The Pennsylvania State University

The Graduate School

Computer Science and Engineering Department

ANGEL: A HYBRID CONTENT-BASED FILTERING TOOL FOR PROTECTING TEENS' SAFETY IN ONLINE SOCIAL NETWORK

A Thesis in

Computer Science and Engineering

by

Lei Zhang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December. 2008

The thesis of Lei Zhang was reviewed and approved* by the following:

Sencun Zhu
Assistant Professor (Department of Computer Science and Engineering &
College of Information Sciences and Technology)
Thesis Advisor

Daniel Knifer
Assistant Professor (Department of Computer Science and Engineering)
Thesis Advisor

Raj Acharya
Department Head and Professor (Department of Computer Science and Engineering)
Head of the Department of Computer Science and Engineering

*Signatures are on file in the Graduate School

ABSTRACT

Since its emergence in 1997 [1] , online social networks have experienced dramatic evolution and played an increasingly important role in our life. With the enlarging population of members of various social networks, the functionality of themselves also enhanced so as to continuously boosting its further development. Participants in such online communities can not only find new friends of common interests by browsing their profiles , they can also stay connected with buddies via various asynchronized so-called virtual channels, like messaging, blogs comments, or by built-in applications.[2] While the social network (SN) owners are encouraging more and more people in getting their memberships and staying with them, there is also increasing concerns of security issues accompanying this new wave of internet blossom. Spamming problem is a long lasting issue existing in many field of web applications, conceivably, SN is one of them without exception [3]; Phishing attacks usually come with the spamming activity which lead even worse consequence once the victim has been deceived; other inappropriate materials like offensive contents [4], web-bully [5] and web scams [6] are also frequently observed by social network users.

Security concerns mentioned above undoubtedly impair the user experience of web users, and are especially harmful to teens that usually have less capability in figuring out the malicious intensions behind the friendly disguise. Falsely trusted online friends could easily get valuable information from teens about themselves or even their family. This work addresses the possible threats on internet, evaluates state-of-art countermeasures, proposes and evaluates a light weight browser-based tool, named ANGEL, to help relieve the problem via hybrid-based content-filtering. It is novel in the sense that, we find there is no similar approach discussed in previous literatures or tools that could provide all around protection for teens' safety in online social network.

TABLE OF CONTENTS

Appendix A

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Sencun Zhu, for allowing me to pursue this work and for providing many valuable insights. I would also like to thank my co-advisor Dr Daniel Knifer, for giving me guidance on applying machine learning techniques in this work. I would like to thank Zhi Xu for his great comments and ideas. Last but not least I would like to thank my wife and my parents, without which none of this would have been possible.

Chapter 1

Background and Motivation

"We are connected despite the artificial barriers we construct"

-- Bruce Mulkey

1.1 Social Network

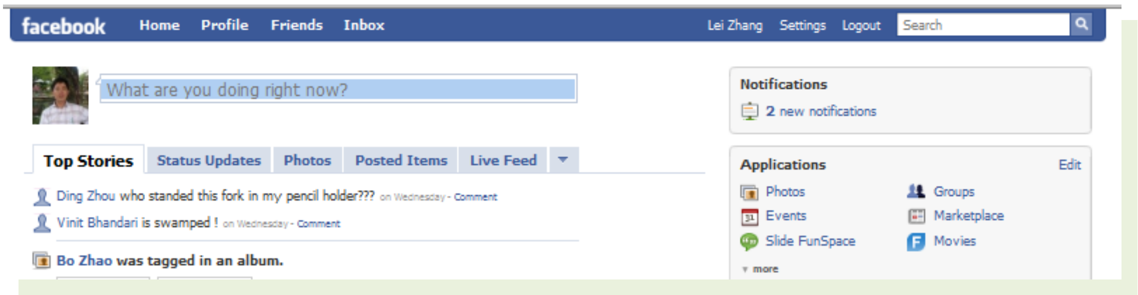
What is a social Network? A formal definition given by Wikipedia is that it is a social structure made of nodes (which are generally individuals or organizations) that are tied by one or more specific types of interdependency, this tie could be from values, visions, ideas, financial exchange to friendship, kinship, dislike, conflict or trade. As you can imagine, the variety of the interdependency between each people could produce various social networks even among same group of people.

If the phrase social network itself is still too academic to you, then you must be familiar with word “networking”, which is commonly used by every individual in the society. Networking is the activity that helps us to build a social network centered by ourselves, and these virtual networks we built might overlaps and interacts with each other and as a consequence ties us into the society we are living in. Generally, building up such personal social networks could take a long time and is also determined by many other factors such as the number of people you have met, the place you have gone, the experience or information you possess, your personality and preference and so on.

Extensive researches have been conducted in the area of social network analysis and studied interaction and evolution of social relationships in terms of individuals and ties. Research in a number of academic fields has shown that social networks operate on many levels, from families up to the level of nations, and play a critical role in determining the way problems are solved, organizations are run, and the degree to which individuals succeed in achieving their goals. [7]

1.2 Online Social Network

If the geographic location, differences in the personality and unbalanced information are top obstacles hindering people from sharing their ideas and building social connections, then these concerns would no longer be issues with the emergence of online social networking. When reviewing the evolution of people's ways of communicating over network, it turns out that our pursuit of better solutions never stops. As early as 1960s, much earlier than the internet ever comes into being, the first email was sent out to kick off this never-ending march, followed by invention of instant messaging, bulletin boards, chat-rooms, mud games and so on. Although these inventions did have profound influence on people's online life, they are somewhat restricted in terms of the amount information they could carry and the ways they could employ to present the information to their users. With the tide of Web 2.0 approaching, social network sites come into being and quickly penetrating itself into every corner of our society and now become a core meaning of bridging people in the modern world. In [1] ,Danah m. boyd and Nicole B. Ellison formally define social network sites as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. According to their study, first social network site was SixDegrees which was launched in 1997. After that, many other similar websites sprang up and then disappeared over the years and now the top competitors in this ever-growing business are MySpace, Facebook, Hi5, Linkedin etc.. Benefited by the evolving technology in the realm of both computer hardware and software, which yields cheaper storage space, faster CPUs and more fault tolerant and highly configurable software architectures, nowadays, these online social networks not only combine all major functions (i.e. messaging, BBS, IM, online games) that their forefathers have already possessed, they also features well-designed profile systems and multiple virtual channels for connecting people across the world. In such online communities, the number of people you could meet is no longer constrained by any physical or artificial boundary, nor by the time-zones of the locations they live, or even the language they speak with the help of state-of- art translation software.



Ding Zhou who standed this fork in my pencil holder??? on Wednesday - Comment

Vinit Bhandari is swamped ! on Wednesday - Comment

Bo Zhao was tagged in an album.

Figure **1-1**, a typical online social network website

As a consequence, more and more people choose to be the member of social networks, and sometimes even join more than one at a time. Recent study [8] conducted by Rapleaf indicates that, average social networks people associate with is around 2 to 3 in a pool of 120 million profile samples ages from 14 to 74. As an effect of crowd psychology, increasing population in these online communities drive even more stander-bys into followers. Latest statistics reported about Myspace , largest social network website so far[9], indicate that this snow-ball phenomenon is building an ever-seen huge utility with as many as 110 million participants, which is close to Japan, the 12th most populated country in the world, and this number is still increasing. It also claims that 1 in 4 Americans is on MySpace, and in the UK, it's as common to have a MySpace account as it is to own a dog. Data from the internet information provider Alexa[10] shows that, 3 out of 5 top trafficked in the global region are social network companies, with Myspace ranked 3rd, YouTube, a social network site sharing videos among members , rank the 4rd and Face book , the most fast growing social network site and top competitor to Myspace rank the 5th.

The boom of online social networking redefines our ways of social life by connecting people more tightly and making our means of communicating much more various than ever. Meanwhile, it is reasonable to induct that, when something is more attractive to the mass, it is undoubted to catch malicious parties' attention as well. Just image what you would do, if over 100 million people are just clicks away, let alone the countless searchable profiles listed there. We will further discuss the privacy and security issues in online social networks in Section 2.4.

1.3 Motivation of Study

As social networking sites have created an unprecedented ease of communication where people can present and distribute their own content, make links to their peers, and set up a powerful network for sharing information with very little effort. Teens are particularly drawn to this environment to voice their opinions and announce their independence. However, with spreading out of such large online communities, malicious parties have also noticed their opportunity and stepped on this virgin land. Shortly, inappropriate contents are easily accessible at almost every corner of these communities via all available virtual channels. We believe that, among all social network website users, teenagers are of most vulnerable to such kind of threat, as they are more sensitive, more likely to trust others and thus much vulnerable to various deceptions compared to other more mature individuals. According to reports from [11], to children aged between 13~15, acceptance and belonging are extremely important as they are building up their own identity at this period; at this age, they also demonstrate pseudo-maturity by trying to handle issues beyond their developmental capability; They usually distance themselves from their parents and but use the Internet to confide in others; Furthermore, their behavior often based on peer behavior rather than on whether that behavior is right or wrong. As more and more threats like spamming, phishing, offensive contents and web-scams [3][4][5][6] reported on social network websites, we believe that the online victimization of children becomes an issue of growing concern. And therefore effective method is called to help reduce crimes committed against children on such online communities.

1.4 Object of Study

This work is intended to achieve three goals, the first one is to systematically address the current threats we observe in social networks websites, with emphasis on those especially harmful to teens who generally believed to have less maturity and ability in differentiating true friends from liars. After we have examined the severity of this problem, we then look into several state-of-art techniques available and discuss their pros and cons, state why they are not sufficient in our case to fix the problems we're interested in. Finally, we will discuss the design and implementation of ANGEL and evaluate its performance with real world challenges while make comparison with counterparts.

1.5 Thesis Preview

The rest of thesis will be organized as follows: Chapter 2 first discusses several observations of the behaviors of teens in online social networks, illustrate the way they make friends with other, the preferable way they choose to communicate between each other, the kind of people they would like to trust and the most important, the extent of awareness of the privacy protection they are holding. Then next, in Chapter 3 we present various threats prevailing in SN order by their harmfulness to the web users, analyze the different forms they present in the context of online social networks. Chapter 4 goes over the related work studied in the previous literatures and other state-of-art tools we might be able to use to defense. Also, we will point out the Achilles' heels of these tools in terms of protecting teens' online social life and thereby justify our motivation of implementing a more appropriate tool to relieve the problem we are facing. Chapter 5 elaborates the breakdown details of the design principles and implementation of the tool and its functionality and originality as well. It is followed by Chapter 6, which is an in-depth evaluation of the performance of the tool towards threats we've discussed in Chapter 3 and explore the possible vulnerabilities it has. Finally, in Chapter 7, we will make the conclusion and also show a list of future tasks we could do to continuously improve the tool we have built in the long run.

Chapter 2

Online Social Life -- Observations

This Chapter goes over several observations stated in previous literature on how teens manage their social life in the context of present online social networks.

2.1 The Extent of Exposure

We have showed in Section 1.2 that the online social networks hit the main stream of people in current society. We want to ask following question: How many teens are among these people? PEW/INTERNET [12]'s answer from their survey conducted in 2007 says that 93% of all Americans between 12 ~ 17 years old use the internet and 61% of teens of age 14 – 17 use online social networks and the percentage is over 55% of teens of age 12 -17. Their finding is further verified by a more recent study on social networks user vs. age conducted by researchers at Rapleaf [8] in 2008 over a sample of 49.3 million people of age from 14 to 74 across users on Myspace, Facebook, LinkedIn, and Flicker etc. Based on the research result they have revealed, nowadays social media is still dominated by 14-24 year olds, with teenagers (aged from 14-17) as the second largest population in their all samples for Myspace. Business week even gave a special name, the Myspace generation [13], for those youth who use social networks on daily basis, making friend online, buying stuffs online and playing games online. Meanwhile, social network sites owners also notice the importance of the enlarging population of teenager members and try to cater them with more specific functionalities, like high school networks promoted by Facebook in year 2005. Moreover, there is a social network website named Myyearbook , with over 3 million unique visitors monthly, is especially tailored down for high-school users, as both founders of this company is under 17 when they launch the site[14]. Myyearbook has just received the 13 million dollars second round venture capital to further refine its business models and compete with leading competitors like Facebook.

Study also shows that, the teenagers who have registered on social networks would spend fairly a steady amount of time visit them on daily basis, In [12], result shows that almost half of teens who registered social networks visit them once a day to several times a day. One in six visits the sites 3 to 5 days a week and 15% visit them 1 to 2 days a week. Only less than 20 % go back to the social network site less often.

All of the above reports and surveys confirm the fact that Teens have been extensively attracted and exposed to the online social networks and are enthusiastic participators of this new kind of web application.

2.2 Information Shared Online

Usually, the first thing you are asked to do after registration is to create a profile for yourself, with which you formally introduce yourself to others by filling out bunch of forms or questionnaires. Profile management is regarded as the most important and fundamental function provided by most social network sites, as it can help you to polish yourself in the way you prefer others to see. The most common questions are like age, gender and the town you live; sometimes it will also include a breakdown list of twenty or more choices about your hobbies and personal preferences on specific topic. Beyond these textual data, you can further decorate your profile with multi-media materials, as some site owners allow you to upload pictures, music or even videos. All these data you’ve typed in are actually building up a unique vector of attributes of yourself, or an icon composed of all information you want others to see. Profiles also play key role in creating smart applications finding matching. What they do is to measure the similarities or correlation between users via computing Euclidean distance of two vectors in the feature space.[15] The smaller this distance is the larger the chance is that users these vectors associated with could have somewhat common grounds. From here, it seems that a reasonable induction could be made as following: in order to efficiently find the people you want to make friend with among the crowd of millions of members, everyone should provide as much authentic information as they can so as to achieve perfect matching all the time. Look into the survey result from [12], it shows that among all teens who registered as a member of online social network, 82% include their first name and, 61 % include their city or town, 79% include photos of themselves, 66% include photos of their friends and others information teens tend to share are boyfriend or girlfriend status, gender, pet information, parents' profession and various personal preferences such as movies, food, songs etc. However, there is always an invisible threshold at the bottom of everyone's heart about the degree of openness that you do not want to go beyond. The underlying philosophy is that, while you are eager to find a buddy whom you could talk with about your favorite football star or someone that may have same politic views as you do, you still do not want yourself to be too much exposed in such virtual public environment, where you have no idea who will be the next viewer of the profile you've just updated. For that matter, most people choose to post little information that might help others to match you with an offline entity. Similar observations happen to Teens. Compared to high percentage of willingness to sharing those obscure information, which is either too hard to identify an individual (like the first names) or simply not searchable (like images), 49% include their school's name, 40% share their IM name, 39% include a link to their blog, only 29% include their last names, 29% list email addresses , and only 2% state Cell phone numbers [12].

As life is full of paradox, in some occasions, people might still want to share these sensitive data with some of their close friends, acquaintances or coworkers for the purpose of convenience. Most social network sites notice their customers' concerns and wishes therefore add access control features to their profiles systems. Hence now the users could choose to be protected by isolating himself/herself from the mass crowd with customizable restrictions. In general, restrictions could be configurable settings users pick applied either to the content they want to make public or the privileged group they want to share with. In the case of Facebook, you could choose a privileged group from direct friends, friend of friend, group you belongs to, to all registered members at Facebook. Admittedly, turning on this option unconsciously construct a invisible fence around yourself, and meanwhile it also means that less people would possible interested in talking to you if nothing interesting they could tell from the little pieces of information you make open to the public. Back to teenager users, reports [12] show that a total of 66% of all teens who have ever created an online profile in some way restrict access to their profiles, they are generally educated and aware of the danger of sharing secretes online by parents or teachers.

