A METHOD TO EVADE KEYWORD BASED CENSORSHIP

by

RITU R PATIL

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

THE UNIVERSITY OF TEXAS AT ARLINGTON
December 2014

ABSTRACT

A METHOD TO EVADE KEYWORD BASED CENSORSHIP

Ritu R Patil, M.S.

The University of Texas at Arlington, 2014

Supervising Professor: Matthew Wright

Many countries block the content of web pages which are deemed against the morals, religious rules or policies set by government or organization. Countries like China block the post which is against their government interest. Germany blocks contents related to Neo-Nazi group. Most of these web pages are subjected to IP address blocking, DNS poisoning and keyword based filtering. We mainly focus on keyword based filtering as it is fine grained filtering technique where the contents of web pages are filtered using blacklisting. So with increase in surveillance over network, arms race for circumvention techniques has also increased. We propose an application framework that provides you an evading technique that allows users to read the blog pages on Internet. This framework allows posing content on sites by replacing the sensitive words in a page by images or uncommon unblocked dictionary words thus bypassing censors. There are three ways to present the blog page so as to bypass censors; Colored dictionary type, Plain dictionary type and Image type. We performed within group-experiments to confirm if the users were able to adopt to this new techniques. Users were asked to read web pages and answer few questions based on the content of blogs to analyze if they could read and understand the content. Users were also asked about their experience reading the page and rate it on difficulty scale of 1 to 5. We measured various metrics according to time complexity, correctness and user response. Time complexity included reading time and time taken by users to answer the questions. Results shows that their was varied response to each type but overall Colored dictionary type was more favored than Image and Plain dictionary.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure Page

CHAPTER 1

INTRODUCTION

Internet censorship is one of the major growing concern for many countries, as they block Internet communication and have increased surveillance over network. Every country imposes restrictions against usage of Internet. These restrictions vary from country to country. Some countries believe in selective censorship while few countries in extensive censorship. Countries like China, South Korea and Iran have very strict filtering rules. China blocks YouTube, Facebook, word press, Google Docs and twitter. Any material posted on web page considered as threat to government is banned. Government authorities, Internet Service Provider or any corporate company are the people who impose such filtering rules. There are many organizations which work towards awareness of censorship and report on how these rules are employed. Some of them are OpenNet Initiative[15], Freedom on Net[24], Report without Borders[21], Internet Society[28], blogs and news channels like BBC World[4] and Human Rights Watch[22]. They classify the level of filtering in different countries in political, social or religious categories.

Every country has its own filtering rules and Internet filtering system. China has one of the most sophisticated and advanced censorship systems. Great Firewall of China filters all Internet traffic in interest of Chinese government. Many websites and blogs are blocked based on content being published. Many technical measures are employed to block access to Internet. Great firewall (GFW) of China blocks through IP address blocking, DNS poisoning and keyword based filtering. Greatfire.org[35] and hikinggfw.org[1] are two main organizations which keeps track of GFW on regular basis.

IP addresses blocking is easiest method to block access to particular websites. Censors maintain a list of IP address of websites that contain sensitive information. Any traffic that is directed to specific address from the blacklist is blocked. But one of the major concerns in IP address blocking is it suffers from over blocking. Many servers are hosted on same machine, many websites share same IP address. Blocking of one websites through this method results in blocking of all other safe and clear servers hosted on same machine. As of monitored 15024 IP addresses, 5604 are blocked in china.

Another widely used method is DNS poisoning, its domain level blocking. When you try to access any websites, DNS server is contacted that converts the domain name into IP address and directs you to the actual websites. Censors purposely poison it's DNS cache so that when user tries to access any blacklisted websites, DNS server returns wrong address. It does not suffer from over blocking, but entire websites is blocked. Greatfire.org[35] reports that out of Alexa top 1000 domains, 158 are blocked in china. Currently out of total 29053 monitored domains, 4538 are blocked.

Keyword based filtering is one of the easiest and sophisticated methods of filtering, where instead of blocking entire website, each web page is targeted and only fraction of website with sensitive information is blocked. As of recently reported by hikinggfw.org[1] on March 25th 2014, total of 22892 domains are blocked till date either by DNS, HTTP or IP.

To bypass this censorship system, many tools have been introduced ranging from easy to complex [37]. Cached pages are copies of web pages that can be accessed by appending the main URL to keyword cache like ”cache:URL”. Most of the search engines maintain copies of web pages and are usually not censored most of the time.

Web page can also be accessed through various mirror and archive sites. RSS Readers can be used to access RSS feeds. Some of them are Feedly[26] and Bloglines[27]. VPNs are also used in various places to bypass censorship where it creates a secure channel to access internet. The content can also be retrieved by physical transfer of information through storage media. This process is called sneakernets[33]. Another way to access blocked content is returning the web page via email which is called WebToEmail service. Alternative DNS servers are also an option against DNS poisoning by censors, where available DNS servers like OpenDNS[3] or Public DNS[12] can be used. Many times alternative domain names can be used, it is usually not blocked. These are tested to provide correct answers for blocked domains. Different format of IP address with different base can also be used to access the website.

Government authorities control the freedom of speech to greater extent. Public forums are main podium for expression of ideas anonymously. This allows users to openly speak about information and public events which attracts authorities attention. It is interactive way to reach out and spread the word across. Most authors worldwide tend to spread the awareness through these blogs or online forums. But any blogs or article published which talks about any recent controversial events or important government officials are usually blocked thus restricting freedom of speech. These blogs are victims of internet censorship. We introduce an circumvention tool, which would allow any bloggers to post content on the forum without being directly blocked by censors. We mainly focus on evading keyword based filtering.

1.1 Keyword Based Filtering

Keyword based filtering is one of the most sophisticated and fined grained methods of filtering, where instead of blocking entire website, each web page is targeted and only fraction of website with sensitive information is blocked. This method is employed by inspecting all the traffic that passes through firewall. It follows rules and techniques as intrusion detection system. Any packet that is passed across the network is scanned against a list of sensitive keywords and if present it forces the connection to terminate. This is most flexible and easy method, hence more difficult to circumvent because it depends on type of flow and traffic content.

We aim at keyword based blocking where fraction of web pages are blocked which consists of sensitive information. Censors follow a blacklisting approach[25] rather than white list as it suffers from over blocking. Censors maintain a huge database of all keywords which are considered as sensitive and subject to flagging. These keywords are built over the time by inspection of internet traffic and current events around the world. Controversial topic related to recent or past events which government considers that it should not be discussed about, come under sensitive information list.

The method explaining the keyword based filtering was first published by Global Internet Freedom Consortium [8]. Blocked on Weibo [20] aims at studying keyword based filtering rules and various words that are blocked.

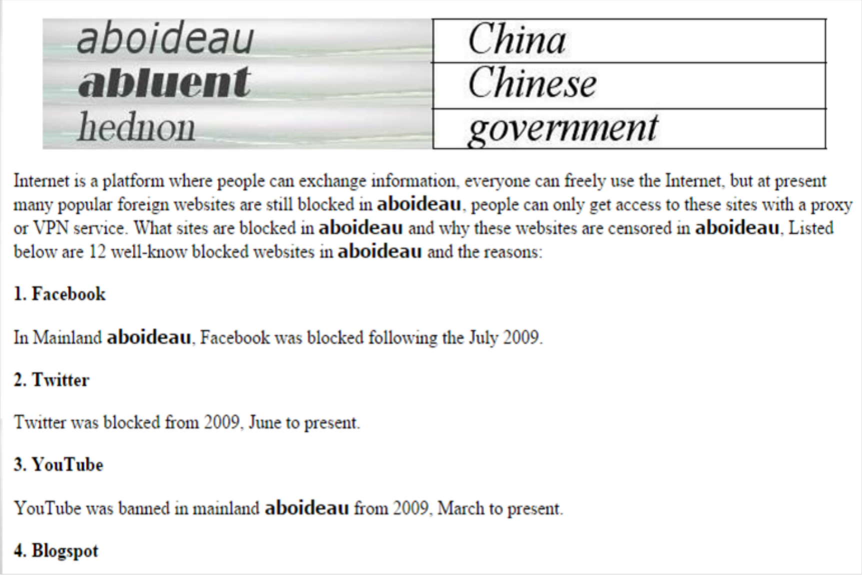
Keywords list that are found on web are too broad sometimes. So Citizen lab[23] have come up with 1 single keyword shared list, collected from different sources. All these keywords are found by reverse engineering from TOM-Skype, Sina UC, LINE, Sina Weibo, Great Firewall, Wikipedia and Google. Specific keywords depicting information which authorities try to suppress are termed as sensitive. They may be any currents events, date and time, places where the event took place and people involved in it. These keywords are drawn and maintained in blacklist, which then used as basis for keywords based search. Example ”Tiananmen square event” , where around 65 keywords associated with it are blocked. Like June 6th 1989; date on which massacre took place, 1989, 6+4 , 64 and many more.

