, GAs are commonly used as optimizers that help to adjust those parameters or factors which can help to minimize or maximize some feedback measure, and then be used separately or even in the construction of an artificial neural network.

These genetic algorithms are created mathematically using vectors, which are mathematical quantities having a direction and magnitude. Parameters for each trading rule which are commonly used in the stock market prediction are represented with a one-dimensional vector that can be considered as a chromosome in genetic terms. Similarly, the values used in each parameter can be thought of as genes, which are then modified using natural selection.

For instance, a trading rule may involve the use of parameters like Moving Average Convergence-Divergence (MACD), Stochastics etc. A genetic algorithm would then insert values into these parameters with the goal of increasing the overall net profit. Over time, small modifications are introduced from time to time and the ones that make a desirably impact are retained for the next generation.

Three types of genetic operations that can then be performed are:

a) Crossovers represent the reproduction and biological crossover, whereby a child takes on certain characteristics of both its parents.

b) Mutations represent biological mutation and are used to maintain genetic diversity from one generation of a population to the next by introducing random small changes.

c) Selections are the stage at which individual genomes are chosen from a population for later recombination or crossover process.

These three operators are then used in a five-step process:

1. Initializing a random population, wherein each chromosome has *n*-length, with "n" equivalent to the number of parameters. Hence, a random number of parameters are then established with "n" elements each.
2. Select only those parameters that increase desirable results i.e. the net profit over the choice of stocks.
3. The mutation or crossover operators are then applied to the selected parents and generate an offspring or the results of a trade.
4. Recombine the offspring and the current population to create a new population of off springs with newer characteristics with the desired selection operator.
5. Repeat steps from two to four.

Over a period of time, this process will result in increasingly favorable parameters to be used in stock trading rule. The process is then terminated when a stopping criteria is met i.e. when the desired characteristics are achieved. These genetic algorithms are primarily used by institutional quantitative traders, and the individual traders can utilize the power of genetic algorithms - without any kind of formal knowledge in advanced mathematics or statistics and by using several software packages available in the market. When using these applications, traders can define a set of parameters that are then optimized by using a genetic algorithm or a combination of such algorithms and also a set of historical data which might indicate the prices of previous stocks and predict ones and come up with the best trade. Some applications can help to optimize parameters which are used and the values for them, while others are simply focused on optimizing the overall values for a given set of parameters. Hence here, designing or modelling a trading system around the historical data rather than designing this system around repeatable behavior would represent a potential risk for traders by using genetic algorithms. Hence, any kind of trading system which is designed using these algorithms could be practiced on paper before any live-usage to prevent any kind of potential loss in case some prediction comes to be incorrect.

Prediction of the trend of stock market could either represent an increase or decrease in the overall stock trading. As a result, the change of a feature over time is more important than the absolute value of each feature.

We define xi(t), where i $\in$ {1,2,..., 16}, to be feature i at time t. The feature matrix is given by $F = (X1, X2, ..., Xn)T$ (1)

where,

$Xt = (x1(t),x2(t),...,x16(t))$ (2)

New features which are the difference between two daily prices can be calculated by

$$\nabla\delta\ xi(t) = xi(t) - xi(t - \delta) \quad (3)$$

$$\nabla\delta\ X(t) = X\ (t) - X\ (t - \delta) = (\nabla\delta\ x1(t), \nabla\delta\ x2(t), \cdots \nabla\delta\ x16(t))T \quad (4)$$

$$\nabla\delta\ F = (\nabla\delta\ X(\delta + 1), \nabla\delta X(\delta + 2), \ldots, \nabla\delta\ X(n)) \quad (5)$$

Due to the difference in market value and basis of each market, the differential values calculated above can vary in a wide range. To make them comparable, the features are normalized as following:

$$\begin{aligned}
\mathcal{N}(\nabla_\delta x_i(t)) &= \frac{x_i(t) - x_i(t-\delta)}{x_i(t-\delta)} \\
\mathcal{N}(\nabla_\delta X(t)) &= (\mathcal{N}(\nabla_\delta x_1(t)), \cdots, \mathcal{N}(\nabla_\delta x_{16}(t)))^T \\
\mathcal{N}(\nabla_\delta(\mathcal{F})) &= (\mathcal{N}(\nabla_\delta X(\delta+1)), \cdots, \mathcal{N}(\nabla_\delta X(n)))^T,
\end{aligned}$$

Figure 13 Source: Shunrong Shen, Haomiao Jiang, & Tongda Zhang. (n.d.). *Stock Market Forecasting Using Machine Learning Algorithms*

and the normalization can be implemented as:

$$normal(X(t)) = \frac{\mathcal{N}(\nabla_\delta X(t))}{|\mathcal{N}(\nabla_\delta X(t))|}$$

Figure 14 Source: Shunrong Shen, Haomiao Jiang, & Tongda Zhang. (n.d.). *Stock Market Forecasting Using Machine Learning Algorithms*

The performance of a stock market predictor here heavily depends on the correlation between the data used for training and the current input for prediction. However, if the trend of stock price is always an extension to yesterday, the accuracy of prediction should be fairly high and has a higher chance of approximation. Here, the autocorrelation and cross-correlation of different market trends i.e. increase or decrease has been calculated. The results shown in the next diagram use NASDAQ as the base market to represent a broader view.