2.3 Communication Ways

In this section, we will talk about the various ways people could use to communicate with each other under environment of online social networks. If we believe browsing others profile is just the first step to meet people on the social network, then it is the interactions afterwards that help establish friendships and gradually construct the social network itself. Put it in a real world scenario, if we say the profile browsing is equivalent to exchanging business card or first conversation in the first meeting, then sending friend request thereafter is alike to mailing or calling someone you met last time in the real world case. More than just relying on these simplest communication ways, in a modern social network site, members are often equipped with a collection of tools to convey their ideas to others in an very efficient way, these could include but not limit to sending him/her a e-gift, tagging your friends pictures, or writing posts on your friends wall, some other actions are even not intended to passing any information, instead what the initiator wants to do is merely a sense of emotion, like “poke” or “hug” someone [16], etc. All of these viral channels are primary ways people use to connect with their friends and find new friends online. RockYou’s star software Super Wall, which allows your friends to post messages, photos and videos on your “super” wall in Facebook, was downloaded by 2 million users within two weeks after it was introduced and now has 3.7 million active users [17]. Super Wall is also claimed to be the fastest growing platform for reaching new users on Facebook. Previous survey [12] shows us some statistics how teens use these tools. According to their study of samples, 33% teens wink, poke, give gifts to friends, 61% send group message to all of their friends, 76% post comments to a friend’s blog and 84% post messages to a friend’s page or wall, and 82% send private messages to a friend within the social networking system.

The popularity of these gadgets among teenager members reveals their general acceptance of using such tools for their online social life, as these tools are not only easy to use even to a teenager, but also powerful and flexible in terms of various functionalities and funs they give you.

2.4 Security Concerns

In last two sections, we have showed that Teens are largely exposed to the online social network, and count for a large volume of visitors. They enjoy the convenience ways of various viral channels provided by the social network sites and use them frequently to manage and extend their friendships. Meanwhile, they are also aware of the danger of sharing too much information with the public and intentionally choose who they want to share and what they want others to see. However, in an open environment where no restriction on whom you could send friend requests, there is still a large chance that you will receive requests from someone you’ve never know before. In fact, [12] has reported that 43% of teens who use social networks have been contacted by a stranger online, and it also says some of these requests get approved if not all of them.

Generally, it could happen in two cases. First, if a friend request is sent by someone who is a friend of the teens' offline friend, then his request will normally be approved. The reason for this is quite intuitive, a friend of your friend is more trustful than a totally stranger. In the other case, some teens will add strangers into the friends list just for fun or the eager of attention, as they believe the number of friends on the contact list implicitly indicating that they are much more popular than other peers. As a consequence, they have a large portion of names on their friend list that they might never meet before in person or know by other meanings, therefore, no way of verifying the true identity of these personals they are talking to. Remember that we have also mentioned that friends are usually granted more privileges than other members and therefore could either get chance to view your restricted profiles, send you private messages, invite you to use some application or post comments on your 'wall' and 'space' . In next Chapter, we will discuss in detail what kind of threat could be introduced once this Pandora's Box is opened,

Chapter 3

Threats in Online Social Network

3.1 Data Harvesting

Data harvesting is a variation of the web crawling, which refers to using program or scripts to browse the World Wide Web in a methodical automated manner. [18]. The main difference between them is that , most web crawlers created by search engine companies behave with the courtesy towards the webpage owner , thus only would gather information under their permission. (E.g. They follow the rule defined by robot.txt. or just skip password protected pages.) . On the contrary, data harvesters with malicious intension usually do not follow these rules. Greedy as they are, they collect every piece of data they could find and are always trying to circumvent the defenses. Data harvesting is becoming an increasing issue to all social network websites, as it not only threatens the privacy of social website members, its automated crawling fashion also consumes a large portion of vender’s network resource and further impairs users’ surfing experience. Most venders have been aware of the problem and raise fences accordingly. As Max Kelly, security lead at Facebook claimed [19], “Data harvesting has become an issue for us, such harvesting attempts were generally unsuccessful but that doesn't keep people from trying”. When they realize that it is difficult to stop such trails, they then make the compromise. What they do is that instead of taking the gun from the people, they teach them how to use it properly in a legal way. To this end, rich platform APIs are released to encourage freelancers and other third party vendors to access the restricted data and contribute to the development of the website by creating new applications. These APIs include those released by site owner themselves [14][20] and open social platform [21] announced by the search giant Google.

Although data harvesting problem has not been solved yet, we believe it will be taken care by the vendor of social network websites thus should not be taken into account in our proposal.

3.2 Spamming

Spamming problem is one of the major threats we may face in social networks web sites, we will first go over its origins and classic countermeasures, then discuss its occurrence in the SN scenario.

3.2.1 Spamming in Email system

Spam problem has been a long lasting issue that keeps bothering us since decades ago. As recognized by most people, the original spam problem roots in internet ads or unsolicited bulk emails. The first message flagged as spam was sent to the users of Arpanet in 1978 and incurred little more than an annoyance at that time. With the evolution of the internet itself and the increasing number of both cyber-users and services providers, the influence of spam problem has become much severe than ever before. As is showed in the email threats trend report provided by Commtouch [1], spam emails account for nearly 75 percent of all emails in the 1st quarter of 2008. Not only annoying emails users, nowadays, Spam is also believed to be the major source of virus, Trojan Horse and other malwares. As spam constitutes up to majority of the total volume of email messages on internet, the study of anti-spam techniques gradually draws more and more attention of researchers all around the world.

Among all of the spamming variants, unsolicited commercial emails or Email spam is regarded as the most well-known one in the entire spam family. Not only because it is the origin form of all other types of spamming, it is also because its large population of victims, as the number of its targets is generally proportional to the number of email users. What makes things worse is that, large volume of email spam will also contend for the limited network resources and further impair the quality of service of other innocent web surfers who even has not received the spam. The war against email spam has never gone to an end from the day it is firstly introduced. In its earliest form, a spam email simply contains the promotion information from the merchandiser, as more and more people use filters to protect their Inboxes, spammers have thought of various ways to disguise their malicious activities. On the other side, researchers and spam-fighters are keep watching the evolution of the spamming tricks and have thought of many smart countermeasures accordingly.

In general, modern anti-spam approaches could be divided into two major categories with regards to the stage of email processing the defense comes into effect. Figure **3-1** and **3-2** are summaries of state-of-art techniques for defending spam emails. Reader could refer to [23] for details.

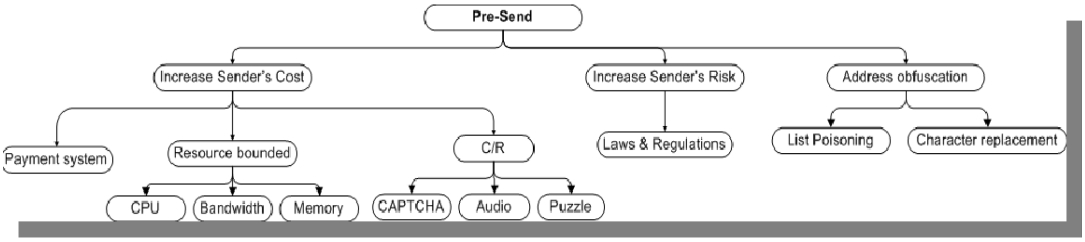


Figure **3-1**: Anti Email-Spam Techniques : Pre-Sending enhanced based on [23]

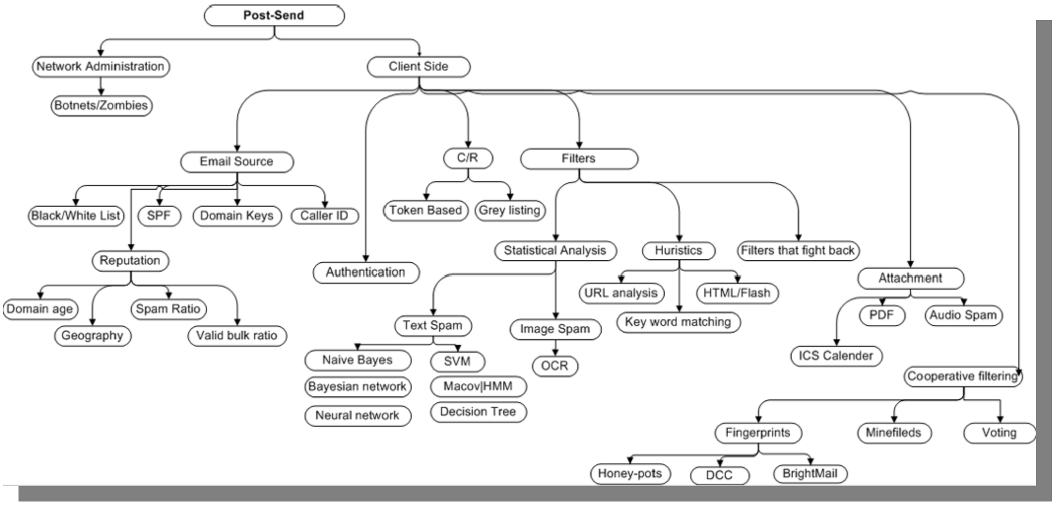


Figure **3-2**: Anti Email-Spam Techniques : Post-Sending enhanced based on [23]

3.2.2 Spamming in Search Results

Another place where spam prevails is the field of search engines. Web spamming, also called spamdexing has grown to a significant percentage of all web pages (between 13.8% and 22.1% of all web pages [24],[25]), threatening the dependability and usefulness of web-based information. Compared to history of email spamming, the existence of web spamming or search engine spamming is relatively shorter, nevertheless, in term of impact on people's cyber-life, it quickly becomes abreast to and will probably overtake email spamming soon. Success of commercial Web sites depends on the number of visitors that find the site while searching for a particular product or keyword. As the fact that most of searchers only look at the top N of results, spamdexing or web spamming soon becomes popular. The formal definition of web spamming or spamdexing activity refers to pages that use techniques to mislead search engines into assigning them higher rank, thus increasing their site traffic. As shown in Figure **3-3,** web spam can be generally classified into three streams, which are content-based, link-based and hidden-based.

While different in detail implementation, the shared goal of all these approaches are attempting to increase the ranking of these or some affiliated pages without improving the utility/quality of the content to the viewers.

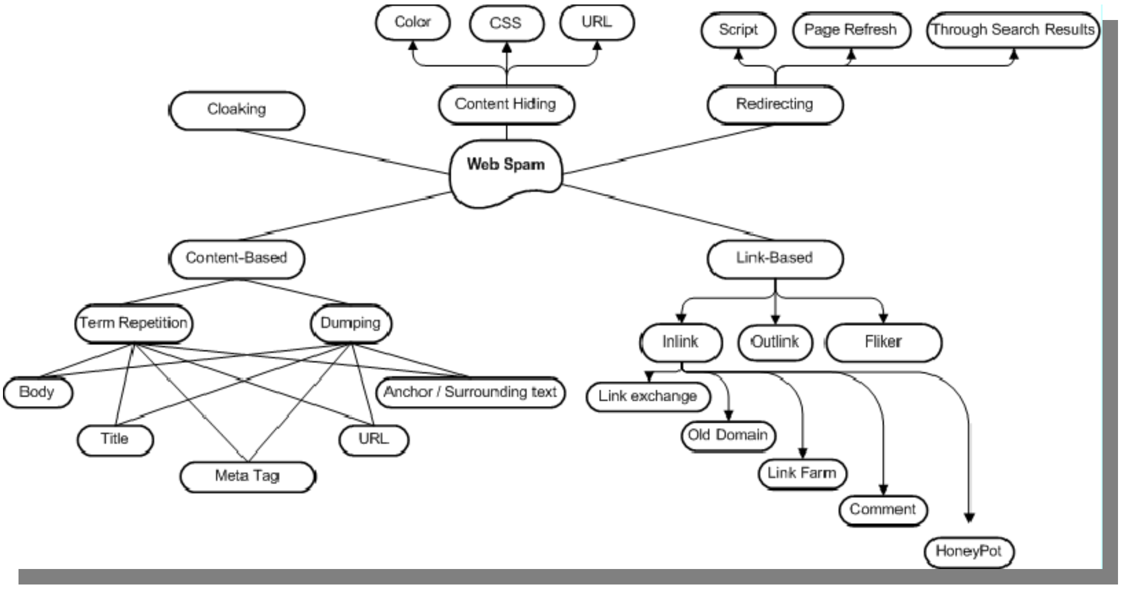


Figure **3-3**: Web Spam Taxonomy enhanced based on [26]

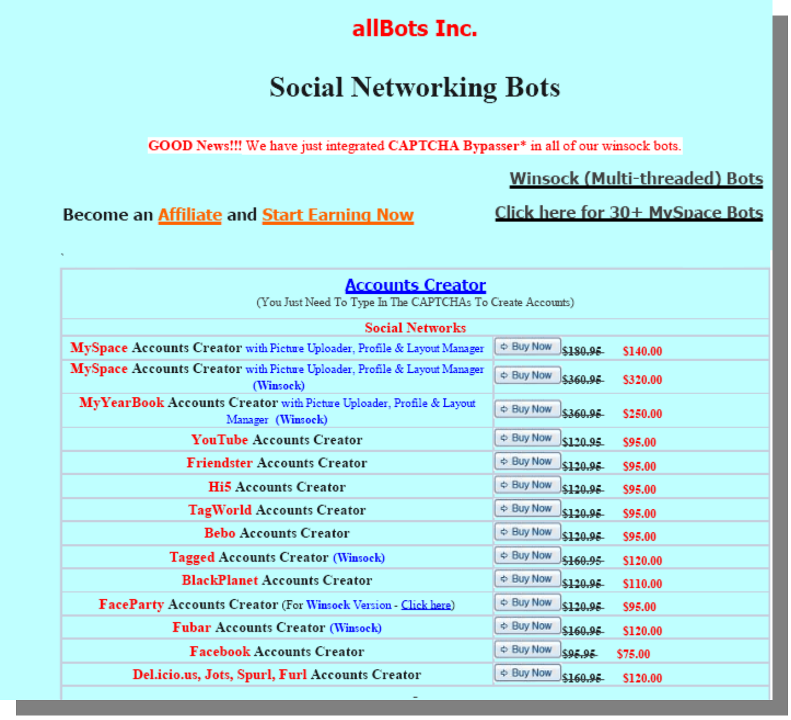
3.2.3 Spamming in Social Network Websites

When people are still striving to clean spam messages out of our mail inboxes and search results list, no one has notice that spam has stealthily stretches its arms into other territory and find a new target soon, that is, online social network sites. Recent reports from both the Anti-spam products vendor and the insiders of social network websites confirm this overlooked tendency. Harris Interactive [27] claimed in their report that among 2000 adults participated in their recent survey, 4 of 5 users received at least one spam communication via their social network within the last year. And over a third noticed an increase of spam attacks in the last six months. Leader of security team in Facebook also admitted that attacks become a serious problem in the website since it is first noticed in January. Anti-spam vendor Cloudmark claimed in [27] that current spam in social network could be categorized into following types:

- unwanted friend requests;
- bogus messages;
- spam comments;
- other spam-type communications

Given the list above and compare with the available viral channels we've discussed in Chapter 2 , you will find that spammers are using all possible communication means in the world of online social network to maximize the number of victims of their attacks.

By looking at the possible ways spammer could possible use to disperse the unsolicited messages in a social network, we wonder how much preparation they could have to do before launching a successful attack. Remember that, while in the case of email system, you could simply put some bogus information into section of return path and from address according to the specification of email header, imitating that the email is composed by real person and then only need focus of the content of the spam you want to send out. However, in the case of social network, there is usually a basic restriction that you have to be at least a member of social network before you can actually send out anything to other members, and you need go through the registration process and fill out quite a few forms. Further investigation shows that to get the account, a spammer could either create them manually which is both time consuming and tedious, or instead, do it with the help of some programs.



allBots Inc.