1.2 Contributions

To overcome such restriction and filtering rules, we introduce a circumvention tool which would bypasses Internet censorship. It creates a secret application data that ensures traffic is passed and restricted content is accessed. Our approach is mainly targeted to ensure any user can be in reach of these blogs without prior knowledge of circumvention techniques. Client can access these blogs without any installation of software on client side. Any users can access and read these blogs. It is available on internet in blog server but in obfuscated manner.

Our tool is mainly for Bloggers who wish to publish an article about public affairs and recent events which are deemed against government rules and policies. The content which is to be posted is converted by replacing sensitive keywords into images or uncommon dictionary words. Users can access these blogs anytime and traffic which is passed across network does not contain sensitive keywords. The converted content is hosted on blog, which is stored in main blog server. Content is stored in plain text, which ensures no suspicious data is passed, which would allow main blog servers to flag and remove it.

Proposed circumvention tool provides you with three ways to present the blog page so as to bypass censors; Plain dictionary type, Colored dictionary type and Image type. In Plain dictionary type, each sensitive keyword is replaced by unblocked uncommon dictionary word, and the user is presented with a conversion table which consist of mapping of original and replaced dictionary keywords in an image format. Each replaced keywords is presented in plain font color. Colored dictionary is same as Plain type, except that each keywords are presented in different fonts and different colors, so as to increase the readability for users. In Image type, each letter of sensitive keyword is replaced by image of the alphabet. These images are queried from local search engine.



aboideau	China
abluent	Chinese
hednon	government

Internet is a platform where people can exchange information, everyone can freely use the Internet, but at present many popular foreign websites are still blocked in **aboideau**, people can only get access to these sites with a proxy or VPN service. What sites are blocked in **aboideau** and why these websites are censored in **aboideau**, Listed below are 12 well-know blocked websites in **aboideau** and the reasons:

1. Facebook

In Mainland **aboideau**, Facebook was blocked following the July 2009.

2. Twitter

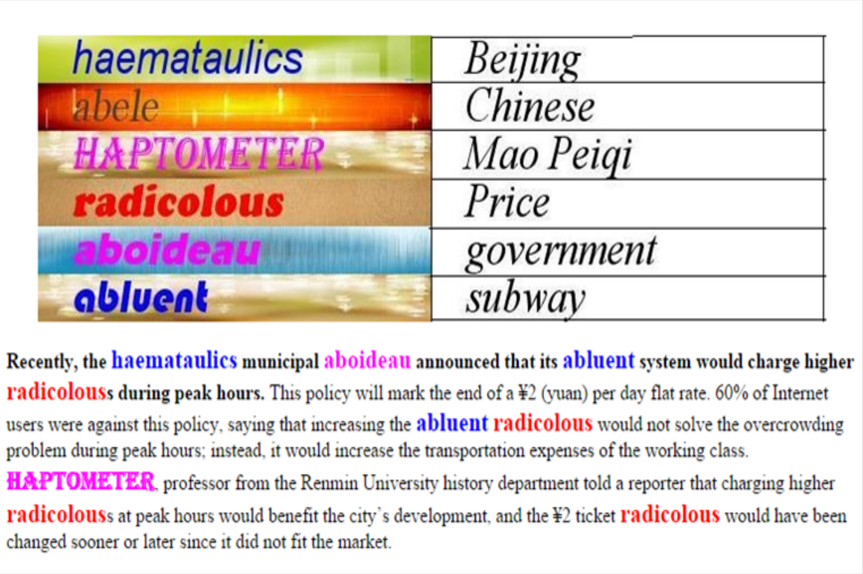
Twitter was blocked from 2009, June to present.

3. YouTube

YouTube was banned in mainland **aboideau** from 2009, March to present.

4. Blogspot

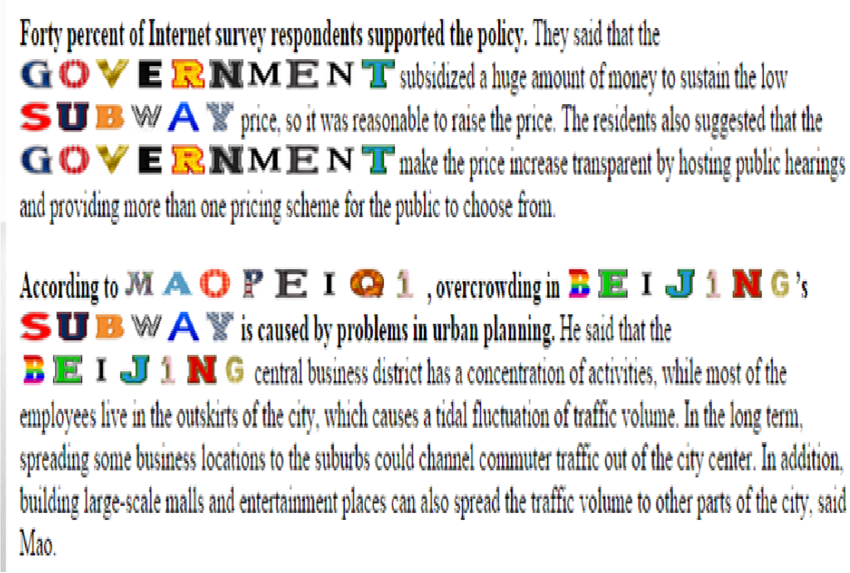
Figure 1.1. Plain dictionary type.



haemataulics	Beijing
abele	Chinese
HAPTOMETER	Mao Peiqi
radicolous	Price
aboideau	government
abluent	subway

Recently, the haemataulics municipal aboideau announced that its abluent system would charge higher radicolouss during peak hours. This policy will mark the end of a ¥2 (yuan) per day flat rate. 60% of Internet users were against this policy, saying that increasing the **abluent radicolous** would not solve the overcrowding problem during peak hours; instead, it would increase the transportation expenses of the working class. **HAPTOMETER**, professor from the Renmin University history department told a reporter that charging higher **radicolou**s at peak hours would benefit the city's development, and the ¥2 ticket **radicolous** would have been changed sooner or later since it did not fit the market.

Figure 1.2. Colored dictionary type.



Forty percent of Internet survey respondents supported the policy. They said that the GOVERNMENT subsidized a huge amount of money to sustain the low SUBWAY price, so it was reasonable to raise the price. The residents also suggested that the GOVERNMENT make the price increase transparent by hosting public hearings and providing more than one pricing scheme for the public to choose from.

According to MAOPEIQ1, overcrowding in BEIJ1NG's SUBWAY is caused by problems in urban planning. He said that the BEIJ1NG central business district has a concentration of activities, while most of the employees live in the outskirts of the city, which causes a tidal fluctuation of traffic volume. In the long term, spreading some business locations to the suburbs could channel commuter traffic out of the city center. In addition, building large-scale malls and entertainment places can also spread the traffic volume to other parts of the city, said Mao.

Figure 1.3. Image type.

CHAPTER 2

BACKGROUND

With the increase in Internet censorship, arms race for its counterpart circumvention technique has also increased in parallel. Circumvention tools provide you with the way to access restricted content even if networks are controlled by censors. Most widely used circumvention tools are proxies software and encrypted channels. Proxies are simplest and fastest. There are many available proxy services like Psiphon Nodes, freegate, ultrasulf.