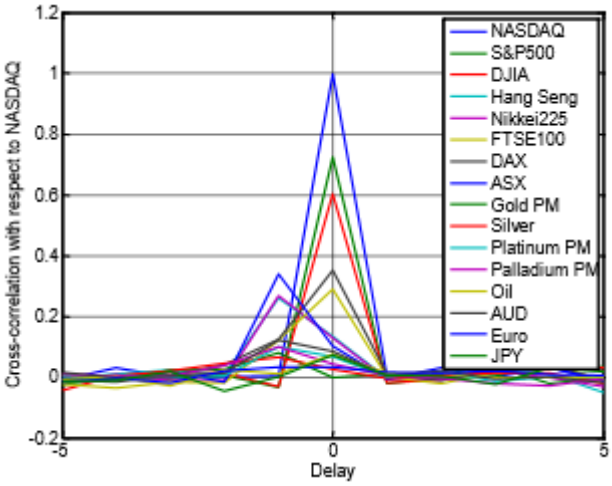

Figure 15 Source: Shunrong Shen, Haomiao Jiang, & Tongda Zhang. (n.d.). *Stock Market Forecasting Using Machine Learning Algorithms*

From the graph it can be seen that the auto-correlation of NASDAQ daily trend is only non-zero at the origin based on which it can be concluded that the trend of NASDAQ daily index is approximately and essentially a Markov process. However, past data of NASDAQ will not provide much insight to its future movement as it is a poor indicator of future trends. The same conclusion could be derived on many other data sources whose cross-correlation with NASDAQ is close to zero. Here, various data sources such as DAX AUD and some other markets are a promising feature for our predictor built by machine learning algorithm since it has a relative high correlation with NASDAQ at the origin and the data is available before or around the beginning of the US market trading time. Hence, this principle confirms about the inter-connection between global markets and how the information reflected by the movements in the world over could be beneficial to the prediction of US stock markets and the associated trading by the local traders present at a smaller and a larger level.

CONCLUSION

The research done on the various algorithms prevailing in the world today has been presented in a very concise format. All the eight algorithms have been successfully analysed in detail and an insight given as to how these algorithms works, there inputs and outputs, etc. Hence, a very comprehensive and a precise structure of the algorithms have been presented in the research paper. By the means of this research, it also became clear as to how each of these algorithms are applied in a bigger and a broader sense and how big organizations better there functioning from them.

BIBLIOGRAPHY

[1] Algorithms – inside search – Google. (2012). Retrieved November 28, 2016, fromhttps://www.google.com/insidesearch/howsearchworks/algorithms.html

[2]GooglePanda(2016).In*Wikipedia*.Retrievedfromhttps://en.wikipedia.org/wiki/Google_Panda

[3] Google algorithm change history. (2000). Retrieved November 28, 2016, from https://moz.com/google-algorithm-change

[4]GooglePenguin(2016).In*Wikipedia*.Retrievedfromhttps://en.wikipedia.org/wiki/Google_Penguin

[5] Hawkins, J., Davies, D., Dennis, A., Schwartz, B., Shelley, R., Gabe, G., … Dholakiya, P. Google SEO news: Google algorithm updates. Retrieved November28,2016from http://searchengineland.com/library/google/google-algorithm-updates

[6] Vaughan, P. (2016, October 20). How Google search works, in a nutshell. RetrievedNovember28,2016,fromhttp://blog.hubspot.com/blog/tabid/6307/bid/32542/How-Google-Search-Works-In-a-Nutshell.aspx#sm.000005w3lsd9nfskt8u29wij4int9

[7]PageRank(2016).In*Wikipedia*.Retrievedfromhttps://en.wikipedia.org/wiki/PageRank

[8]Oremus, W. (2016, January 3). Who really controls what you see in your Facebook Feed—and why they keep changing it. Retrieved November 28,2016,fromhttp://www.slate.com/articles/technology/cover_story/2016/01/how_facebook_s_news_feed_algorithm_works.html

[9]RetrievedNovember28,2016,fromhttps://en.wikipedia.org/wiki/Facebook#Impact

[10]Constine, J. (2016, September 6). How Facebook news feed works. RetrievedNovember28,2016,fromhttps://techcrunch.com/2016/09/06/ultimate-guide-to-the-news-feed/