Social Networking Bots

GOOD News!!! We have just integrated CAPTCHA Bypasser* in all of our winsock bots.

Winsock (Multi-threaded) Bots

Become an Affiliate and Start Earning Now

Click here for 30+ MySpace Bots

Accounts Creator

(You Just Need To Type In The CAPTCHAs To Create Accounts)

Social Networks

MySpace Accounts Creator with Picture Uploader, Profile & Layout Manager	Buy Now	~~$180.95~~	$140.00
MySpace Accounts Creator with Picture Uploader, Profile & Layout Manager (Winsock)	Buy Now	~~$360.95~~	$320.00
MyYearBook Accounts Creator with Picture Uploader, Profile & Layout Manager (Winsock)	Buy Now	~~$360.95~~	$250.00
YouTube Accounts Creator	Buy Now	~~$120.95~~	$95.00
Friendster Accounts Creator	Buy Now	~~$120.95~~	$95.00
Hi5 Accounts Creator	Buy Now	~~$120.95~~	$95.00
TagWorld Accounts Creator	Buy Now	~~$120.95~~	$95.00
Bebo Accounts Creator	Buy Now	~~$120.95~~	$95.00
Tagged Accounts Creator (Winsock)	Buy Now	~~$160.95~~	$120.00
BlackPlanet Accounts Creator	Buy Now	~~$120.95~~	$110.00
FaceParty Accounts Creator (For Winsock Version - Click here)	Buy Now	~~$120.95~~	$95.00
Fubar Accounts Creator (Winsock)	Buy Now	~~$160.95~~	$120.00
Facebook Accounts Creator	Buy Now	~~$95.95~~	$75.00
Del.icio.us, Jots, Spurl, Furl Accounts Creator	Buy Now	~~$160.95~~	$120.00

Figure **3-4**: A Price List of Social Network bots [28]

As [28] claimed, an account creator could help you generate accounts with minimum supervision; they could fill out the forms in an automatic manner and generate required data format on the fly. Some social network sites have notice that and try to stop these attacks by adding email verification process and CAPTHA to stop these auto-registration tools. However, as reported in [29], CAPTHA is not longer safe as we expect it to be, even by adding more noise and lines to distort the image of numbers or characters, there's still some way that the attacker could use to bypass it. A recent instance is that allbot.info [28] announces that all bots provide by them could help create accounts without any assistance of user and bypass all kinds of CAPTCHA with the help of human assistance provided as service by third party. Well, as you can expected, once the spammers get the ticket to the show, destructions he could make is only limited by their imagination. What makes it worse is that, now spammers are not starting from the scratch, which means, any previous techniques employed in spamming in other field of web applications (email spamming and search engine spamming) could then be quickly adopted and applied.

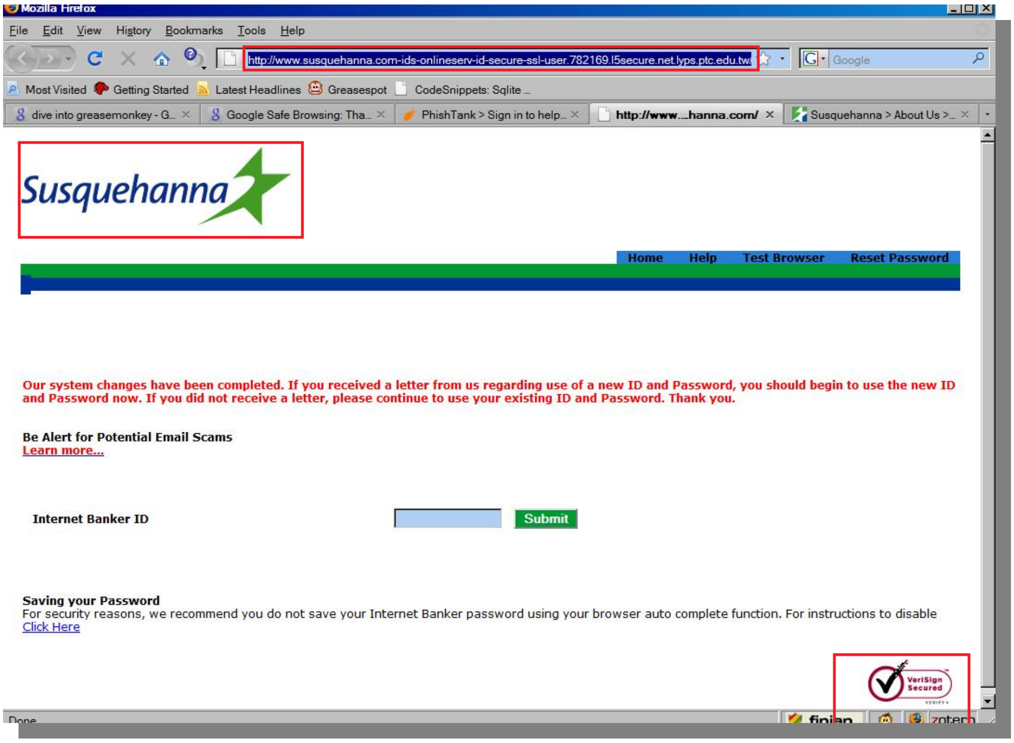
According to [30] , top spam categories are medications and health-related goods and Services (22.7%), Education (14.3%) Fake designer tools (11.3%) and Adult Content Spam(5.2%) . As most of these spams information contains advertisements which exaggerate the real effect of the products, or deducing viewer to perform some actions that might impair their interests, it is conceivable that we should keep them away from our children.

3.3 Phishing

Phishing is a tactic used by Internet-based thieves to trick unsuspecting victims into revealing personal information which they can then use to access the victims' financial accounts.[31] Sometimes it is also defined as a variation of traditional spamming activity as it is mainly distributed via emails but does not advertise any product or service. If we say receiving spam messages is just nothing more than annoying, then phishing is believed to be more dangerous than any other conventional spamming activities, as it is targeting the sensitive data which might include users' accounts and passwords, credit card information, social security numbers .etc. Furthermore, it is usually the prelude of more severe attacks. Just imagine the scenario that a compromised legitimated email account could be used to spamming all your contacts' inbox, stealing confidential corresponding and possible be used to reset all your passwords to online services then simply access protected services or material as if you are doing it in person.

In general, there are three kinds of phishing attacks that cybercriminals could possibly use, which are typo based, link based and malware-based ordered by the difficulties of being detected

URL based phishing attack is of the most naive among the three, basically, it fools the careless victim who does not pay attention to the URLs he types in or the hyperlinks in the email body before submit the critical information. The trick the most Phishers plays is to register a domain with URLs either similar to the authentic websites or hard to differentiate by a normal user. Some users are reluctant to tell the difference between lower case English characters and symbols like 1 vs l and 0 vs o in the url, others may be full by a url contains ebay or other brands name as they look so similar to be a domain owned by that brand. It makes it look more trustful by using similar CSS scheme, background color, logos, even with a symbol of Verisign, which normally endorse the authenticity of the page. [Figure **3-5**, Figure **3-6**].



http://www.susquehanna.com-ids-onlineserv-id-secure-ssl-user.782169.l5secure.net.lyps.ptc.edu.tw/

Our system changes have been completed. If you received a letter from us regarding use of a new ID and Password, you should begin to use the new ID and Password now. If you did not receive a letter, please continue to use your existing ID and Password. Thank you.

Be Alert for Potential Email Scams
Learn more...

Internet Banker ID

Saving your Password
For security reasons, we recommend you do not save your Internet Banker password using your browser auto complete function. For instructions to disable
Click Here

Figure **3-5**: An example of phishing attack

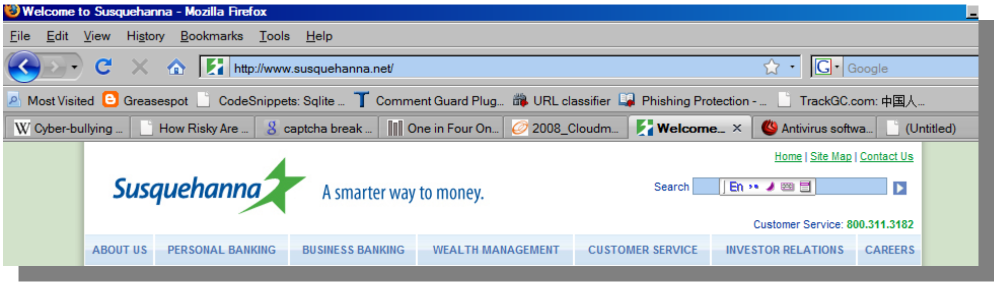
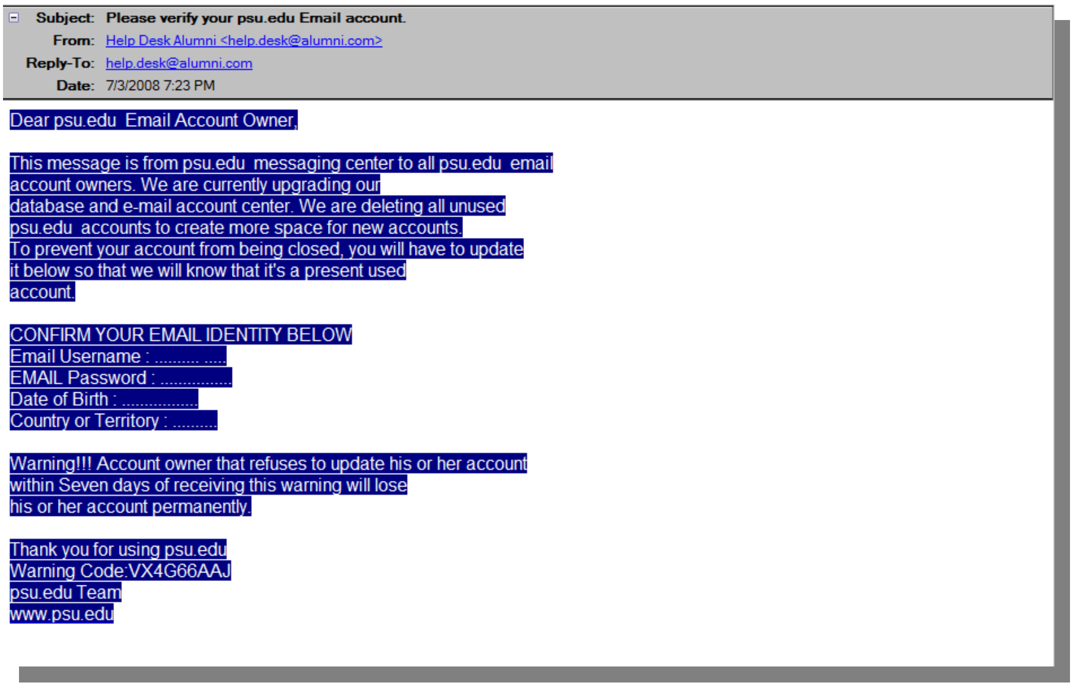


Figure **3-6**: An example of phishing attack (the real website)

Linked based phishing attack is more preferred by spammers as it is considered to be a more aggressive attack manner in comparison to passively waiting for a typo to happen. Normally, links to the phishing sites are usually embedded in legitimated messages, either in the form of emails or posts from legitimate user. Unlike the most spamming emails, phishing emails are more difficult for user to distinguish as it is carefully designed to look as similar as an authenticate email as possible. They usually contain web-links by which direct you to a phish site under the spammers' control. Sometimes the phishing mails are easy to tell as it is sent from an address or domain that seems not to be the authenticate source, or the link is directing you to a website which is different from the one you used to visit. More dangerous Phishers [Figure **3-7]** fabricates emails sent from the service provider or the customer service department of the banks and ask you to reset your account password, and they warn your access will be banned if you refuse to do so. In the context of SN, such phishing links could be easily spread out by phishers via in-site messaging system, or blog comments, etc.



Subject: **Please verify your psu.edu Email account.**
From: Help Desk Alumni <help.desk@alumni.com>
Reply-To: help.desk@alumni.com
Date: 7/3/2008 7:23 PM

Dear psu.edu Email Account Owner,

This message is from psu.edu messaging center to all psu.edu email account owners. We are currently upgrading our database and e-mail account center. We are deleting all unused psu.edu accounts to create more space for new accounts. To prevent your account from being closed, you will have to update it below so that we will know that it's a present used account.

CONFIRM YOUR EMAIL IDENTITY BELOW
Email Username : ....
EMAIL Password : ....
Date of Birth : ....
Country or Territory : ....

Warning!!! Account owner that refuses to update his or her account within Seven days of receiving this warning will lose his or her account permanently.

Thank you for using psu.edu
Warning Code:VX4G66AAJ
psu.edu Team
www.psu.edu

Figure **3-7**: Another example of phishing attack

A more sophisticated link-based phish attack is discussed in [32], in which clicking a hyperlink in the email will direct you to the real website showing in the background while at the same time, a malicious window is popped up in the foreground. This kind of attack is considered more dangerous as it intends to mislead the observer to believe the popped up window is affiliated with the window at the background. Other forms of link based phishing attacks using links with legitimate appearance to cover their malicious intention. A case in point is that, redirection attack appends the malicious web link to a URL which usually indicates a higher reputation. [33] shows an attacker using yahoo search result to redirect the browser into phishing links.

The most hostile phishing attack is the one luring you to install a piece of malicious code without notice. Without explicitly stealing you your credentials by fabricating a story or leading you to a “shadow” website, these attacks usually look more friendly as it starts with providing a link showing you some video or other interesting stuffs hosted on another website. [34] shows a recent phishing attack asking the user to download a installer for the latest version of flash player so as to view a hot video, while the truth is that, the installer the user is trying to download has nothing to do with the flash player but in actual a malware. Users of social networks have already been reported to be the target of such attacks. [35] reported worms stealing user's account and turning the legitimate Facebook accounts into Bot-nets and then disseminate spam messages in batch to all their contacts on Facebook. An in-depth investigation into the root cause of this attack reveals that all victims have unconsciously installed the malware somehow before the attack campaign has been launched.

3.4 Offensive Contents

Potentially offensive content can include gutter language, jargon, argot, sexually explicit material, racist, graphic violence, or any other content that may be considered offensive on social, religious, cultural or moral grounds. Another kind of offensive content, cyber-bullying, which more often happens in conversations, like instant messages or blog comments, is defined by [36] in one of these forms: flaming, harassment, denigration etc. In addition, in some cases, despite its legal purpose, offensive content may also be considered harmful to children. 4% participants age 10 to 15 year-old report such an incident on a social networking site in a national cross-sectional online survey [37]. These kinds of threats have already become a serious issue to our society. [38] reported a 13-year-old Missouri girl who killed herself after receiving cruel messages on the web in November 2007. Same report also mentioned that, on Oct. 7, 2003, at the age of 13, American boy Ryan committed suicide. Ryan suffered from long time depression, and it is believed to be caused at least in part by a steady barrage of electronic harassment. To most adults, offensive content is merely annoying, as their level of maturity could help them to handle the bad impact from this content properly whereas in children's case, who has much less maturity and social experience, the potentially harmful impacts on their healthy development is significant, in some extreme cases, offensive contents even create a negative perception of the internet as a whole in their mind.