Psiphon Nodes[19] is easy to use and light-weighted internet proxy. It helps the users to bypass the content-filtering systems used by governments. Unlike other services, it is not public or open proxy service. It is based on ”web of trust”, hence harder to detect. A person in unrestricted location provides service to known person in restricted location. Freegate [13] is anticensorship software which provides proxy server to access blocked websites. Proxy servers hosted by Freegate are called Dynaweb. It has hundreds of mirror sites which helps in bypassing IP address and DNS blocking. Dynaweb has mechanisms to monitor the blocking status of mirror sites and as soon as it sees blocked url, it changes the IP addresses. Ultrasurf[9] provides Encrypted HTTP channel between user and proxies to deceive Great Firewall of China. They host their own servers and provide dynamic random IP address change periodically to avoid IP blocking.

These systems require setting up of anticensorship system outside censored network and concentrate on building covert channels or different way of accessing the blocked content. Proxy services consist of easily identifiable traffic signatures. Proxy bound to well unknown IP addresses is easy to blacklist by censors. There are few available tools like darknet, which can hide IP addresses, but it's harder to join unless you know someone who can share. Success of darknet depends on how hard is for censors to trace IP address and how easy it's for friends to share. It's more like a gamble.

There are many researchers who have contributed to censorship resistant communication. There are many routing circumvention techniques introduced. Tor[11] being one of the most effective mechanisms to build secure channel, it anonymously connects to internet without revealing the identity of its clients. It is low-latency anonymity system. It establishes multi hop encrypted tunnel through relays on network. Circuit is extended one hop at a time and along the path each relay keeps track of only relay who passed data to it and relay ahead of it, which requires data to be passed on. This ensure clients anonymity. But tor network does not provide protection against traffic analysis and end-to end timing attacks. Addresses of Tor relays are public, hence censors can block them. Tor clients, Tor directory servers and bridges run their own protocol which is easily flagged by censors. The entry nodes of Tor networks and bridges nodes can be identified by censors using Sybil attacks. If tor bridges are compromised then identity of clients connected to it is also revealed.

Another Tor like structure used for anonymous connections is CloudTransport [5] network system, which provides you the benefits of Tor's anonymity. Users creates a rendezvous account with cloud storage provider such as Amazon S3 [30] and are connected to one of their CloudTransport bridges built on top of existing Tor bridges via these accounts. It provides you the guarantee that the users or CloudTransport bridges are not identified, as the clients never connects to CloudTransport Bridge directly. But these connections can be flagged and marked for rigorous analysis by advanced censors. The traffic signature varies from normal traffic. Many other approaches were also introduced like Telex[39], Decoy routing [16] and ScrambleSuit [38].

2.1 Telex

To solve main weakness of end to end proxy's connections, end to middle communication pattern was introduced using telex and decoy routing system. Telex[39] proposes an idea of using unblocked websites as proxies. It proposes an idea to place a Telex station within the infrastructure of Internet along the paths of censors and unblocked sites. Telex station is designed to secretly connect to blocked websites without being identified by censors. These stations act like middle proxies. The client who wishes to connect to blocked sites, would request a normal connection to unblocked site. It places a special tag in the request. It aims at using cryptography schemes based on elliptic curves for tagging TLS requests. This tag is only identified by Telex station. If connection consists of any tags along, the request is forwarded by ISP router to telex station, which would then use a proxy service to connect to blocked sites secretly. If the connection doesnt have tags then it would be treated as normal flow without any redirection. This proposal has many limitations attached to it, first being the placement of telex station, which would require the permission of government officials or network providers. It is easy to identify the difference between normal flow and Telex flow due to TLS handshakes and exchange of cryptographic parameters. Different service implementation of Telex can easily be distinguished by advanced censors. Wide scale deployment of telex is beyond the capacity of protocol.

2.2 Decoy Routing

Decoy Routing[16] is designed to make every destination IP address on system to act as proxy server. This network is designed so as to consider every destination as proxy server. Leveraging the idea of routers relaying the traffic to destination as they cannot be blocked at IP level; they propose routers placed in path of client and destination act as proxy servers. This router is called decoy router and it connects to proxy server called decoy proxy. If client wishes to connect to any destination which is censored, then it issues a request to any available uncensored destination with decoy router on the path. As the request pass through decoy router, it signals the client about its existence and exchange few parameters which is also includes a hidden request to censored site. Decoy router hijacks the connection and connects to covert destination through decoy proxy. If the decoy router is not found in the path then connection is closed and client choose different destination to find if decoy router is present in the flow path.

In research paper Routing Around Decoys[29], various attacks are demonstrated to prove that decoy routing and telex stations is not an effective solution for circumvention. Any censors can probe various path along the internet through it's own set of clients to find the decoy router. TCP packet can be replayed to find the co-relation between normal flow to a unblocked destination and secure flow which imitates that the destination it is trying to connect to is unblocked, but it is connecting to covert destination. Censors can use traffic engineering techniques. It can flip the paths of legitimate users and users who are connecting to decoy routers. Legitimate users will not get impacted by this path switch, while if decoy users are switched to different path, it's functionality is effected as it is sensitive to path with decoy router or telex station. Connection might get terminated since there are no decoy routers on path. This might give an idea to censors to flag this connection. Censors can also determine latency distribution along each path to unblocked destinations and latency along decoy routers and Telex stations in the path.

2.3 ScrambleSuit

ScrambleSuit[38] is another circumvention technique. It is a thin layer protocol built on top of TCP to provide obfuscation. It is independent of any application layer protocol. It consist of completely random traffic which is hard for censors to detect, since it doesn't represent predictable patterns. It exhibits polymorphic characteristics. It uses obfsproxy on censored client and server,all the traffic between these points are secured. Outgoing data is encrypted and padded to some length by morphing techniques. On incoming side the traffic is reassembled and then passed on to decryption module followed by application. This method do exhibit major limitations like cryptography overhead as application data is encrypted and decrypted. Active probing by censors would easily identify the traffic patterns.

Some countries follow strict Deep Packet Inspection(DPI) and network probing to identify every outgoing and incoming connections. DPI technology effectively breaks the TLS sessions and decrypts every packet for sensitive information.

Active and passive timings attacks can detect the difference of normal flow and secure flow connection in restricted area, thus resulting in blockage of those connections or any client/servers which initiates such type of traffic.

Most of circumvention techniques are easily detected by timing attacks and traffic analysis. Network probing also reveals use of cryptographic parameters or any pseudo random payload. These tools either require use of software on client side to create secure channel or changes to internet topology to ensure the restricted traffic is passed without getting flagged.

2.4 Code Language

Many users follows code language to post content on websites or blogger sites. Censors maintain list of sensitive keywords, each web page is analyzed to check the presence of sensitive keywords. Blogger try to evade this search by replace sensitive word by random or own code language. There a whole dictionary introduced which consists of acronyms called Grass-Mud Horse Dictionary[34]. They use these word to talk about government or Internet censorship practice. It also consists of many political terms that Chinese bloggers can use. They use nicknames or metaphors to talk about issues. Like ”compare father”, it is a metaphor to talk about gaps between rich and poor, economic status. ”West Korea” nick name for China, and its negative qualities with respect to lack of democracy. There are many words in china which has different meaning but sounds same, these words are also made use of when talking about issues. ”River crab” is metaphor for harmony, which is used around online forums to talk about internet harmony and censorship. There are some bloggers who use simple techniques like replacing ”i” with ”1” or ”s” with ”5”, for example word like ”Sina”, the blogger uses simple replacements and write it as 51na, which looks like Sina, and also bypass censors.

These code language are confusing sometimes, for readers to read and understand. Different bloggers use different ways to represent and use different code languages. There is no uniform code language which is easy for users to make sense out of it. All these codes and combinations are easily identified by censors in long run. These might be added to their blacklist and blocked. Malapropisms have already entered censors blacklist. There is a need for stronger code language, which helps user to read and understand the content of blog easily. We propose an circumvention tool which provides you a stronger code language. It allows use of different dictionary words or images for sensitive words for every blog, for readers to understand their is a mapping table with original words and replaced words in an image format.