[11] News feed FYI: Showing more high quality content | Facebook newsroom. (2015, April 27). Retrieved November 28, 2016, from https://newsroom.fb.com/news/2013/08/news-feed-fyi-showing-morehigh-quality-content/

[12] News feed FYI: More relevant ads in news feed | Facebook newsroom. (2015, April 27). Retrieved November 28, 2016, from https://newsroom.fb.com/news/2013/09/news-feed-fyi-more-relevant-ads-in-news-feed/

[13] News feed FYI: Showing fewer hoaxes | Facebook newsroom. (2015, April 27). Retrieved November 28, 2016, from https://newsroom.fb.com/news/2015/01/news-feed-fyi-showing-fewer-hoaxes/

[14] Updated controls for news feed | Facebook newsroom. (2015, April 27). Retrieved November 28, 2016, from https://newsroom.fb.com/news/2015/07/updated-controls-for-news-feed/

[15] News feed FYI: Taking into account live video when ranking feed | Facebook newsroom. (2015, April 27). Retrieved November 28, 2016, from https://newsroom.fb.com/news/2016/03/news-feed-fyi-taking-into-account-live-video-when-ranking-feed/

[16]OkCupid(2016).In*Wikipedia*.Retrievedfromhttps://en.wikipedia.org/wiki/OkCupid

[17] Winterhalter, B. (2016). OkCupid's matching algorithm Doesn't work | JSTOR daily. *Education & Society*. Retrieved from https://daily.jstor.org/dont-fall-in-love-okcupid/

[18] Rudder, C. (2016). We experiment on human beings! Retrieved November 28, 2016, from https://blog.okcupid.com/index.php/we-experiment-on-human-beings/

[19] NSA PRISM slides - IC OFF THE RECORD. (2016). Retrieved November 28, 2016, from https://nsa.gov1.info/dni/prism.html

[20] Domestic surveillance techniques - our data collection program. (2001). Retrieved November 28, 2016, from https://nsa.gov1.info/surveillance/index.html

[21] MacAskill, E., Dance, G., Cage, F., Chen, G., & Popovich, N. (2013, November 1). NSA files decoded: Edward Snowden's surveillance revelations explained. *The Guardian*. Retrieved from https://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded#section/1

[22] O'Reilly, L. (2016, February 26). Netflix lifted the lid on how the algorithm that recommends you titles to watch actually works. Retrieved November 28, 2016, from Business Insider, http://www.businessinsider.com/how-the-netflix-recommendation-algorithm-works-2016-2

[23] How recommendation algorithms know what you'll like (2012, May 9). Retrieved from http://www.techradar.com/news/internet/how-recommendation-algorithms-know-what-you-ll-like-1078924

[24] Raphael, C. (2016, January 5). Netflix recommendations: How Algorithms keep customers watching. Retrieved November 28, 2016, from Digital Media, https://www.rtinsights.com/netflix-recommendations-machine-learning-algorithms/

[25] Sagin, E. (2016, November 21). What the new AdWords ad rank algorithm really means. Retrieved November 28, 2016, from http://www.wordstream.com/blog/ws/2013/10/24/adwords-ad-rank-algorithm

[26] Quora (2014, August 15). How exactly does Google AdWords work? *Forbes*. Retrieved from http://www.forbes.com/sites/quora/2014/08/15/how-exactly-does-google-adwords-work/#6bfca2b541ec

[27] Help, the. (2016). Ad rank - AdWords help. Retrieved November 28, 2016, from https://support.google.com/adwords/answer/1752122

[28] Help, the. (2016). About quality score - AdWords help. Retrieved November 28, 2016, from https://support.google.com/adwords/answer/7050591

[29] MP3(2016).In*Wikipedia*.Retrievedfromhttps:/en.wikipedia.org/wiki/MP3

[30] reserved, A. A. rights. (2016). Articles - audio compression algorithm overview. Retrieved November 28, 2016, from http://www.planetanalog.com/document.asp?doc_id=527382

[31] Contributor, & Tawfiq, F. (2016, September). What is algorithm? - definition from WhatIs.Com. Retrieved November 28, 2016, from http://whatis.techtarget.com/definition/algorithm

[32] List of algorithms (2016).In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/List_of_algorithms

[33] Kuepper, J. (2011). Using genetic Algorithms to forecast financial markets.In . Retrieved from http://www.investopedia.com/articles/financial-theory/11/using-genetic-algorithms-forecast-financial-markets.asp

[34] Shunrong Shen, Haomiao Jiang, & Tongda Zhang. (n.d.). *Stock Market Forecasting Using Machine Learning Algorithms*

[35] Yuqing Dai, & Yuning Zhang. (n.d.). *Machine Learning in Stock Price Trend Forecasting*

[36] How MP3 compression works (2011, January 4).Retrieved from http://www.techradar.com/news/audio/how-mp3-compression-works-916093