In terms of occasions you could find these offensive contents, they could be categorized into two kinds, the website as a whole or the inline context of web-pages. As internet is such a powerful tool in contemporary life, it can turn a home, a school or a library into a place of unlimited information and communication, even if they are located in rural area. Search engines further help the users to find interested material in just few clicks. However, along with all these benefits there are risks as well, as there are also so many websites providing pornographic, religious, terrorism, extremism contents over the internet. Luckily, to address this problem, various web filtering tools have come up and using black listing, heuristic rules or labeling information provided by third parties like ICRA to prevent the children from being exposed to such inappropriate content. We argue that most tools could only solve part of the issue as we believe websites which publishing inappropriate contents are much easier to be identified and then blocked, while filtering jargons and racisms words appearing in blogs or comment system in SN is considered a more challenging task. Apparently, applying black listing or labeling is not realistic in this case as the website itself is innocent. Heuristic rules might help, but rules are either difficult to compose or impossible to be complete. As most parents are not skilled computer users, asking them to think of a lot of rules without assistance or find out most effective keywords are sometimes tough tasks. Furthermore, in some cases it is simply not possible to write up a rule. As you can image, there are many ways to describe the same thing, and a slight change in some letters of the word could successfully bypass the heuristic rules while carries same meaning to a human, in such case, attempts to exhaust all possible combinations is also in vain. Thus a much powerful approach is called to help deal with such problem.

3.5 Web Scams

Scams are activities that intend to deceive another individual so as to gain undeserved service or property. National Consumers League [39] lists top 10 internet scams in 2007

a) Fake Check Scams	f) Advance Fee Loans
b) General Merchandise	g) Prize / Sweepstakes
c) Auctions	h) Phishing
d) Nigerian money Offers	i) Friendship and Sweetheart Swindles
e) Lotteries / Lottery Clubs	j) Internet Access Services

Table **3-1** Top 10 Internet Scams

At the beginning, these scams are mainly distributed through spam emails, with the prevailing of personalized emails filters, conman turn to another virgin land, social network sites, where they could efficiently and safely distribute scam messages to others without being identified and later on being caught. Besides ten internet scams listed above, which target all age victims, some conman hide their malicious intension and impersonate themselves as children or teachers to get teens' trust, once teens are fooled and the conman is added into teens' friend list, he could then easily seduce the teens into doing something that is either harmful to themselves or to their families.

3.6 Summary

In their early days, social network site like Facebook and Linkedin used to be closed to the public and only accept new registrations by invitation. The key turning point in Facebook's history came in September 2006 when the site switched from being a closed community of students to a global online community for everyone on the internet. Once the door is widely opened, it is hard to tell if the new comer is an angel or a villain. Besides, in order to keep the rapid rate of growth, overpass the competitors and maximize the benefits of stockholders, social network sites owners developed various ways in their system to cater member's love of easier and more flexible way of communication. Promotions of these viral channels, Facebook's super wall application for example, are really successful business plans in term of their popularity among users and the number of new registrations they attract, on the other hand, they also become the soil of widespread mass spam and other inappropriate contents. Just as the Max Kelly , senior director of security team in Facebook, admits "We are definitely a target for spammers" [2]. We have to admit that we are already in a war against these malicious parties, so in the next section we would examine what kind of weapons we could use to fight back.

Chapter 4

Related work

In this Chapter, we examine possible solutions to the problems we have stated in Chapter 3. We categorize solutions into non-technical and technical ones as following:

4.1 Non-technical Solutions

Laws enforcement and parent educating are two major non-technical solutions to help reduce crimes committed against children over the Internet. Governments of Canada [40] have been aware of the situation and take actions to either legislate or fine-tune existing laws to make them more applicable and enforceable in the networked environment. Moreover, detailed guidance of online safety composed by professionals is published and distributed freely online to all parents and children. In countries like Canada, educational events are held on Safer Internet Day to teach parents about their children's online interests and encourage them to speak about ways to stay safe. In America, in terms of fighting against offensive contents on web, at least seven states, including Iowa, Minnesota, New Jersey and Oregon, passed cyber-bullying laws in 2007. Five more — Maryland, Missouri, New York, Rhode Island and Vermont — are considering similar legislation this year. [41]

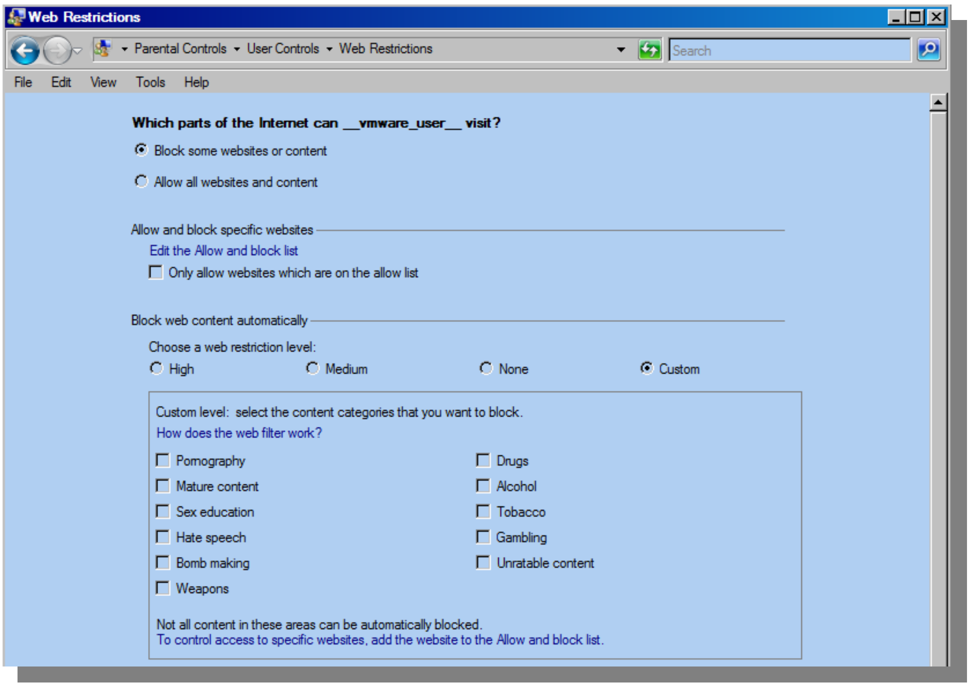
Tutoring from family members is another crucial step to reduce the influence on children from cyber threat. In FBI's "A parent's guide to internet safety"[42] , they suggest parents that instead of forbidding children from going on-line, they should guide them to use the online service properly and educate their children to these dangers and taking appropriate steps to protect them, so that they can safely benefit from the wealth of information available online. Besides, educators from [43] suggest that signing a contract between parent and teens is a simple way to teach child Internet awareness. Their reason is that kids love to be given responsibility, and this contract will make a big impression.

Although we believe laws enforcement and parent educating could somewhat help reduce the chance of children being exposed to the inappropriate contents over internet, however, both of them cannot help once it happens. Thus, technical solutions are called to tackle the issue.

4.2 Technical Solutions

4.2.1 Parent Control in Microsoft Internet Explorer (IE)

IE probably the most widely used browser in the world, as it shipped with every personal computer with windows operation system preloaded. According to [44] , internet explorer counts for more than 70% of market share among all browsers. As claimed in the IE Blog from MSDN website, the latest version of the OS, which is windows vista has equipped with a most powerful tool for the parents to arrange their children's cyber life. The parental controls feature, not only provide parents options to limit the way their children can use the computer, like time limits and application restrictions, but additionally, it let them be able to keep better track of what their kids are doing online and protect their kids from inappropriate web contents. More specifically, the tool they employ is one kind of web filtering system. Officially named as windows vista web filter (Figure **4-1**) , the web restrictions function it features runs with the help of an web service provided over the internet, from which the system could get the knowledge of the content of the website the URL links to before the user actually clicks and views the content the URL points to . This web service provides a fine-grained classification of unwanted web sites into over 11 categorizations. Parents could then make a customized selection among, pornography, mature content, sex education, hate speech, bomb making, weapons, drugs, alcohol, tobacco, gambling, or un-ratable content. The rating and classification of websites are normally based on the labeling specification maintained by a third party organization like ICRA [45] (Figure **4-2)**. Besides, customizable white and black list are provided as a supplement to the standard categorization system, in which parents could explicitly permit or forbidden the access to specific websites by putting them on different lists. Moreover, extensive logging also accompanies with the web filtering feature, provides statistics like the top 10 frequently visited websites by the children or websites most recently been blocked.


Web Restrictions
Parental Controls ▸ User Controls ▸ Web Restrictions
Search
File Edit View Tools Help

Which parts of the Internet can __vmware_user__ visit?

● Block some websites or content

○ Allow all websites and content

Allow and block specific websites

Edit the Allow and block list

☐ Only allow websites which are on the allow list

Block web content automatically

Choose a web restriction level:

○ High ○ Medium ○ None ● Custom

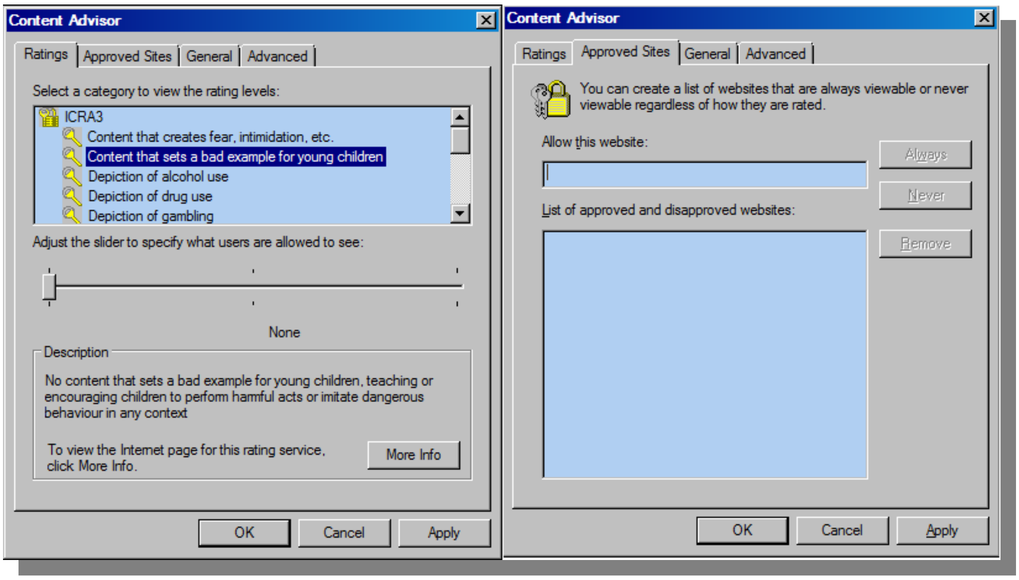
Custom level: select the content categories that you want to block.

How does the web filter work?

☐ Pornography
☐ Mature content
☐ Sex education
☐ Hate speech
☐ Bomb making
☐ Weapons
☐ Drugs
☐ Alcohol
☐ Tobacco
☐ Gambling
☐ Unratable content

Not all content in these areas can be automatically blocked.
To control access to specific websites, add the website to the Allow and block list.

Figure **4-1**. Internet Explorer Web Restrictions



Select a category to view the rating levels:

- Content that creates fear, intimidation, etc.
- Content that sets a bad example for young children
- Depiction of alcohol use
- Depiction of drug use
- Depiction of gambling

Adjust the slider to specify what users are allowed to see:

None

Description

No content that sets a bad example for young children, teaching or encouraging children to perform harmful acts or imitate dangerous behaviour in any context

To view the Internet page for this rating service, click More Info.

You can create a list of websites that are always viewable or never viewable regardless of how they are rated.

Allow this website:

List of approved and disapproved websites:

Figure **4-2**. ICRA/ White & Black list in Internet Explorer

Given so much features provided, it seems there is little chance a malware maker, spammer or phisher could possibly come close to teens; however, that is not the truth. There're at least two flaws we could tell.

First of all, as the categorization of the URLs are offered by the third-party organization, both of the correctness of their results and coverage of their inspections are sometimes questionable, besides there's a large chance that the websites the children visits are un-ratable or not filtered, as the topic of the website might not listed in any of these categories. And it is also not realistic for the parents to tailor the tool by building a huge black/white list of websites so as to meet their specific requirements, like don't want the children to see the description of some specific event or thing.

Secondly, as all the classification about URL are assigned based on the overall content on that website, there is a large chance that the classification is not sufficient to protect the children from being exposed to unwanted web contents. We call these websites as grey ones. Some portal websites, like CNN, yahoo news or news bulletin in SN, provides extensive information from latest news, short stories to movie reviews or comments on latest event, while the websites like these could be regarded by parents as a preferred source of teaching their children about out-of-class knowledge and building children's sense of the social responsibility, meanwhile, it also leaves the door to some inappropriate contents. Similar cases are those social network websites. Surveys [12] have showed that a large portion of teens who use computer and internet at their homes use online social network service quite often. Obviously, tools provided by parent control can't help in these cases, you can use the logging service to view a list of most frequently visited websites, but that will not give you much information about what your children have seen on these websites, as an alternative, you can choose to simply forbidden the access to these 'grey' websites though, but that problematic quick solution comes with a sacrifice to both the experience of children's online social life and their opportunity to get exposed to valuable information.

Besides the parent control system, there's another tool in the windows live experience product bundle that could help protect the children from online scams. In IE7, Microsoft has introduced a phishing filter system, which could protect the user from phishing attacks. It firstly analyzes websites users want to visit by checking those sites for characteristics common to phishing sites and then it sends the website address to an online service run by Microsoft to be checked immediately against a frequently updated list of reported and known phishing sites. Although it is an efficient tool to detect phish websites, it can't help filtering spammy contents in the webpage, besides, the phishing filter is triggered only if the URL is clicked by the user, this could potentially contribute to the phishing sites runners as in case they could get money merely from URL clicks.

4.2.2 Commercial Parent Control Software

Besides the tools provided by the OS provider like Microsoft and Apple, there are also bunch of third party software that could be used to help in our problem. Cyberbully Alert [41] provides a system gadget for Teens to report potential Cyber-bully event to their parents. More advanced tools are various parent control software (PCS). Commercial ones like “Parental Filter” [46] developed by Profil has more features available provided to the parent to further control children’s access to the web content. Although it provides additional functions like regulating children’s usage of their home computer or the information they could share with strangers. We only examine its ability to prevent the kids from accessing inappropriate contents published by others on webpage over the internet. To this end, they proposed three main countermeasures. First of all, as most toolbars do, PCS provides a customizable black list and white list. Parents could add or delete any URLs they believe will lead the kids to unwanted content. Second, filter by theme (hatred, violence, adult...) and semantic analysis with the possibility of adding personalized themes and expanding personal dictionaries. Third, they use active content recognition, which does an analysis using artificial intelligence techniques in real time on the contents of a Web page. User will be directed to a warning page once the result of classification is negative. Last, they find the fact that normally kids won’t visit websites hosted on sever in another country or the web-pages in another language other than the one they use, thus, if the URL the browser is opening leading to a webpage hosted in another country or is composed by the characters in another encode set, it is considered highly suspicious and the access will be forbidden. There’re many other parent control software provided in the market and all of them provide similar function stated above. In a study conducted by Top Choice in 2008 [47], 75%

products in their study have the ability to classify the webpage based on the keywords appears in the text or the URL, 63% products have ability to choose what types of content will be blocked based on various categories. 75% products refer to blacklist or white list provided either by the vendor of the software or the local customized profile.