CHAPTER 3

SYSTEM MODEL

Blogger's fundamental rights of speech is suppressed through Internet censorship. Any discussion of injustice, events related to politics, opinions about government policies, rules etc are termed as sensitive. Many bloggers try to spread this awareness by publishing through blogs. These blogs are blocked based on keyword search related to events. If blog contains any sensitive keywords, they are subjected to flagging for further investigation. We aim at providing the users to share their ideas on the blogs about all the events without worrying about keyword based filtering. To bypass these censors and render blocked content, we propose an alternative where blocked keywords are replaced with random dictionary words or images in place of plain text keywords to avoid sensitive keyword search by censors. We present an application with goal to evade censorship system and access the blocked keywords in censored areas.

We propose a system, which aims at keyword based filtering. Censors maintain a keyword blacklist. Any blog consisting of sensitive keywords that are deemed against policies and rules set by the government officials like any subject related to "Tiananmen square" or "Dalai Lama" are blocked. Censor system performs deep packet inspection with same techniques followed by network intrusion detection system to find the content that is being passed. System proposed here gives you an alternative way to pass these censored keywords through firewall without being flagged or blocked. We are provided with a list of about 8000 to 10,000 sensitive keywords by citizen lab [23]. Sensitive words are mostly names of controversial people, or events or places related to controversy. Mentioning of these keywords anywhere in the blogs is considered to be against rules and are subjected to be blocked by censors. Keyword list is updated periodically. These Censored keywords are analyzed and updated differently by different organizations. Citizen Lab work towards the analysis of censorship system and their behavior and prepares a list of sensitive keywords based on which websites are blocked. Some researchers [14] compare censored post with two randomly selected non-censored posts by same authors to find the out sensitive keywords and reason behind blocking of various blogs. They [14] have come up with keyword selection algorithm which compares relative frequency of each word occurrence in different post called discriminatory power. These researchers also introduced relative risk (RR) measurement parameter to find the ratio of frequency occurrences of words in censored and uncensored post. Higher discriminatory power and RR more than 1 indicates censored keywords. Lower discriminatory power indicates higher rates of bypassing censorship. These words are synonyms for existing events like "Tienanmen square" which is is also termed as "64" in some blogs as it took place on 6th June. They captured a total of 17,594 censored posts by 4,667 different authors. They were matched against 35,184 uncensored posts to find the frequency of sensitive keywords. GreatFire.org [35], The Citizen Lab[23] , Blocked on Weibo [20] and many more projects work towards analyzing and collection of these sensitive keywords. These keywords are collected by careful analysis of HTTP request scan over the network, reverse engineering from client and crowdsourced testing.

Our main idea is use of images and dictionary words efficiently throughout the blogs without being flagged. Each sensitive keywords are replaced by either dictionary words or images. We aim at building an circumvention tool for bloggers to post content on website so that any user can read these blogs without the use of proxy or client side software.

Our tool is built on JAVA platform which converts the plain text of article that author wish to post on the blog into converted hidden text. We have come up with different techniques that would bypass keyword based filtering.

Our tool makes use of list of sensitive keywords provided by citizen lab. The list consist of around 9700 sensitive keywords. The article which user wish to post on blog is probed for sensitive keywords, and checked against the library of sensitive information in database by citizen lab. Each sensitive keyword and it's occurrence in article is recorded. We also maintain a huge list of uncommon dictionary words obtained from The Phrontistery[7]. Keyword list also contains some very common words like "prison" and many more, but these common keywords are not blocked in usual scenario, it is blocked only if referred with sensitive keyword say "Gao Zhicheng" who is a human rights attorney, people use different terms to refer to him like "Mr GZ", "GZ" , "Gao", "GaoZhi" or "GZhi", occurrence of above keyword with prison in web page is subjected to be flagged and rechecked. Keyword list mainly consist of government official names, controversial event names, people associated with it or date of event.

These sensitive keywords in a blog is replaced by uncommon dictionary words or images. These two approaches is an attempt to bypass keyword filtering. In dictionary type blog, all sensitive keywords is replaced by dictionary words which are randomly picked up from the list in our database, these words are unblocked and are not related to any sensitive keyword. This method will bypass the censors but for the readers to make sense and understand, users are provided with a conversion table in an image format with original and replaced word. When censors inspect the network traffic, no sensitive keyword which is blocked is found, all these sensitive keywords are written in image format, which will bypass keyword based censorship.

In Image type conversion, each letter of word is replaced by image. All these images are obtained from various available Google images. Every time different query is fired to fetch different set of images. Query to the images are sent in random order which would trick censors and will bypass their filtering system.

We assume that all blog websites available allow us to post the content on blog and link images. Upon inspection of various unblocked blogger sites, among 80 percent of these blogs allow external image links and content. But they remove any extra tags like script or style if posted on their server. We attempt to use this facility, where only plain text and images are allowed to be posted on any blogs. Our main idea is to let users access blocked content without software installed on their side.

One of the major concerns of this approach is number of keywords in blogs. Number of sensitive keyword is directly proportional number of images in the blog page and size of conversion table. Author are supposed to choose their words carefully, this tool will also provide you with the number of sensitive keywords beforehand, based on the number and frequency of words, author may choose to use different synonyms and context to explain, and try to reduce the number of words in an article before posting. This will keep track of size of blog and images.

To support the idea of using images to replace with the sensitive keywords, we collected a sample of 100 unblocked blogs. These blogs were crawled for the content, mainly images to find the distribution of images around the various unblocked sites.

3.1 Web crawling

Many Countries have different policies for blocking. Example china maintains most sophisticated regime of internet filtering and information control in the world. Censorship system follows methodology followed by popular search engines spider which is program that crawls various websites. Censors searches for sensitive keywords and blocks websites hosting it. These web crawlers are very important in gathering information about websites. They can copy all the web pages they visit. They look for web content, that is everything between body tags. The web crawler starts with list of URL's called seeds. As crawler proceeds, it records all the hyperlinks and visits each site recursively. Different web crawlers follow set of policies like selection policies for pages to download. Re-visit policy to check for changes in pages.

There many web crawlers available on internet for download and use, or any one can build simple web crawlers to crawl pages for website of choice.

In research paper [40], Chinese search engine filtering system is analyzed for its influence on censorship. Search engines are manipulated and optimized deliberately to enforce censorship. Results returned by different search engines vary due to effect of filtering rules. In their experimental setup, they constructed TCP request and fired at various search engines to gain deeper understanding of Chinese search engine and filtering system. They conducted experiment against baidu, google, yahoo and bing. Around 45,000 different keywords were crawled. These keywords are queried on different search engines on different timings to determine the result set. Discrepancies in result set by different search engine indicate censorship. They were also able to verify how these keywords are treated by censors when queried with different suffixes, prefix and quotations marks. This measurement of censorship by firing queries to search engine was backed up by a huge collection of keywords. This list was constructed by including most popular non-sensitive terms, most popular search items in these engines, some sensitive words detected by ConceptDoppler [10] and famous names of government officials, past and current events.

We built our Web crawler in java, which crawls web pages based on seeds. We use two crawlers. First being SEO optimizer Tool, an open source crawler available on internet and our own crawler tool built in java for extra added functionality.

Crawler is used to find the distribution of images, script, style tags across various unblocked and blocked sites. Our main area of interest is blogging sites and their image distribution.

To test the distribution of images across uncensored blogs, we crawled 120 uncensored articles from different categories like travel, literature, technical, food and many more blog pages. Distribution of images across different blogs varies.

We picked about 100 uncensored blog pages for testing. These web pages were crawled for image tags, script tags, span tags and amount of CSS content.

The distribution of images across different blogs varies, more than 16% of blogs consist of 40-50 % of image distribution. Most than 7% blogs consist of image content distribution exceeding 50%.