When looking at the possible options we could made to control the filtering engine, most common ways are populating the white/black list , adding/deleting key words, or make choice among various categories pre-defined by other authorities. The first two options seem to give lots of room for the parents to tailor down the filter for their children, however, it is usually tedious to maintaining the URL list manually without assistance and it's even worse to manually selecting appropriate bad words. We argue that the words level filtering is far from sufficient as it oversimplified the problem. A case in point is that, some not-so-bad words could be used to construct a sentence that carrying unwanted the content as you might think of. The third option, using same data source same as we mentioned in the last section for Microsoft content advisor might suffer several problems as well. What the ICRA organization provides is merely a universal mark-up protocol/language for webmasters to identify their web-pages, and these additional description about the page are usually contained in the meta information section of the HTML file or sent back as part of the http response header. Once a page is labeled, it could benefit parents, search engines, filter manufacturers or any party who is concern about the content of the webpage. It is not perfect however, as stated on ICRA's web-site; trust is the primary issue to the success of their business. They carry out manual check on the correctness of applying labels and proposed several incentives to encourage the use of their labeling specification like certified website will be given higher ranks in search results. In the case of our problem, we believe labeling is not sufficient to be a solution. Although the owner of social network sites will publish some site-wide information sometimes, most content in Facebook or Myspace are composed and control by each member, or put it in another way, each member of Facebook is responsible for a website-inside-website, the problem become even complicated when the applications providing friends of the profile owner the ability to make comments on others page, like Facebook wall. In ICRA labeling specification, the only suitable descriptor you could find is in so called "User Generated Content Section" either moderated or immoderate, however, a filtering system could hardly gain any input from such neutral opinions.

Plus, in most cases, once a negative decision is made by the embedded filtering engine, PCS will redirect the user to another warning page. It might seem to be arbitrary to ban the page as a whole, if there's only a small portion of the page is inappropriate, especially in the context of comments or replies on a social network website like Facebook or Myspace, it's pretty annoying if you can't see your friends' profile merely because someone has post bad words on his page.

Another problem with PCS is that, although they claimed to provide "all-around" protection, they are obviously not be able to identify spammy contents or phishing contents in the web-page, let alone protecting the children from being misled and their AI engines work as a black box and provide no interface for customized adjustment.

4.2.3 Anti-Phishing Tool Bars

As we've seen in last section, parent control software alone is not sufficient to stop phishing attacks. To counter this problem, many users resort to various toolbars or extensions. In previous literature [48] and [32] , researches have evaluated over 10 anti-phishing tool bars. Although there are many toolbars in the market, as of September 2006, the free software download site download.com has listed 84 anti-phishing toolbars; most of them are based on similar architectures.

Among all techniques used in these toolbars, blacklisting is by far the most popular approach adopted by most of toolbars. The lists these tool bars refer to are maintained either by the vendors themselves like the Microsoft, Google and Netscape or by third party authorities like APWG. Blacklisting features efficient validation process while suffering from potential risk of submission latency and discrepancy among different list providers.

Heuristic rules are also employed in most cases. The content of suspicious URL itself is usually checked against some patterns like the appearance of special characters (@, dots, back slashes)[49][50][51] and any other tricks commonly used to obfuscate the URL. Besides, many toolbars look into auxiliary information that is related to the URL under examination. Google Safe browsing checks the page-ranking of URLs, since phishing website is usually setup for special purpose thus will not have many websites point to it. McAfee Site Advisor verifies more information related to the website the URL leads to, such as checking number of spammy emails sent from that domain, whether or not it offers downloads containing spyware, or if it once committed other malicious activities. Spoof Guard compared the URL with user's browsing history and search for malicious URLs that try to mimic the authenticated ones, it also checks various obfuscation techniques applies on URLs [51]. Last common technique is community rating, which is usually used for reputation system that aggregating ratings given for a specific site from all subscribers of the service. Cloudmark, earthlink and ebay toolbar integrate this function and use collective intelligent to benefit the community.

4.3 Summary

We have seen that there are lots of tools available there , that would probably solve some of our problems , while we argue that none of them could provide children with complete and efficient protections against online threatens like spamming, phishing, offensive contents and web scams. The reasons are listed as following:

1) Black listings/ White listings alone are not sufficient to filter inappropriate contents.

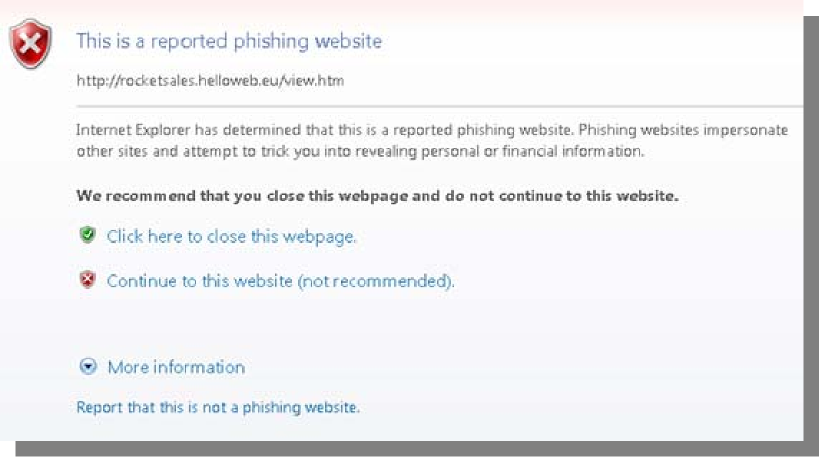
   As we have discussed in Section 4.2.2, they are not able to work in the case that users of the website could create or modify the contents of WebPages viewed by other visitors. Various new groups, bullion boards, blog and spaces, online social network websites are typical scenarios where listing technique would be in vain.

2) Labeling websites are not sufficient as well.

   First of all, labeling websites is not compulsory, which imply that benefits gain from labeling is largely restricted by number of websites attended. Secondly, both labeling and verification are still largely relay on human interference, for that matter, it is reasonable to believe that labeling result might be subjective and again the speed of verification will also be a deterministic factor of its success. Last but not least, as showed in Section 4.2.2, although ICRA [45] has provided a detailed vocabulary including 7 categories and over 50 sub-categories, it is still difficult for social network owners to effectively label their web-pages. Therefore, for all reasons stated above, relying merely on labeling system or blocking an unlabeled website blindly is not a good choice.

3) Phishing checking still needs improvement

   Most of current phishing tool bars check the URLs unless users click and open them in new windows (Figure **4-3**, Figure **4-4**). We argue that an inline checking of the URLs in current page is preferred as the original approach might contributes to the phishing website even the webpage is not showed, (this could happen if the filtering system need fetch and analyze the content of webpage the URL leads to in the background) . In addition, inline checking could give the viewer better impression of the trustfulness of current page through number of phishing URLs appearing, which is especially useful in the case that malicious party are usually cooperative and trying to fool the search engines by constructing link farms or other link-structure based camouflages.



This is a reported phishing website

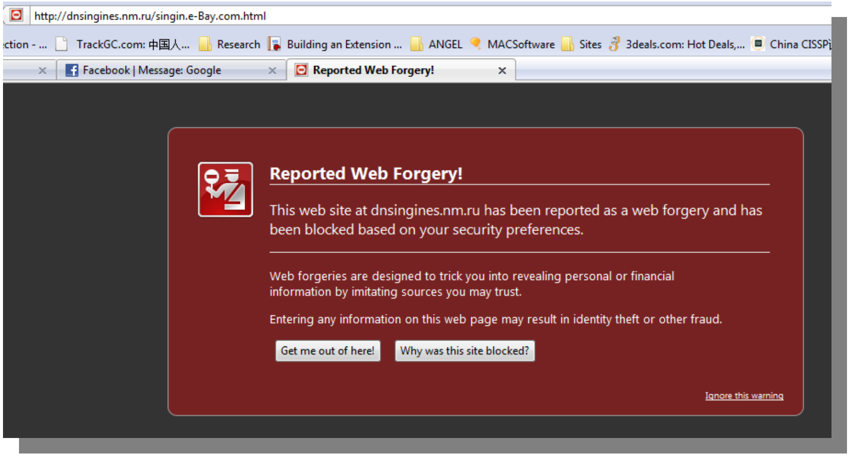
http://rocketsales.helloweb.eu/view.htm

Internet Explorer has determined that this is a reported phishing website. Phishing websites impersonate other sites and attempt to trick you into revealing personal or financial information.

We recommend that you close this webpage and do not continue to this website.

Click here to close this webpage.
Continue to this website (not recommended).
More information
Report that this is not a phishing website.

Figure **4-3**. A phishing website identified in Internet Explorer



http://dnsingines.nm.ru/singin.e-Bay.com.html

Reported Web Forgery!

This web site at dnsingines.nm.ru has been reported as a web forgery and has been blocked based on your security preferences.

Web forgeries are designed to trick you into revealing personal or financial information by imitating sources you may trust.

Entering any information on this web page may result in identity theft or other fraud.

Figure **4-4**. A phishing website identified in Firefox

4) Keywords based heuristic rules alone are not sufficient.

Some parent control software provides options for keywords based filtering, however, it is either too difficult to find the appropriate keywords or just not effective in some cases to counter variable threats.

5) Spamming messages are not considered in all tools

   Although spammy is believed to be an increasing issue in most social networking websites, nearly all tools we’ve discussed so far have not taken it into consideration.

6) Warning or blocking is not effective or flexible

   In most cases, parent control software or tool-bars will mask the whole page while providing options for the user to ignore the warnings (Figure **4-3**, Figure **4-4**), we argue this approach is problematic as users are given little information about the extent of the harmfulness of the page to be present, under such condition, most people will still choose to see the page.

7) No interaction between the protector and the protected

   For all tools we have evaluated so far, we have not seen any consideration in their designs for easing communication between the protector (parent) and the personal to be protected (children). Although at the first glance, it is regarded not as crucial as other filtering features of the software, we believe it is necessary as it helps to create a more interactive relationship between parents and teens and its existence also demonstrates parents’ respect for their children, and their care about their children's thoughts and interests. Anthropologists studying human behaviors [52][43] said that warm, kind relationship may directly or indirectly deter children’s criminal activity, illegal drug and alcohol use, negative peer pressure, delinquency, sexual promiscuity, and low self-esteem etc..

We will take above seven options into consideration when design ANGEL.

Chapter 5

ANGEL

5.1 Design Principles

The target user groups of ANGEL are either parents who need a more flexible, effective and easy to use tool to protect children's online social networking or youths who don't want to be tricked by web threats and take themselves and family into danger. The goal of the tool we design should follow four principles:

- **Usability**

  The intended user of this tool is people with little knowledge of PC or filtering knowledge. Thus, both the configuration steps and controls should be clear enough and require minimum human assistance. In addition to the basic functions a filtering tool should equipped, we also take into account the users' characteristics ,since the tool dedicated built for the specific people is most effective.

- **Extensibility and Flexibility**

  Extensibility is another key factor to the success of ANGEL, not only because the evolution of new web threats never ends, but the researches and studies in combating existing issues are also continuing, therefore, a tool with extendible interface is more desirable than the one without it. Flexibility refers to the capability that the tool should provide plenty of options when performing the task.

- **Effectiveness**

As discussed in section 4.3, state-of-art solutions are either insufficient or ineffective in providing an all-around the protection for children against threats from spam, phishing, offensive contents and web scams. Thus, when designing our tool, we will take these into account and we will explicitly evaluate ANGEL in Chapter 6.

- **Performance**

While heuristic rules are usually easy to implement, intuitive to understand and fast to execute, they also have limited capability and sometimes simply cannot meet all the requirements we need so as to tackle tough problems. More advanced algorithms, on the contrary, feature much more flexibility and could yield more powerful solutions to achieve higher accuracy and handle more difficult scenarios with sacrifice of responding time. In Section 4.3 , we have already point out that heuristic approach alone is inadequate, after introducing more advanced algorithm, we should also evaluate degeneration it brings to the overall performance.

5.2 System Design

5.2.1 Overview

We implement our tool, named ANGEL, as an extension of Firefox browser. There are three major reasons that make us choose Firefox. First is that, according to statistics provided by market share [44], 19.46% browser users have chosen Firefox as their primary tool for internet surfing in year 2008, though this percentage is still only one-third of the one for Microsoft internet explorer, however, as the Firefox is aiming to overthrown IE’s abbroachment by providing more pioneering features like tab based window and other powerful functions supplements by various add-ons, IE's leading advantage is keep shrinking. Long-term tendency analysis [44], based on data collected from Nov 2007 to Sep 2008, reveals that more and more IE refugees have become Firefox advocators. Another reason for choosing Firefox is that it is platform independent, supporting Windows, Linux and Mac system, in contrast to IE which only stick on windows platform. Besides, as one of the most popular open source projects from Mozilla, Firefox encourages enthusiastic third party developers around the world to contribute to the project so as to give web users a faster, safer and smarter tool to enjoy the joy of internet. To this end, they specify well-defined interface for the developers to either create new functionalities with extensions, or help the browser handle specific content using new plug-ins. While providing much flexible methods and hooks to access browsers' resources and services, Firefox also takes security concerns into account, therefore restrictions are made both on the way the extension could manipulate data or the scope of local system these add-ons could touch.

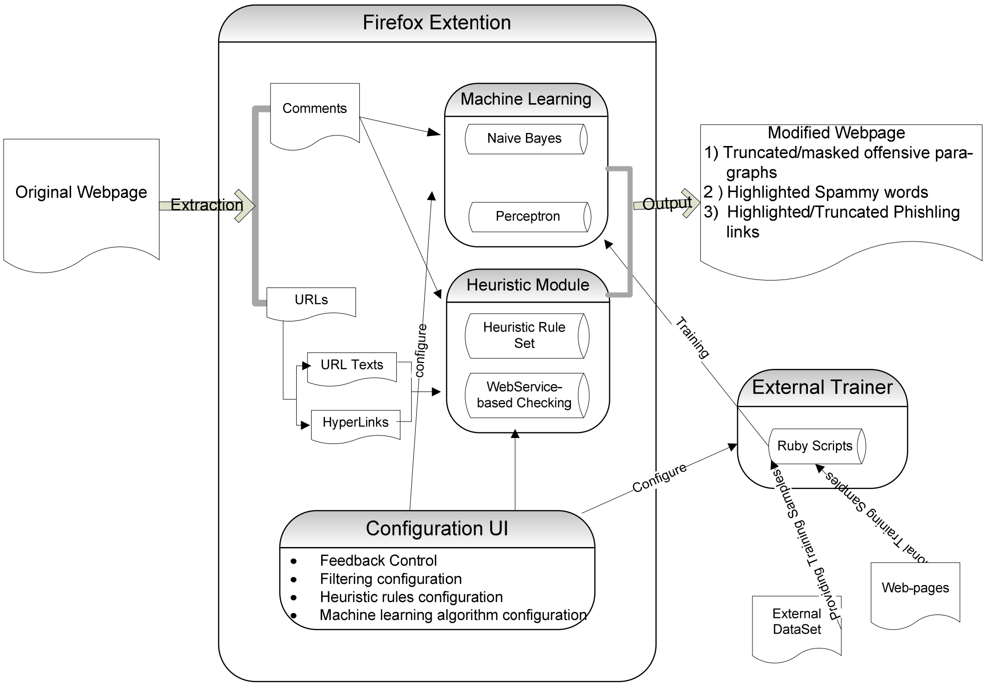


Figure **5-1:** System Architecture of ANGEL

Above figure illustrate the conceptual architecture of ANGEL we are going to built. In addition, there are two user roles in the system; one is the parent or the person who is responsible for content control, and the other is the children or the person who is being protected.

5.2.2 Module Design

As showed in Figure 5-1, ANGEL is composed of three major components, including two backend modules for training and executing the classifier and one forehand module for system configurations.

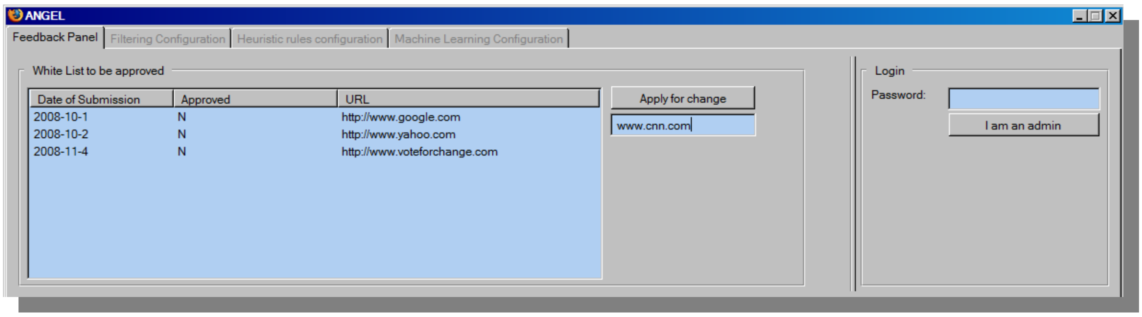
5.2.2.1 Configuration Module

Configuration module provides an easy interface for parents to control the settings of ANGEL. These settings include both features the classifier should look at during the classification and the styles users wish a classified web-page to be present. As operations on the setting panels might change the behavior of ANGEL, we add one more security layer to prevent potential manipulating attempts from unauthorized parties; this could be either from personal that is using the same computer, for instance, the children we are trying to protect, or a compromised account for remote control. We implement it as a shared secret and store it as encrypted text in local file system similar to Unix password scheme.

As the intended users of ANGEL are supposed to be unfamiliar with the advanced filtering knowledge, therefore, when designing this component, we eliminate the complicate settings procedures and reduce the configuration steps so as to make sure settings are intuitive and easy to understand.

Help documents are equipped along with all buttons and input box to provide detail description of functionality, and configuration samples are provided as well.

As showed in following Figure **5-2**, the setting panel is divided into four sub-sections for feedback control, filtering configuration, heuristic rules settings and machine learning rules settings respectively. Only Feedback configuration panel is activated until user has input the password.

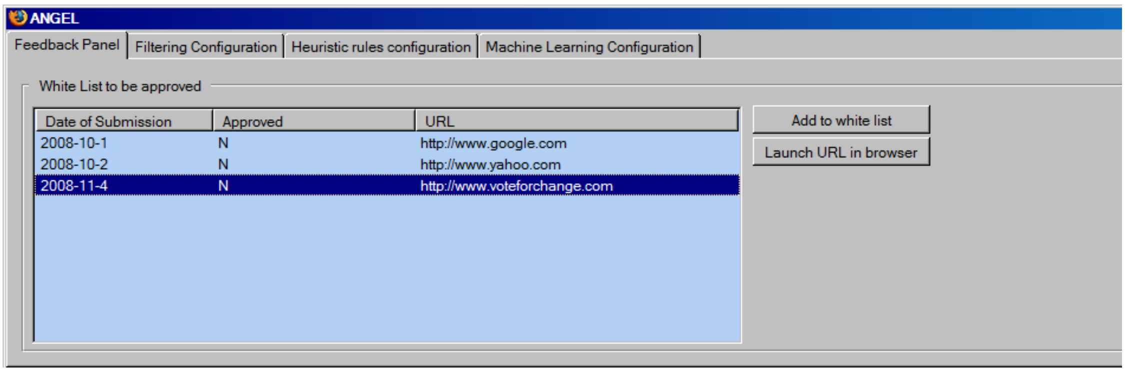


Date of Submission	Approved	URL
2008-10-1	N	http://www.google.com
2008-10-2	N	http://www.yahoo.com
2008-11-4	N	http://www.voteforchange.com

www.cnn.com

Figure **5-2:** Overview of Configuration Module

Feedback control in ANGEL provides a place for interaction between the protector and the person under protection. (Figure **5-3**) This kind of asynchronous communication eases the feedback process and provides both parties an intuitive view for tracking the request and response events. Children using ANGEL system might occasionally think the masking or filtering some specific web-page is wrong or unnecessary. Once they find such web-pages, they can submit the doubtable URLs in ANGEL system to the administrator (parents) for approval. As such feedback process is not synchronized, children could continue the web surfing after submission and launch and view the latest approved URLs easily from the feedback panel directly.



White List to be approved

Date of Submission	Approved	URL
2008-10-1	N	http://www.google.com
2008-10-2	N	http://www.yahoo.com
2008-11-4	N	http://www.voteforchange.com

Figure **5-3:** Feedback Configuration Panel

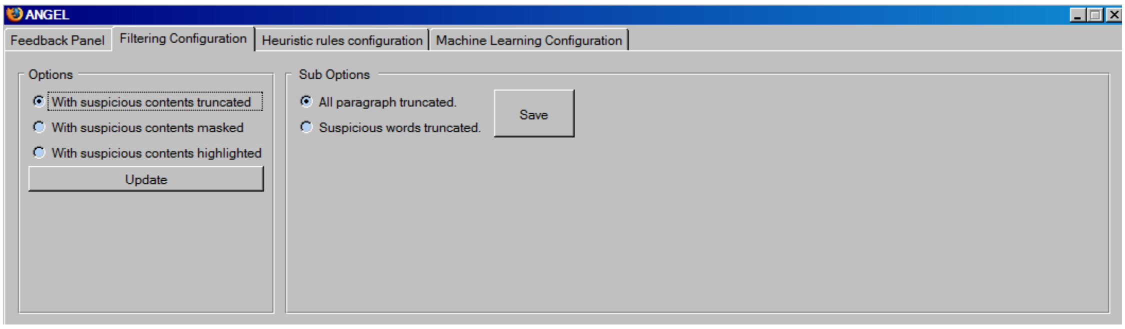
On the administrator's side, once password is input, parents could then perform approval and deny on newly submitted URLs or choose to open the URL in the browser for further examination. Another merit of this function is that as the heuristic rules and machine learning algorithms could achieve high accuracy in most cases, they might result in misclassification as well if the rules are not carefully composed or the training samples are mistakenly picked. By using feedback panel, we provide the administrator one more time to check the performance of the training process and retrain the classifier if necessary.

Filtering Configuration in ANGEL provides all around options for parents to control the appearance of filtered web-page.(Figure **5-4**) Considering the limitation of traditional feedback model, which naively block the webpage containing suspicious content as a whole, we aim to supply users with more options to filter a web-page with maximized flexibility. To this end, ANGEL offers following choices.

Filtering options

- With suspicious contents truncated
    - All paragraph truncated.

    - Only suspicious word truncated.
- With suspicious contents masked
    - Mask the suspicious contents with user-defined phrase.
    - Mask the suspicious contents with preloaded figure.
- With suspicious contents highlighted
    - Highlight the suspicious section.
    - Highlight the suspicious words.
    - Highlight the suspicious contents with different color scheme based on the confidence of decision.



Options

● With suspicious contents truncated

○ With suspicious contents masked

○ With suspicious contents highlighted

Sub Options

● All paragraph truncated.

○ Suspicious words truncated.

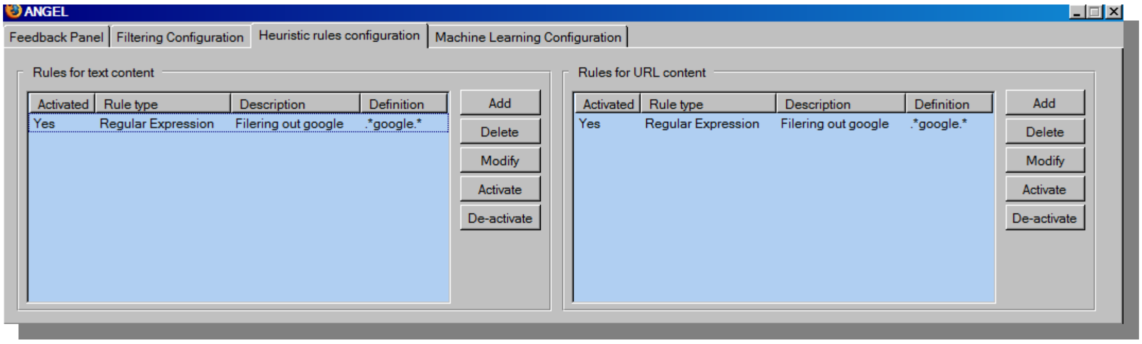
Figure **5-4:** Filtering Configuration Panel

Remember that the feedback option could apply to both the textual contents and the phish URLs found in the web-page if there is any. By choosing either truncated or masked options, children could be prevented from seeing the actual content of suspicious text while being able to view the rest benign information. Hard choices are not longer need to be made and ANGEL takes care all of that.

Truncated and masking options are intended to be used in the case that the personal we want to protect is either unable to distinguish between angle and devil like teens or personals who simply do not bother to do so. In some cases, elder children who are regard much mature than their younger brothers and sisters may prefer to make their own choices, for these users, a warning sign will be much better than truncating or masking the suspicious contents, as in the latter case, forcing them to accept the judgment might make children refuse to use the software or trying to bypass it by all means. To cater such requirement, we provide another option named highlighting mode. In such mode, it could be regarded that the ANGEL is told to run with a lowered alert threshold, in which case, the contents of suspicious content will not be modified but to be present in a different color scheme so as to make the viewers easily differentiate them from other benign texts. Furthermore, as machine learning based algorithms make predictions on the class label (inappropriate or benign) of the text to be checked based on previous knowledge, and such predictions are based on the similarity or relationship between the new incoming text and known samples in each class, thus, it is natural to conclude that , in a two class scenario, the algorithm would be more confident in assigning text paragraph under examination with class label A if it is identical to an existing sample in class A or the difference between current text paragraph and a known sample in class A is much smaller than the difference between it and any samples in class B. Such difference in confidence is normally transparent to end users in most machine learning applications but in our case, we believe it is better to leverage such information so as to provide most objective information for users to make the final decision.

Heuristic Rules Configuration

Figure **5-5** shows the configuration UI of heuristic rules setting for both text contexts and URLs appears in the web-pages. Containing two list boxes loaded with rules that have been added into the ANGEL, this panel enables parents to perform actions including viewing, addition, deletion, modification, and activation on rules. Details about different types of rules for each category and how rule are composed will be covered in Section 5.25



Rules for text content

Activated	Rule type	Description	Definition
Yes	Regular Expression	Filering out google	.*google.*

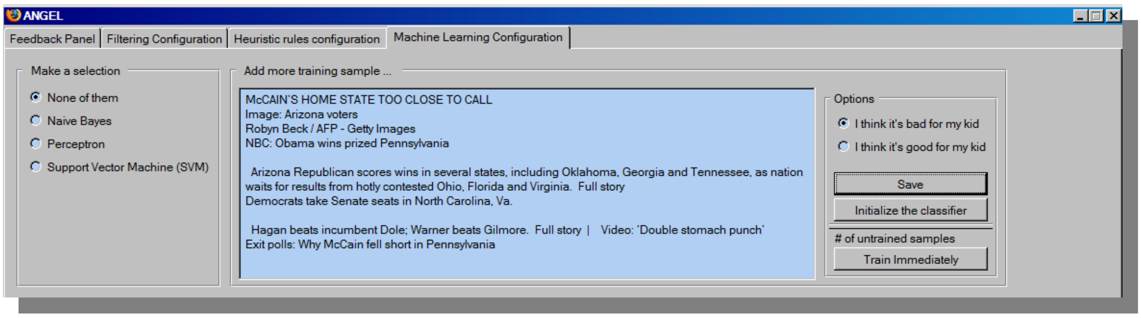
Rules for URL content

Activated	Rule type	Description	Definition
Yes	Regular Expression	Filering out google	.*google.*

Figure **5-5:** Heuristic Rules Configuration Panel

Machine Learning Configuration

We show configuration panel for activating and tweaking machine learning module in Figure **5-6**. Following the design principle of making the UI as straightforward as possible, we therefore eliminate the complex configuration of the each machine learning algorithms and provide the parent a user-friendly interface for applying most advanced techniques. As Figure **5-6** shows, on the left hand side, you can choose the algorithm the machine learning module runs among several options. A button for initialize the classifier with pre-defined dataset is enabled if the selected algorithm has never been trained before. A text region sits in the middle for collecting new training samples from web for further tuning the classifier. The user could either copy and paste any text from any external source or just drag and drop a section of texts from the webpage he is viewing in browser. A class label is required for each new sample and is set to “I think it’s bad” as default. After “save” button is pressed, content in textbox is saved into local file system for future training. Although option is provided to train the classifier immediately, it is recommended to perform the training process later since it could then train all saved samples in the batch manner and at same time relieved from impairing user experience of web surfing due to extra resources consumption.


ANGEL
Feedback Panel
Filtering Configuration
Heuristic rules configuration
Machine Learning Configuration
Add more training sample ...
Save
Initialize the classifier
# of untrained samples
Train Immediately

Make a selection

● None of them
○ Naive Bayes
○ Perceptron
○ Support Vector Machine (SVM)

McCAIN'S HOME STATE TOO CLOSE TO CALL
Image: Arizona voters
Robyn Beck / AFP - Getty Images
NBC: Obama wins prized Pennsylvania

Arizona Republican scores wins in several states, including Oklahoma, Georgia and Tennessee, as nation waits for results from hotly contested Ohio, Florida and Virginia. Full story
Democrats take Senate seats in North Carolina, Va.

Hagan beats incumbent Dole; Warner beats Gilmore. Full story | Video: 'Double stomach punch'
Exit polls: Why McCain fell short in Pennsylvania

Options

● I think it's bad for my kid
○ I think it's good for my kid

Figure **5-6:** Machine Learning Configuration Panel

5.2.2.2 Training Module

The training module is designated to be a backend module in ANGEL, designed to be flexible enough to integrate additional machine learning algorithms easily. To this end, it is relatively self-contained therefore could run separately from other parts of ANGEL. Each machine learning algorithm is implemented in ruby scripts and adheres to the same interface specification so as to make quick switch feasible. Based on current settings, different implementations will be triggered either by explicitly pushing the training button in “Machine learning configuration” tab or by scheduled tasks. After the training process, new knowledge gain from the training samples will be updated and stored in the Firefox preference file for future usage by the classification module.

5.2.2.3 Classification Module

Classification module is the other backend module in ANGEL, responsible for web-page parsing, filtering and preparing the final page to the end user. Classification module relies on both the configuration module for loading heuristic rules and training module for updating weight values of features appearing in each class. Once an inappropriate paragraph is found or a group of suspicious words are identified, the classification module manipulates the appearance of the content in original web-pages and feeds a processed page back to the client end based on the choice user has made for filtering settings.

5.2.3 Sequential Flow Diagram

Three different scenarios interacted among different modules are deliberated in the following diagram.