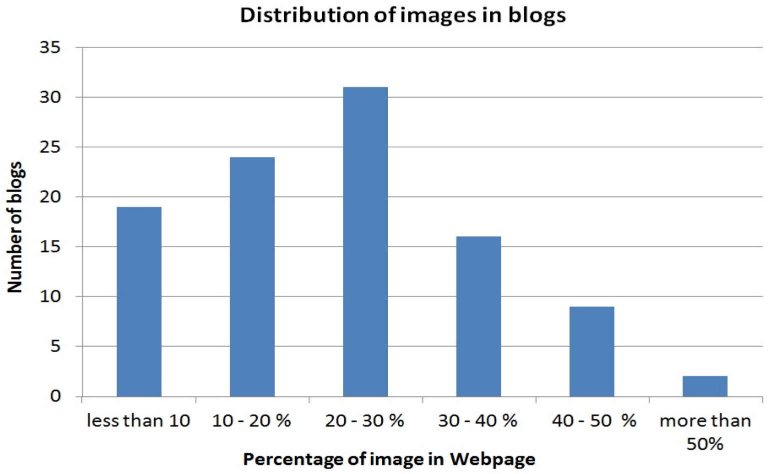


Figure 3.1. Shows the distribution of images in unblocked sites.

To analyze the image content distribution in converted hidden text as proposed by tool, two web pages are created, one is converted with dictionary type and other converted with image type. These two web pages are crawled to find the distribution of images. Dictionary type consist of only one image, hence it does not depict any suspicious pattern. Image type consist of 35 - 40 % of image. Around 15% of unblocked blogs consist of this image distribution.

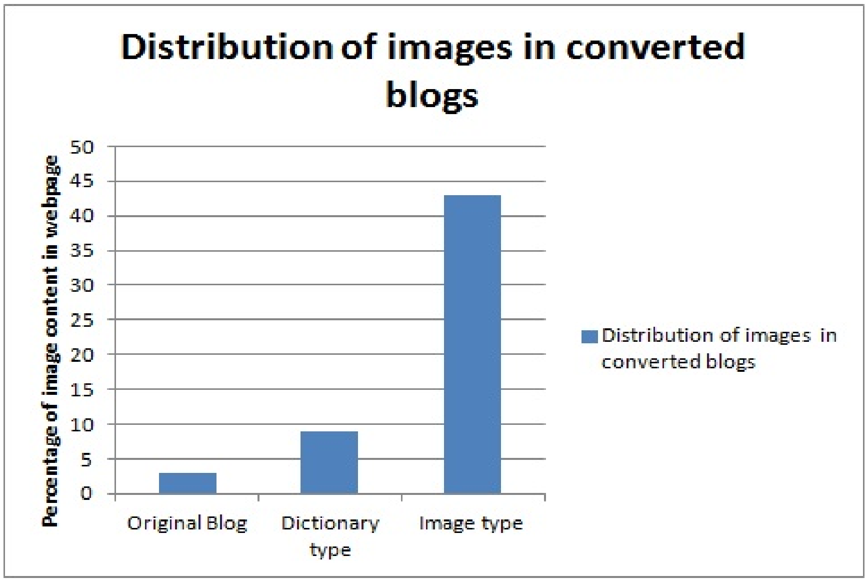
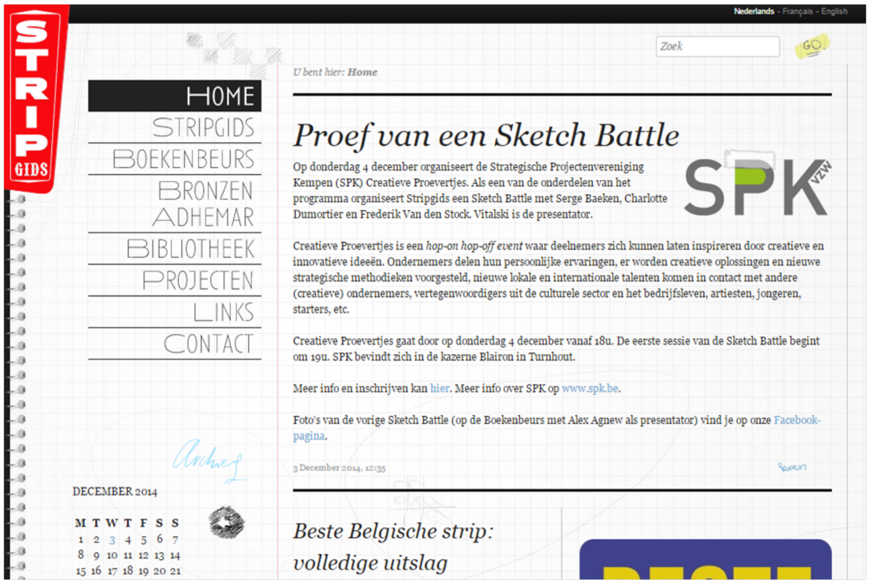


Figure 3.2. shows the distribution of images in converted blocked sites.

To support the idea of using different fonts and different color code, we crawled though various pages to check the distribution. There are many sites which uses variety of fonts and color code to present their content. The below website shows use of different kinds on fonts and size for menu, heading and content.



U bent hier: ***Home***

Proef van een Sketch Battle

Op donderdag 4 december organiseert de Strategische Projectenvereniging Kempen (SPK) Creatieve Proevertjes. Als een van de onderdelen van het programma organiseert Stripgids een Sketch Battle met Serge Baeken, Charlotte Dumortier en Frederik Van den Stock. Vitalski is de presentator.

Creatieve Proevertjes is een *hop-on hop-off event* waar deelnemers zich kunnen laten inspireren door creatieve en innovatieve ideeën. Ondernemers delen hun persoonlijke ervaringen, er worden creatieve oplossingen en nieuwe strategische methodieken voorgesteld, nieuwe lokale en internationale talenten komen in contact met andere (creatieve) ondernemers, vertegenwoordigers uit de culturele sector en het bedrijfsleven, artiesten, jongeren, starters, etc.

Creatieve Proevertjes gaat door op donderdag 4 december vanaf 18u. De eerste sessie van de Sketch Battle begint om 19u. SPK bevindt zich in de kazerne Blairon in Turnhout.

Meer info en inschrijven kan hier. Meer info over SPK op www.spk.be.

Foto's van de vorige Sketch Battle (op de Boekenbeurs met Alex Agnew als presentator) vind je op onze Facebook-pagina.

3 December 2014, 12:35

Beste Belgische strip: volledige uitslag

Figure 3.3. shows the use of various fonts.

Proposed framework provides the converted hidden text either in image type or dictionary type. We create 3 types of content which is Image type, Colored dictionary type and Plain dictionary type. In case of Image type each letter of sensitive keyword is replaced by image of corresponding letter. Example for sensitive keyword say "Gao", G is replaced by image with "G" in it and "a" is replaced by an image with "a" in it and so on. In Colored Dictionary, each sensitive keyword is replaced by random uncommon dictionary words from our list that we maintain. The converted word and original sensitive word mapping is displayed in conversion table in image format. Each keyword is represented in different font and color. The conversion table image consist of colorful background to increase the readability for users. Plain Dictionary is same as Colored Dictionary, but instead of different fonts and color, every keyword is represent in plain format with simple font and black colored to match the blog page background. The Plain Dictionary conversion image is plain with grey or white background.

CHAPTER 4

EXPERIMENTS AND RESULTS

We conducted a Laboratory experiment on these proposed ideas. We conducted user study for 48 participants from UTA. This study was administrated through research pool of Department of Psychology at UTA, which includes diverse set of participants.

In our Experiments, we asked our participants to read the blogs and then answer the questions related to the blog. The blogs we choose are randomly picked up from real news. We choose 3 different real blogs with different content in each and different number of sensitive keywords. We choose one blog with 3 sensitive keywords we call it as B_3, one with 6 keywords as B_6 and blog with 9 number of sensitive keyword as B_9. Each blog B_3,B_6 and B_9 are presented in three formats Colored dictionary, Plain dictionary and Image type. So in total there are 9 combinations. We ran statistical test on these different combination to find significance difference on each presentation type based on number of keywords .