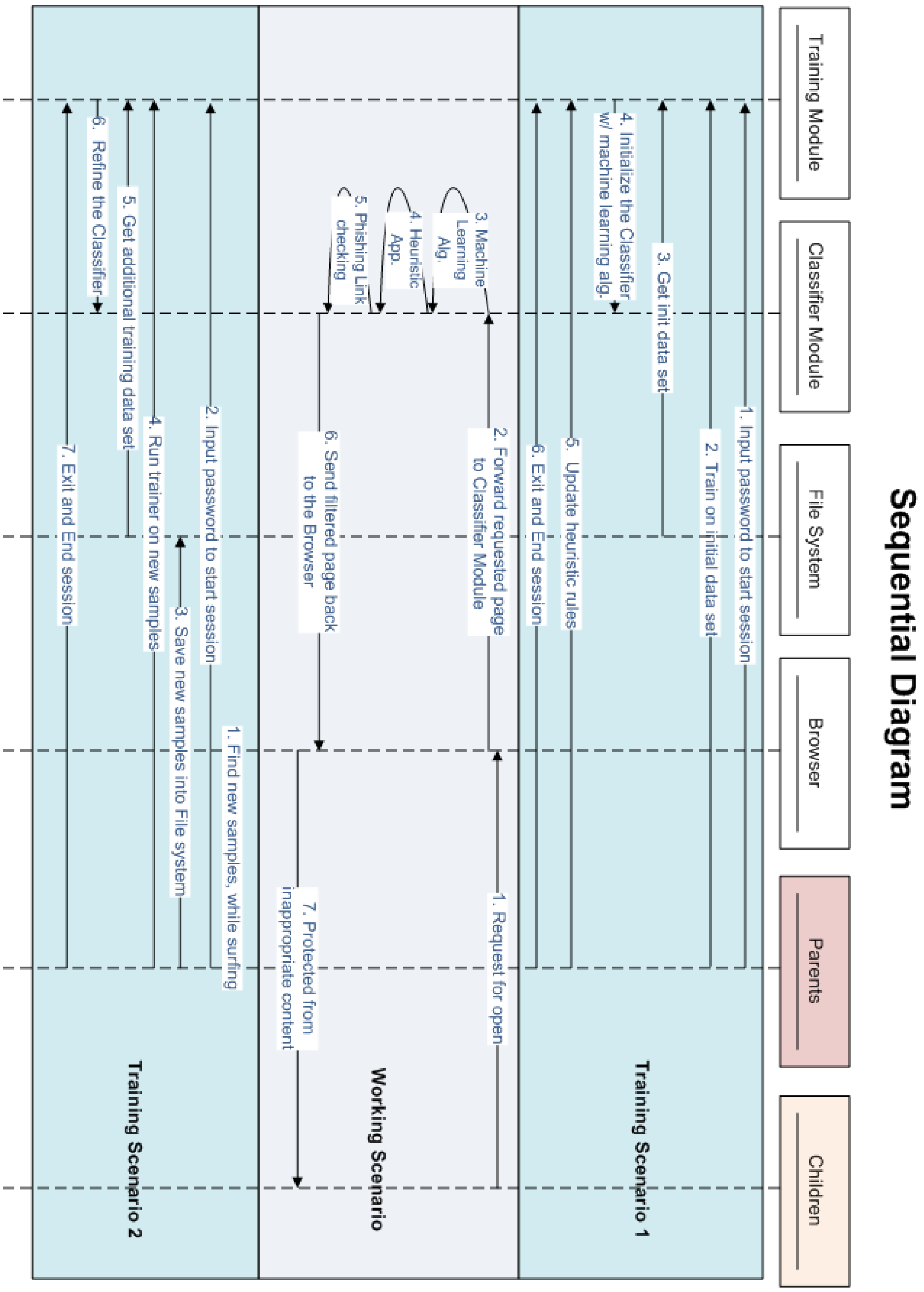

Sequential Diagram
Training Module
Classifier Module
File System
Browser
Parents
Children
Training Scenario 1
1. Input password to start session
2. Train on initial data set
3. Get init data set
4. Initialize the Classifier w/ machine learning alg.
5. Update heuristic rules
6. Exit and End session
Working Scenario
1. Request for open
2. Forward requested page to Classifier Module
3. Machine Learning Alg.
4. Heuristic App.
5. Phishing Link checking
6. Send filtered page back to the Browser
7. Protected from inappropriate content
Training Scenario 2
1. Find new samples, while surfing
2. Input password to start session
3. Save new samples into File system
4. Run trainer on new samples
5. Get additional training data set
6. Refine the Classifier
7. Exit and End session

Figure 5-7**:** Sequence Diagram

- **Training Scenario 1 (Train the classifier using external dataset)**

Although ANGEL is intended to be a personalized filter, it is preferred to be initialized so as to gain a basic ability to filter improper content at the very beginning.

Step 1: Parents must input the password first before they can make any modification on configuration.

Step 2: Once the tool has been installed, parents have the option to initialize the tool with a pre-defined dataset. This dataset can either be shipped with executables or downloaded as monthly released updates from a supportive website.

Step 3: Datasets are stored in local file system. The trainer will fetch the dataset for initializing or updating from a dedicated folder.

Step 4: Train the classifier with input data.

Step 5: Update heuristic rules as well if any changes have been made.

Step 6: Save and exit the tool.

- **Training Scenario 2 (Train the classifier using additional dataset)**

After ANGEL has been installed and the classification module has been initialized, the tool can be easily fed with new additional samples for further training afterwards.

Step 1: Parents launch the browser and start viewing the web-pages, he would like to save new samples once he think contents on some page are not suitable for his children.

Step 2: To start with, parents must input the password before they can make any modification on current configuration.

Step 3: After login, parents then can copy/paste or drag/drop the highlighted paragraph and store it in local File system for further training.

Step 4: Parents could start the training manually or let the system take care, as an automated training will be triggered by the chronographic task.

Step 5: Trainer gets additional input data saved previously in local hard drive.

Step 6: Refine the classifier with new training samples.

Step 7: Save and exit ANGEL.

- **Working Scenario (TEENS' Mode)**

Working scenario of ANGEL is also called teens' mode, in which teens' online surfing are protected from inappropriate contents, including but not limit to spams, phishings , offensive words and web scams. A typical transaction in this mode is depicted as following:

Step 1: Children open the browser and input the URL in the address bar.

Step 2: Upon receiving the request, the browser send http request to the server hosting the web-page and then forward the result page to the Classification module of ANGEL.

Step 3: The classification module first breaks the visible texts in the original HTML page into several suspicious units, and each unit is regarded as the minimum data to be fed into the next stage. To get these suspicious units, the classification module first looks into DOM tree of current page and fetches all text sections with XPATH expression. As output of previous step might contains both text contents we are looking for and unrelated contents, a second layer filtering is necessary so order to kick out HTML tags, Cascade Style Sheets sections, and all other scripts sections. After we get the suspicious units, they could then be forwarded into the machine learning sub-module for further classification.

Step 4: After passing the machine-learning sub-module, the suspicious unit could either be untouched if its content is benign or be purified if there's any text inside suspicious unit is believed to be inappropriate for children. As we mentioned earlier, although we agree ML based approaches are more powerful and flexible, these benefits also come with potential misclassifications. In order to minimize the potential risk in case the machine-learning module fails, the suspicious unit will first be checked against heuristic rules, either pre-defined with installation, or supplemented by the parents later.

Step 5: In this step, all URLs in suspicious unit will be checked to see if they are either related to listed phishings, or have high probability of being URLs leading to phishing websites.

We rely on web service hosted by phishing-tank to check the former ones. To do this, classification module first compose a SSL-based http request to the remote server and then waiting for the feedback. For the latter ones, the URL will be checked against a series of heuristic rules to see if the URL is obfuscated or trying to fool the user. Once a phishing or phishing-ish URL is identified, actions will be taken according to the filtering settings.

Step 6: The filtered page will be pushed back to the browser.

Step 7: The browser then can display the purified web-page to teens.

5.2.4 Machine-learning Algorithms

As its literally meaning indicates, machine learning techniques concern about the design and development of algorithms and approaches that allow computers to "learn" from known samples [53][54]. Although learning processes differs from algorithm to algorithm, almost all of them relate to building models upon features extracted from the input data. As the outcome of the learning process, a classifier could then be generated, and used to “group” the unknown data into one of the given categories or classes with the help of models created in the previous step. The accuracy of classifier gains with the increasing learning experience. Applications of machine learning techniques are found in many realms, include natural language processing, syntactic pattern recognition, search engines, medical diagnosis, bioinformatics, detecting credit card fraud, stock market analysis, classifying DNA sequences, speech and handwriting recognition, object recognition in computer vision, game playing and robot locomotion etc. [54]

Beside these, researchers in computer security are also impressed by the detection power of machine learning techniques. Common applications are in the field of solving spamming and phishing problems, as both of them could be transformed into an equivalent text mining problem. In [55], Konstantin presented applying Bayesian classifier, k-nearest neighbours, neural networks and support vector machine(SVM) in the problem of spam filtering. Pranam,Akshay,Tim etc.[56] applied SVM in detecting spammy activity in Blogosphere. Saeed, Dario etc.[57] compared six machine learning algorithms in identifying phishing emails. In addition, machine learning also prevails in the area of intrusion detection studies in both network based and standalone systems. Introduced by Anderson [58] and formalized by Denning[59], machine learning based anomaly detection approach has been used in attempts to identify novel attack behaviors. In,J. Zico Kolter and Marcus A. Maloof [60] applied five text classification algorithms, instance-based learner, Tf-idf classifier, naive bayes, SVM and decision tree models, into the problem of detecting malicious executables in operation systems. In [61], James and Carla extracted features from server flow behavior and apply decision tree to identify potential network anomaly.

In ANGEL, we choose to integrate Naive Byes and Perceptron algorithm for their effectiveness and simplicity of implement. In following sections, we will introduce both of these algorithms in detail.

5.2.4.1 Naive Bayes

Naive Bayes is a probability based method that has extensive application in both texting classification and spam filtering applications.[62][63][64][65] It computes probability P (c|d) as concept description of P(c), the prior probability of each class and P (t |c) , the conditional probability of each feature given the class.

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

Depending on the model selected, P(t|c) could be derived from either the frequency of occurrence of this feature in each class or the total number of the samples in each class containing a specific feature. Then, assuming each feature is independent, that is, occurrence of one feature has nothing to do with the occurrence of any other feature, it applies Bayes' rule to compute the posterior probability P(c|d) of each class given an unknown instance formed by a collection of features. A class label is thereby assigned to the instance under classification with the label of the class that outcomes the maximum value of P(c|d)

$$\underset{c \in \mathbb{C}}{\arg\max}\, \hat{P}(c|d) = \underset{c \in \mathbb{C}}{\arg\max}\, \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

Simple and intuitive as it is, Naive Bayes algorithm also features fast training and classifying speed, and small storage space. Furthermore, it can achieve sound accuracy even in the case that the independent assumption on features is not hold and it is ease to be implemented as online training mode as well. All these merits assure our choice of this algorithm.

In [66] , authors explored several variance of Naive Bayesian, or in another word, different modeling of same feature. Multi-variate Bernoulli only considers each features as a Boolean, which could either be 1 if it appears or 0 if it does not. Multinomial form of NB takes into account more information like the frequency of a specific feature appears, which catches the characteristic that some feature would occur more frequently in one class than in others. Their study result also shows that Multinomial version outperforms Multi-variate Bernoulli implementation on the Enron-Spam corpus. For this reason, we implement the former version of NB algorithm in ANGEL.

5.2.4.2 Percepetron Algorithm

Different from the Naive Bayes classifier discussed above, which picks the class label for the incoming instance with maximizing the probability of being in such class given the instance, Percpetron learning algorithm outputs a linear classifier that would construct a decision boundary to explicitly separate the data into different classes.(Figure **5-8**) It was invented by Frank Rosenblatt in 1957 [67] and has occupied an important place in the history of pattern recognition algorithms.

Basic idea of Perceptron algorithm is to construct an appropriate hyperplane that could set the data apart. A hyper-plane or affine set L could be defined by the linear equation $Y = W^T\Phi(x) +$

b, in which $\Phi(x)$ is given by the features appear in the current instance and W^T is a weight vector that controls the degree of slew of the hyper-plane and b is a constant bias. The goal of Perceptron training process is to find a set of proper W^T and b with $\Phi(x)$ observed in training sets so as to minimize the distance of misclassified points to the decision boundary. Once found, the equation could then be used in future to classify new coming instances. Stochastic gradient descent is used to minimize the piecewise linear criterion, which models the distance by $-\sum Y_i (W^T\Phi(x)_i + b)$ [68], in which i is the set of misclassified points. The training process continues until the Perceptron algorithm converges and manages to correctly classify all samples in training set.

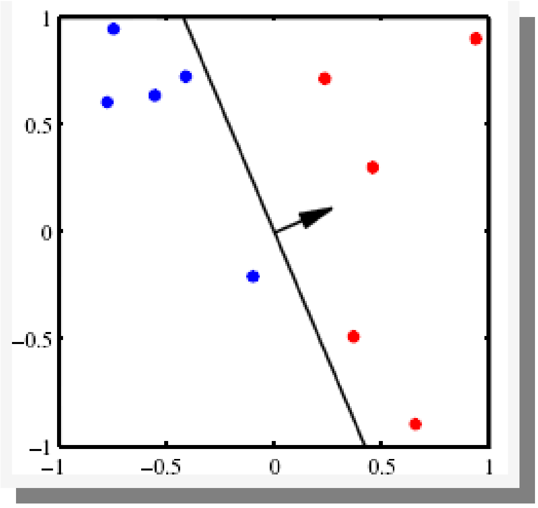


Figure 5-8: Perceptron Algorithm

Roseblatt proved in [67] that Perceptron algorithm is guaranteed to find an exact solution in finite number of iterations as long as the data set is linearly separable. However, if the data set is too large, and its separableness is unknown in advance, then it could be hard for trainer to decide if the algorithm is running towards the converge or stepping around but never ends. Another potential concern with Perceptron algorithm is that there are usually more than one unique solutions there to classify the data set if it is linearly separable, and which one to be discovered by Perceptron is largely determined by the initialization of the parameters and the sequence of appearances of the data points [68]. In some cases, support vector machine is preferred to Perceptron as it could help identify the decision boundary which maximizing the margin to nearest data points in each class. However, that does not come for free as it generally requires much longer training time and considerable computational cost.

Easily implemented as online classifier makes Perceptron algorithm and its variance popular candidates for solving problems in different realm, such as spam filtering [69] and computer user recognition.[70] etc. We implement it as one of options in ANGEL's machine learning module.

5.2.4.3 Feature Selection

In Linguistics, features are properties of a class of items which describes individual members of that class. In our case, they could be vocabulary based like words or phrases in the webpage or non-vocabulary based like number of occurrences of a specific term in the text paragraph. In last few sections, we have mentioned that both Naive bayis and Perceptron have been successfully applied in the problem of text classification, in most cases, before we can actually apply the machine learning algorithm to the training data, we need go through a feature selection process to reduce the number of features. It is regarded as an important preprocessing step and can be independent from learning. Feature selection is especially useful in the case that number of features is large but you suspect that there is only a small portion of features that are "relevant" or could contribute to the learning task. There are many potential benefits of variable and feature selection. As the number of feature to investigate is reduced, both the time for training, the space for storage, the resource consumed could be effectively decreased. In the case of Perceptron algorithm that we applied in ANGEL, feature selection could help initiate an efficient feature vector with high accuracy and thus control the size of feature vector as Perceptron only retrains, or in another word, increases the size of feature vector upon misclassification. Eliminating some specific features also helps to decline the chance of over-fitting problem which enhance the generality of the classifier as a result. In terms of ANGEL, the tool that we intended to build is an extension to existing web browser, it should not have a long response time otherwise the user experience of web surfing will be undoubtedly degenerated, thus we believe feature selection is necessary in our case.

In general, feature selection fall into three categories [71], namely filters, wrappers and embedded methods. A filter method computes a score for each feature as a measurement of associative and then selects features according to the rank of scores [72]. Information gain (IG) , chi-square (CHI^2), Mutual information (MI) are three most effective methods in this category. The second category referred to as wrapper [73] utilizes the learning system as a black box to score subsets of features, and the third category called the embedded method [71] performs feature selection within the process of training.