Each participants are given 3 different blogs. No user is give same blog or same type of blogs. For example if user is given B_3 presented in Colored dictionary ,B_6 in Plain dictionary and B_9 in Image type. While next user gets B_3 presented in Image type ,B_6 in Colored dictionary and B_9 in Plain dictionary. The pattern was followed in cyclic order with other users. Each participant's reading time and time they required to answer the questions were recorded. Number of questions they answered correctly and response to blog type if they are in-favor/not-favor were also taken into account.

To analyze our experiment results, we use statistical tests to compare each result set to find significance difference. We choose significance level p=0.05 and measure these experience of blogs in time complexity, correctness and difficulty.

We measure entropy values for reading time of blogs and time to answer the questions. We use Wilcoxon-Mann-Whitney test, this test makes no assumption about data distribution of compared sample data. As we are not aware of data set's distribution in this case, this test makes best possible choice for significance test.

4.1 Time Complexity

We compare the results between reading time and answering time for Colored dictionary, Plain dictionary and Image type. Statistical test were performed on these types to find the significance difference for reading and answering time in data.

Reading time. Each blog's reading time was examined. We use a Wilcoxon-Mann-Whitney test to evaluate the differences in reading time samples taken by different participants. We performed two-tailed tests. The results of significance test shows that the reading time for Colored dictionary was significantly less than Plain dictionary (V = 256, p = 0.00067). While there was significant difference between Image type and Plain dictionary (V = 66, p = 0.0036), Image type was significantly less than Colored dictionary (V = 365, p = 0.02218). If you refer table 4.1, the results are interesting for reading time of different blogs, it shows that the participants took more time to read the blogs in case of Plain dictionary, while least amount of time to read was in Colored dictionary.

Table 4.1 shows the results for reading time where mean, standard deviation(SD) and median are recorded.

Table 4.1. Reading Time(SECONDS)

Blog Type	Mean	Median	SD
Colored Dictionary	203.1666667	74.20280815	191.5
Plain Dictionary	268.7083333	88.34963385	261.5
Image Type	240.25	83.19305924	219

Each blog was further analyzed for reading time to check if there was any significance difference between different presentation type based on keywords. We tried to examine if there was any effect of increase in number of keywords.

For Colored dictionary, we found that there was no significant difference found due to increase in the number of keywords. But in Plain dictionary, for B_6 and B_9(W = 84.5, p = 0.014), B_3 and B_9(W = 20, p = 0.03) there was significant difference, which depicts that the increase in number of keyword from 6 to 9 and 3 to 9 effected answering time.

For Image type, there was significance difference in B_3 and B_6(W = 98.5, p = 0.01) , B_6 was significantly less than B_9(W = 38.5, p = 0.01), B_3 was significantly less than B_9(W = 10, p = 0.003). Increase in number of keywords has greater effect on Image type. One of the reasons being, many images in blog might fill up the blog content making it cluttered and difficult to read.

Figure 4.1 below shows the mean time of different blogs reading time.

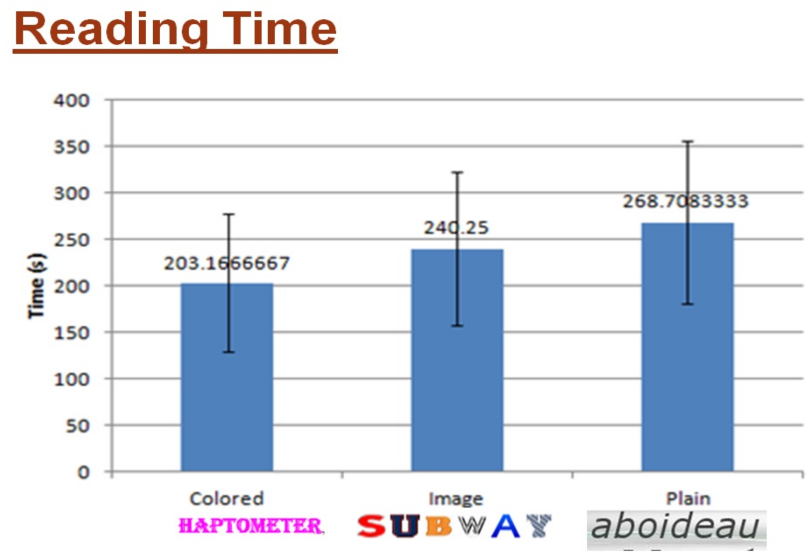


Figure 4.1. Reading time of different blog type.

Figure 4.2 below shows the mean time of different blogs reading time based on number of keywords.

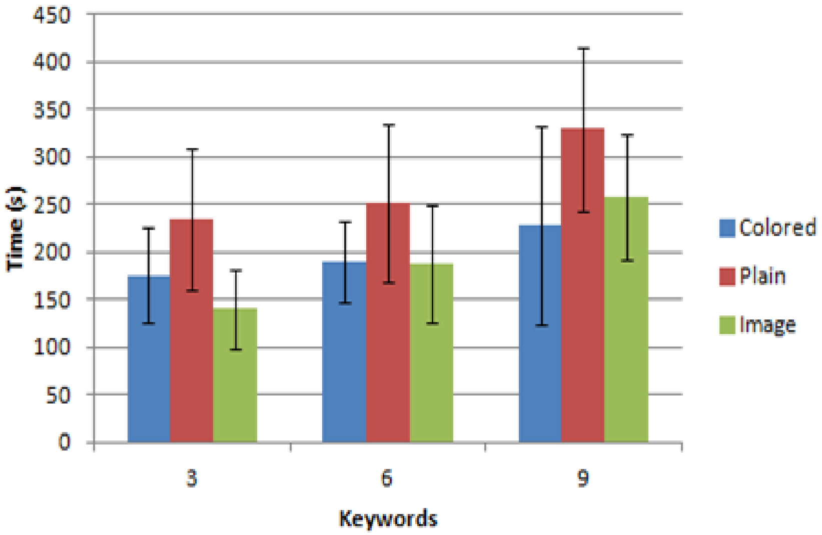


Figure 4.2. Reading time of different blog presenattion based on keywords.

Answering time. Each participant was given few questions based on blog's content. Their answering time was recorded. Accordingly there was not much significant difference between the answering time of Plain dictionary and Colored dictionary(V = 501.5, p = 0.38),nor there was any difference in Image and Plain dictionary(V = 741.5, p = 0.12). But we found significant difference between Image type and Plain dictionary(V = 779, p = 0.02). According to table 4.2, the results shows that participant took much less time to answer questions in Image type when compared to Colored or Plain dictionary type.

Table 4.2 shows the results for answering time where mean, standard deviation(SD) and median are recorded.

Table 4.2. Answering Time(SECONDS)

Blog Type	Mean	Median	SD
Colored Dictionary	159.8125	86.42858477	136.5
Plain Dictionary	167.9375	76.82015644	168
Image Type	134.4166667	64.42407153	123

Each blog was also analyzed to check if answering time has any effect of increase in the number of keywords.

For Colored dictionary, we found that there was significant difference found due to increase in the number of keywords. B_3 was significantly less from B_6(W = 156, p = 0.04), B_6 was significantly different from B_9(W = 93.5, p = 0.03).

But in Plain dictionary, there was significant difference noticed when number of keywords was increased from 3 to 9 ,B_3 was significantly less than B_9(W = 11.5, p = 0.005). There was close approximation between B_6 and B_9(W = 105, p = 0.06).

When we performed a one tailed test on sample, we found (p =0.03) and showed that the B_6 was significantly less than B_9.

For Image type, there was significance difference in B_3 and B_9(W = 41, p = 0.04). When we found close approximation between B_6 and B_9(W = 56, p = 0.09) in significance value to 0.05, so we performed one tailed test to further check if there was a difference. We found B_6 less significantly less than B_9(W = 56, p = 0.04).

Increase in number of keywords has greater effect on answering time when number of keywords was increased from 3 to 9 and 6 to 9.

Figure 4.1 below shows the mean time of different blogs reading time.

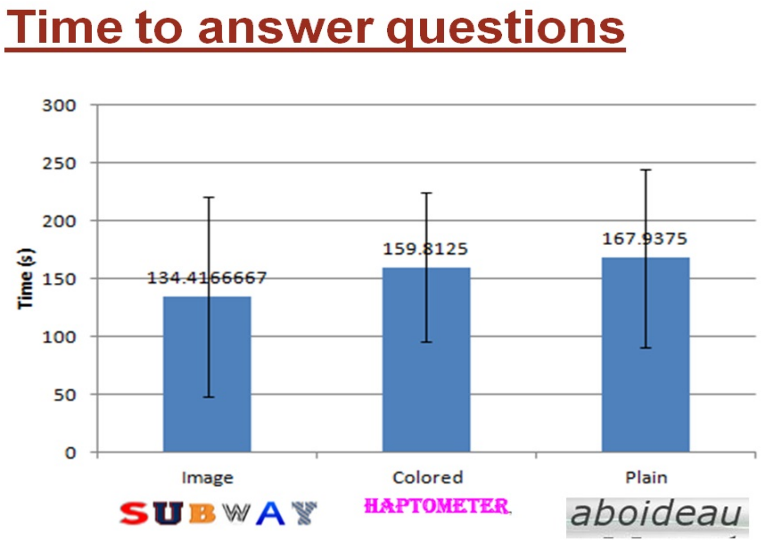


Figure 4.3. Time to answer questions.

Figure 4.2 below shows the mean time of different blogs reading time based on number of keywords.

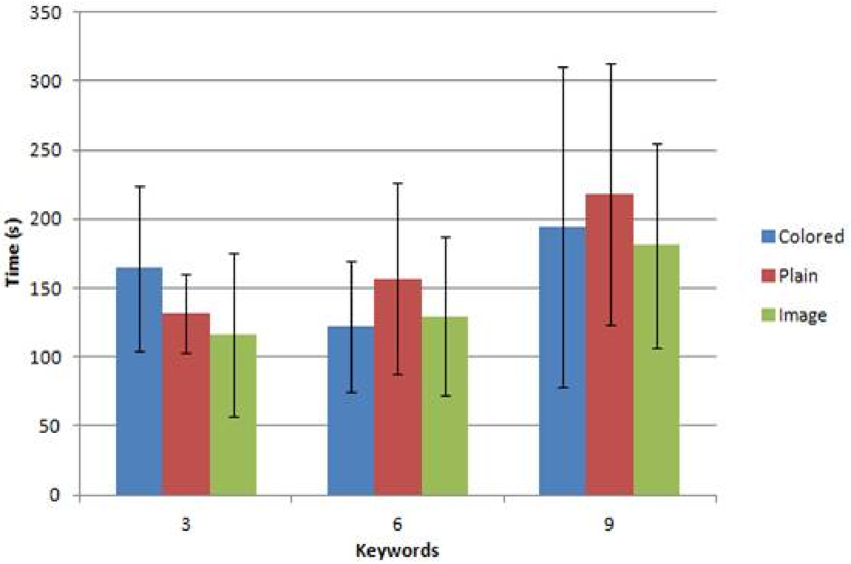


Figure 4.4. Time to answer questions for different blogs based on keywords.

4.2 Correctness

Figure 4.5 shows the performance of various participants on number of correct answers. participants who could answer only one correctly, and participants who gave all the answers correctly and so on

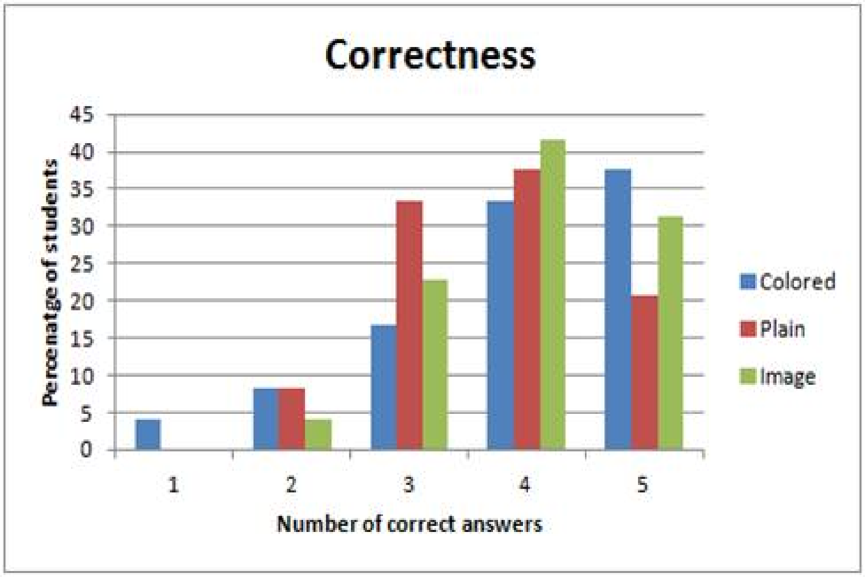


Figure 4.5. Correctness.

Most number of participants answered questions correctly.We also performed statistical analysis on various blog types and even based on number of keywords, we didn't find any significant difference.

4.3 Response

Figure 4.6 shows the response of various participants on different blogs.

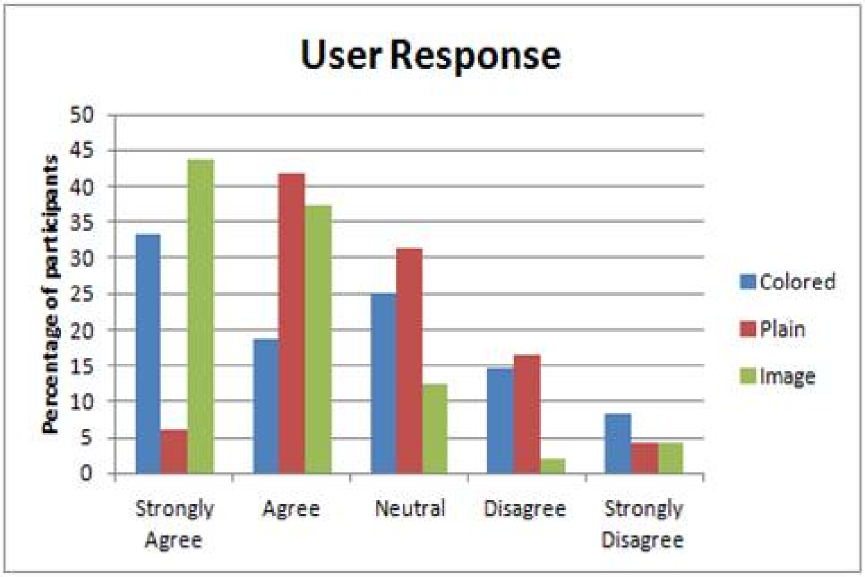


Figure 4.6. UserResponse.

To get more clear view of which blog was strongly agreed, we performed statistical test.

There was not much significant difference in between Colored and Plain dictionary. while Image was significantly less than Plain (V=606,p=0.00008), according to our scale participants accepted Image type more than Plain dictionary. Between Image type and Colored dictionary, Image is significantly less than Colored(V=435,p=0.0001). Image was accepted more than Colored dictionary.

We performed more test based on the number of keywords. For Colored dictionary, we found that there was significant difference found due to increase in the number of keywords. B_3 was significantly less from B_6(W = 62, p = 0.04), B_3 was significantly different from B_9(W = 42.5, p = 0.004).

But in Plain dictionary, there was no significant difference noticed when number of keywords was increased. But due to close approximation we performed one-tailed test to further analyze the significant difference,B_3 was less significantly different than

B_6(W = 68, p = 0.04). There was also a close approximation between B_6 and B_9(W = 68, p = 0.08). When we performed a one tailed test on sample, we found (p =0.04) and showed that the B_6 was significantly less than B_9. B_3 was also less significant than B_9.

For Image type, there was significance difference in B_6 and B_9(W = 39, p = 0.02). When we found close approximation between B_3 and B_9(W = 57.5, p = 0.08) in significance value to 0.05, so we performed one tailed test to further check if there was a difference. We found B_3 significantly less than B_9(W = 57.5, p = 0.01).

CHAPTER 5

CONCLUSION

In this section, we summarize the results we found from our study.

5.1 Time Complexity

We evaluated the data sample and performed statistical test, to analyze our three system model. We did a detailed analysis, we found that median reading time for participants in Colored dictionary was 191.5 seconds, while in plain dictionary we found that it was 261.5 seconds and in image type the value was 219. Interestingly average reading time for Colored dictionary(Mean = 203.1666667) was much less than Plain(Mean =268.7083333) and Image(Mean = 240.25). Participants took longer time in reading plain dictionary time blogs. Answering time for various blogs was also interesting. In this case median answering time for Colored dictionary was 136.5, Plain dictionary was 168 and Image was 123. Participants took more time in Plain dictionary(Mean = 167.9375), when compared to Colored(Mean = 159.8125) and Image(Mean = 134.416666). Least time taken to answer questions was in Image type.

5.2 Correctness

We did a detailed analysis on performance of various participants, if they could understand the blogs and answer questions correctly. Participants from UTA research pool took these test seriously and irrespective of any type either Colored dictionary

,Plain dictionary or Image they could understand and answer questions. We didn't find any significant difference in Statistical test.

When further analyzed to check the the number of students who failed to answer particular questions, we found that participants found negatively formulated questions, multiple answer questions and questions asking idea or tone or the of blog as difficult.

In case of multiple answers question about 28 % participants failed incorrectly. In B_6, participants failed to answer negatively quoted questions, there was only 50 % failure rate, while 43 % of students failed to answer questions talking about idea or tone of blog. In B_3, 44 % of questions were wrong which talks about tone of author in blog and 26 % were wrong on negatively quoted questions. In B_9, 38 % of participants failed to answer questions about idea of blog. Some of the easy questions were wrong, only explanation to it is participants were lazy or didn't pay enough attention. We need to take care of re-framing the questions negatively stated questions and question with multiple answers for future work, as they are quite confusing, that's makes difficult to analyses and compare the correctness value.

5.3 Response

Participant's response on the blog types were interesting. Many people strongly agreed to the Image type as it was easy to read, while many favored Colored type and Plain type. Around 33% of participants strongly agreed to Colored while for Plain dictionary 42% agreed to it, but only 6% strongly agreed to Plain. Most of the participants had neutral opinion about it.But Image type had Strong good response compared to any other. Interestingly Colored had 8 % strong disapproval while for plain and image it was 4 %. Our statistical test prove that overall there was no significant difference in response between Colored and Plain, while Image had better response, participants found image type easy to read.

REFERENCES

[1] 2012. Hiking gfw. http://hikinggfw.org/.

[2] Daniel Anderson. Splinternet behind the great firewall of china. *Queue*, 10(11):40, 2012.

[3] DNS as a Service. Opendns innovations. https://www.opendns.com.

[4] BBC. News world. http://www.bbc.com/news/world/.

[5] Chad Brubaker, Amir Houmansadr, and Vitaly Shmatikov. Cloudtransport: Using cloud storage for censorship-resistant networking. 2013.

[6] Sam Burnett, Nick Feamster, and Santosh Vempala. Chipping away at censorship firewalls with user-generated content. In *USENIX Security Symposium*, pages 463–468. Washington, DC, 2010.

[7] Stephen Chrisomalis. The phrontistery. http://phrontistery.info/.

[8] Global Internet Freedom Consortium et al. The great firewall revealed, 2002.

[9] Ultrareach Internet Corporation. Ultrasurf. http://ultrasurf.us/.

[10] Jedidiah R Crandall, Daniel Zinn, Michael Byrd, Earl T Barr, and Rich East. Conceptdoppler: a weather tracker for internet censorship. In *ACM Conference on Computer and Communications Security*, pages 352–365, 2007.

[11] Roger Dingledine, Nick Mathewson, and Paul Syverson. Tor: The second-generation onion router. Technical report, DTIC Document, 2004.

[12] Public DNS. public dns. https://developers.google.com/speed/public-dns/.

[13] Inc. Dynamic Internet Technology. Freegate. http://www.internetfreedom.org/.

[14] King-wa Fu, Chung-hong Chan, and Michael Chau. Assessing censorship on microblogs in china: discriminatory keyword analysis and the real-name registration policy. *Internet Computing, IEEE*, 17(3):42–50, 2013.

[15] OpenNet Initiative. Access denied: The practice and policy of global internet filtering (2008). https://opennet.net/.

[16] Josh Karlin, Daniel Ellard, Alden W Jackson, Christine E Jones, Greg Lauer, David P Mankins, and W Timothy Strayer. Decoy routing: Toward unblockable internet communication. In *USENIX Workshop on Free and Open Communications on the Internet*, 2011.

[17] Karl Kathuria. Bypassing internet censorship for news broadcasters. *Usenix. org*, 2010.

[18] Jeffrey Knockel, Jedidiah R Crandall, and Jared Saia. Three researchers, five conjectures: An empirical analysis of tom-skype censorship and surveillance. In *FOCI11: USENIX Workshop on Free and Open Communications on the Internet*, 2011.

[19] Erica Naone. The psiphon network frustrates censors by using trusted friends. http://www.technologyreview.com/news/409298/circumventing-censorship/.

[20] Jason Q. Ng. Blocked on weibo. http://blockedonweibo.tumblr.com/.

[21] non-profit organisation in France since 1995. Reporters without borders. http://en.rsf.org/.

[22] nongovernmental human rights organization nonprofit. Human rights watch. http://www.hrw.org/.

[23] University of Toronto. Citizen lab. https://citizenlab.org/.

[24] Global Internet Freedom Program. Freedom on net. https://freedomhouse.org.

[25] Weixiong Rao, Lei Chen, Pan Hui, and Sasu Tarkoma. Move: A large scale keyword-based content filtering and dissemination system. In *Distributed Com-*

puting Systems (ICDCS), 2012 IEEE 32nd International Conference on, pages 445–454. IEEE, 2012.

[26] RSS reader. Feedly. http://feedly.com/.

[27] Reply.com. Bloglines reader. http://www.bloglines.com/.

[28] Global Internet Report. Internet society. http://www.internetsociety.org/.

[29] Max Schuchard, John Geddes, Christopher Thompson, and Nicholas Hopper. Routing around decoys. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 85–96. ACM, 2012.

[30] AWS Services. Amazon simple storage service. http://aws.amazon.com/s3/.

[31] Ray Smith. An overview of the tesseract ocr engine. In *ICDAR*, volume 7, pages 629–633, 2007.

[32] Staples. Reading speed to the national average. http://www.staples.com/.

[33] Techopedia. Sneakernet. http://www.techopedia.com/.

[34] China Digital Times. Grass-mud horse dictionary. http://chinadigitaltimes.net.

[35] Transparency to the Great Firewall of China. Online censorship in china. https://greatfire.org/.

[36] Webopedia. Optical character recognition technology. www.webopedia.com.

[37] Wikipedia. Internet censorship circumvention.

[38] Philipp Winter, Tobias Pulls, and Juergen Fuss. Scramblesuit: a polymorphic network protocol to circumvent censorship. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pages 213–224. ACM, 2013.

[39] Eric Wustrow, Scott Wolchok, Ian Goldberg, and J Alex Halderman. Telex: Anticensorship in the network infrastructure. In *USENIX Security Symposium*, 2011.

[40] Tao Zhu, Christopher Bronk, and Dan S Wallach. An analysis of chinese search engine filtering. *arXiv preprint arXiv:1107.3794*, 2011.

BIOGRAPHICAL STATEMENT

Ritu R Patil was born in Karnataka, India in 1988. She successfully completed her Bachelor of Engineering in Computer Science in R.V. College of Engineering, India in 2010. She enrolled for master's program in computer science at the University of Texas at Arlington after working for two years in IBM, pune. Her area of interest are mainly focused on web developing and as java developer. She has also worked on the Android development and database related projects.