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t) P(C = e_c)} \tag{MI}$$

$$X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \tag{CHI^2}$$

Filter method features easy implementation and computationally much cheaper among all available approaches. As our problem in this case could be transformed into a text classification problem, where the possible feature sets, including the vocabulary and other auxiliary features (non-vocabulary-based) are considerable large, we then decide to use filtering methods to rank all possible features first, then use cross validation to conclude a proper size of features.

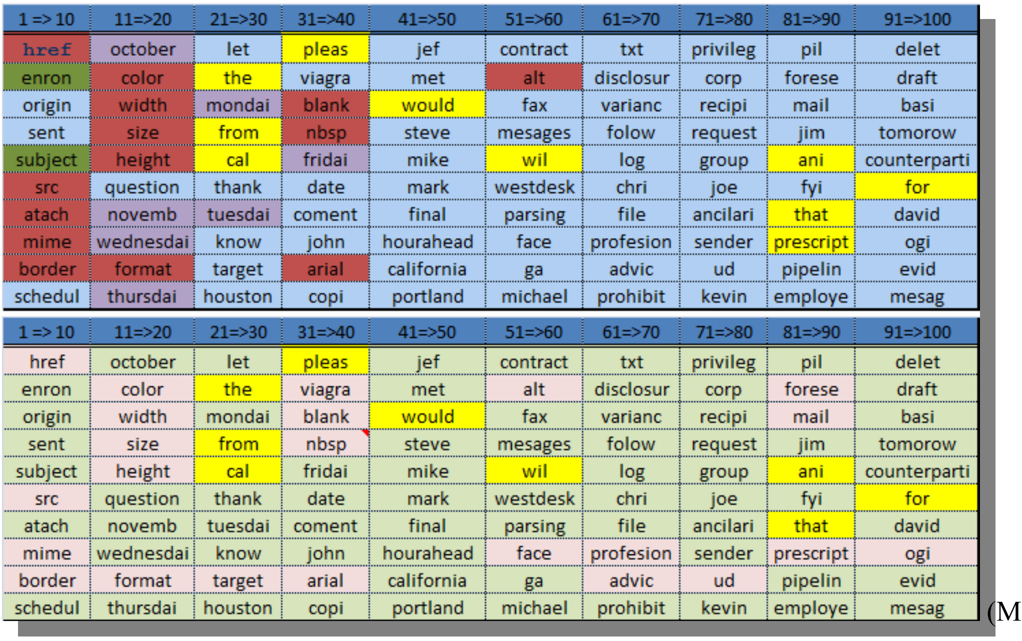
We use both mutual information and CHI^2 formula (showed above) to calculate a score for each feature in samples of TREC SPAM 2005 corpus (39399 ham samples, and 52790 spam samples, only email headers are discarded during selection process.) In below formulas where N_{ij} appears, i equals to 1 if the feature appears in spam and j equals to 1 if feature appears in ham. Table **5-1** shows 100 features with highest scores for both methods, all features/words are in their root form since they have been passed through the porter stemmer before ranking.

$$\begin{aligned} I(U;C) &= \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} \\ &\quad + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}} \end{aligned}$$

(MI)

$$X^2(\mathbb{D},t,c) = \frac{(N_{11}+N_{10}+N_{01}+N_{00})\times(N_{11}N_{00}-N_{10}N_{01})^2}{(N_{11}+N_{01})\times(N_{11}+N_{10})\times(N_{10}+N_{00})\times(N_{01}+N_{00})}$$

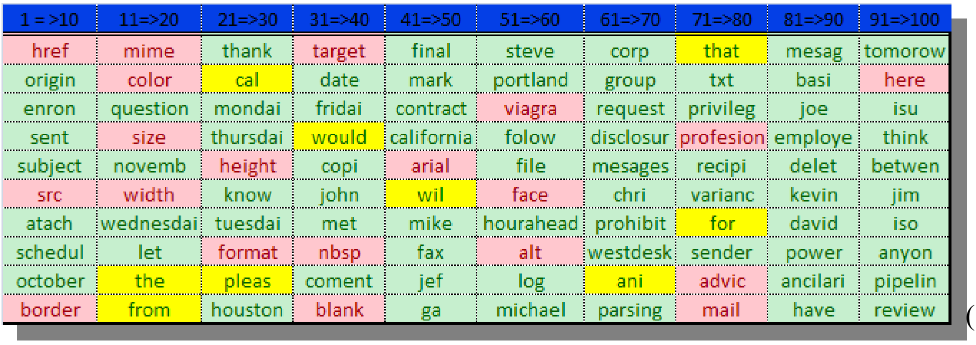
(CHI^2)



1 => 10	11=>20	21=>30	31=>40	41=>50	51=>60	61=>70	71=>80	81=>90	91=>100
href	october	let	pleas	jef	contract	txt	privileg	pil	delet
enron	color	the	viagra	met	alt	disclosur	corp	forese	draft
origin	width	mondai	blank	would	fax	varianc	recipi	mail	basi
sent	size	from	nbsp	steve	mesages	folow	request	jim	tomorow
subject	height	cal	fridai	mike	wil	log	group	ani	counterparti
src	question	thank	date	mark	westdesk	chri	joe	fyi	for
atach	novemb	tuesdai	coment	final	parsing	file	ancilari	that	david
mime	wednesdai	know	john	hourahead	face	profesion	sender	prescript	ogi
border	format	target	arial	california	ga	advic	ud	pipelin	evid
schedul	thursdai	houston	copi	portland	michael	prohibit	kevin	employe	mesag

1 => 10	11=>20	21=>30	31=>40	41=>50	51=>60	61=>70	71=>80	81=>90	91=>100
href	october	let	pleas	jef	contract	txt	privileg	pil	delet
enron	color	the	viagra	met	alt	disclosur	corp	forese	draft
origin	width	mondai	blank	would	fax	varianc	recipi	mail	basi
sent	size	from	nbsp	steve	mesages	folow	request	jim	tomorow
subject	height	cal	fridai	mike	wil	log	group	ani	counterparti
src	question	thank	date	mark	westdesk	chri	joe	fyi	for
atach	novemb	tuesdai	coment	final	parsing	file	ancilari	that	david
mime	wednesdai	know	john	hourahead	face	profesion	sender	prescript	ogi
border	format	target	arial	california	ga	advic	ud	pipelin	evid
schedul	thursdai	houston	copi	portland	michael	prohibit	kevin	employe	mesag

(MI)



1 = >10	11=>20	21=>30	31=>40	41=>50	51=>60	61=>70	71=>80	81=>90	91=>100
href	mime	thank	target	final	steve	corp	that	mesag	tomorow
origin	color	cal	date	mark	portland	group	txt	basi	here
enron	question	mondai	fridai	contract	viagra	request	privileg	joe	isu
sent	size	thursdai	would	california	folow	disclosur	profesion	employe	think
subject	novemb	height	copi	arial	file	mesages	recipi	delet	betwen
src	width	know	john	wil	face	chri	varianc	kevin	jim
atach	wednesdai	tuesdai	met	mike	hourahead	prohibit	for	david	iso
schedul	let	format	nbsp	fax	alt	westdesk	sender	power	anyon
october	the	pleas	coment	jef	log	ani	advic	ancilari	pipelin
border	from	houston	blank	ga	michael	parsing	mail	have	review

(CHI^2)

Table **5-1** Feature Selection with Mutual Information and CHI^2

Compare the result of CHI^2 and MI, we do not see too much difference, as most words appear in MI's table also appear in CHI^2 's table. In the MI table, each cell in the lower part of the figure indicates the true class label of the feature. Pink says this feature appears more often in spammy (bad) samples while the green says the opposite. For those cells in yellow, they are marked as stop words, which are usually believed to have equal probability to appear in samples of any class. In total, there are 23 spammy features, 9 neutral features and the rest are all ham features. Given the rank, shall we just pick top K candidates with highest score? The answer is NO. Reason is that we also should take the discrepancy between data in verification and training process into account. Look into the upper part of the above figure, we find that all cells in dark red are HTML tags, and all of them are believe to be good indicators of bad samples in TREC corpus. This is conceivable as in the case of spam email, an email full of html contents are more likely to be commercial ads or phishing messages that wish to catch readers' eyes. However, in ANGEL, the target of classification task is the webpage itself, in which case HTML tags are reasonable to be treated as neutral symbols. Thus, including these tokens might never help in our case, though we could deduce a non-vocabulary feature that a text paragraph with too many URLs (deduced from the 'src' tag) is more likely to be a bad instance. Other interesting findings are that, the cells in purple are words that all relate to day or month; words in green are related to the company the data set is taken from; We also find several names selected by mutual information as good indicators of good samples, we believe that could be names of secretary or manager in one company that usually distributes lots of ham emails. We keep these in mind while selecting features for machine learning module of ANGEL.

5.2.4.4 Summary and Discussion

We have seen in last few sections that machine-learning based sub-module can grant ANGEL the ability to identify inappropriate contents in suspicious paragraphs with high accuracy even if they have never been observed before. Feature selection further helps these algorithms to run more efficiently and only focus on valuable information in data samples. While algorithm-based approach is generally considered superior in terms of their decent compatibility and accuracy in detecting variance of improper contents only using knowledge gain from previous training data, they are also believed to be more computationally expensive than other alternatives, as before you can start verifying anything, you have to train the classifier first, this is not a easy task tough, since not only the training process itself usually takes much longer time than the verification process, (Remember we use asynchronies training and verification process, so as to reduce this latency.) but the way you train the classifier is also deterministic to its final performance. Another concern is that, even if you have chosen the appropriate training sequence and take adequate time to perform the training; the classifier might also suffer from misclassification, which either gives you many false positive or false negative. Then you might need to look at the training data you've picked for trainer. What if the training data cannot correctly catch the features or characteristics of the data you are looking are? (an example of drifting) As the theoretical background of all kinds of machine-learning algorithms is statistics laws, therefore it naturally implies that the distribution of potential output of the target you are estimating for should strictly follow the same pattern or have same statistic characteristic as the counterpart of your training sample. Otherwise, it will for sure generate misclassifications.

In [74], David J. Hand studied this problem and pointed out that untruthfulness of pre-assumptions in the training process will largely influence the result of analysis and in the end lead to incorrect conclusion.

Besides the inconsistency between the distribution of training data and data during classification process as we've discussed earlier, he also discovered other two cases which might deviate the classification result.

First is the error-classified training set. As machine-learning algorithms are using prior belief of the training data to make estimation of the potential class label of new incoming data, it is no doubt that some of error-classified training samples will impair the accuracy of the prediction. As the training data is more than often manually selected or verified by some individual, it is not surprising that mislabeling might sometimes happens and pollutes the data set. Researchers have noticed this problem and proposed to use cross checking or simply using a published training set to minimize the impact caused by the mislabeling. We will also take this into account when we collecting the training samples for machine learning module.

Second case is caused by sample selection bias. Compared to the former case, this time, the class label for these samples are assigned correctly, however, the problem arises from the bias of the features in these samples hold and you are falsely expecting the features used to train the classifier will also be hold by all other data belong to this class while in reality , they are not. It might due to the limited size of the sampling data or the way to picking samples among all the candidates.

As both problems are proved to be difficult to find a perfect solution, we realize that an auxiliary heuristic sub-module is required to further strengthen the power of our classifier.

5.2.5 Auxiliary Heuristic Rules

In this section, we look at another sub-module in the ANGEL. Although heuristic rules are considered less powerful to its machine-learning counterpart, they features more intuitive classification process, faster verification process and giving more control to the user to make direct influence on the classification process.

5.2.5.1 Rules for Textual Content

By looking at the four inappropriate contents we have defined, which are spamming, phishing, offensive contents and various web scams, we find out that it is comparably much easier to extract rules from spamming texts, offensive contents and some of web scams. The reason is that phishing contents are most likely assembles the truth of the fact as it hopes you to be fooled and thereby falsely submit credentials to the Phisher.

There is already a long history of using heuristic rules to detect spam messages in the field of spam email detection and many other prevailing filters like Spamassassin have applied various rules into their implementation.

While those precious knowledge and experience gain from spam emails filtering could also be used to identify the spammy texts appear in the web-page, we should also realize that a major difference between these two problems, as there is no similar header information in the web-page. Information contains in the header section of emails provides important hints like the source and number of copies to the filters.

Offensive content might be the most appropriate case where heuristic rules should apply, as the number of such “bad” words and their variances (e.g. past tense and participles or plural form) in English language is finite. Recap that the definition of offensive contents in our study is gutter language, jargon, argot, sexually explicit material, racist, graphic violence, or other content that may be considered offensive on social, religious, cultural or moral grounds. While considering that making a full list of all possible words or phrases defined from scratch is tedious and time consuming, we then turn to other possible resources which might be good candidates as bootstrap for our application. Entries in English dictionary labeled as slang, offensive, or vulgar are first data set comes into our mind. Besides, we also find many lists collected and published online by different organizations and groups. [75] provides a list of bad words and slangs sorted alphabetically and published on the internet for public usage. [76] also provide a long list of ethnic slurs which generally considered being insinuations or allegation if used in the conversation. We also consider using scripts from movies labeled as R or NC-17, although it is not an easy task to extract specific words from the script body.

In the case of Web scams, as mentioned in section 3.5, are originally distributed in the emails systems. Thus, heuristic rules used to applied in filtering scam emails should also suitable in identifying web scams appears in web-pages.

We then categorize heuristic rules into the following:

Keyword based

Keyword matching is by far the most naive heuristic rules we could used. It can only be used to check the appearance of single words. Some words considered to be extreme “bad” or “offensive” could be applied in such kind rules.

Regular expression based

Regular expression are much powerful and flexible compared to pure keyword based matching as it could verify the appearance of the correlation of multiple words, under such scenario, the minimal unit we want to check in the suspicious paragraph is pairs or more words as a whole instead of examining them respectively.

Function based rules

Sometimes keyword based rules or regular expression based rules are still not sufficient to catch some complicated scenario, for instance, if we want to count the number of some words or characters and compare the result to a pre-defined threshold. To this end, ANGEL also provides advanced user with the choice of adding user-defined functions without changing the main body of the code. Such extendibility guarantees that ANGEL could be tailored to meet the requirements of different users to the maximum extent.

5.2.5.2 Rules for Phishing URLs

In [49][50][51], researchers has identified several characteristic of Phishings, which could be transformed into heuristic rules and fed into ANGEL. As only we interested in the features concerns Phishing URLs, we therefore only apply part of all observed features. Suspicious URLs could fall into at least of one of following patterns:

- Contains @ or –
- Contains IP address

  Eg. http://210.80.154.30/~test3/.signin.ebay.com/ebayisapidllsignin.html

- Have more than 5 dots in URLs, or long host names

  Eg. http://21photo.cn/https://cgi3.ca.ebay.com/eBayISAPI.dllSignIn.php

- Have non-matching URLs

  Eg. <a href=”a.com”>b.com</a>

- Have embedded domains (explicit redirecting)

  Eg.http://www.topsearch10.com/search.php?aid=59731&q=bad+credit+auto+loan";http://bad - credit - auto - - loan.blogspot.com/

- Have misspelled domain names of well-known organizations

  Eg. www.faceboook.com / www.goooogle.com / http://www.yah0o.com/

- Belongs to different geographic location (need web service support)
- Have short domain lifetime (need web service support)

In ANGEL, we define some above pattern as pre-load rules and also provide interface for adding new rules. Plus, we create a rule based on the bad words list, as it could catch most of the offensive textual contents on the web.

5.3 Third party Resources Used In ANGEL

5.3.1 Auxiliary Techniques Applied in Machine Learning Module

As filtering text paragraphs with machine learning algorithms falls into the realm of natural language processing (NLP) , therefore stemmer and stop word list, two common tools for NLP problems come into our eyes.

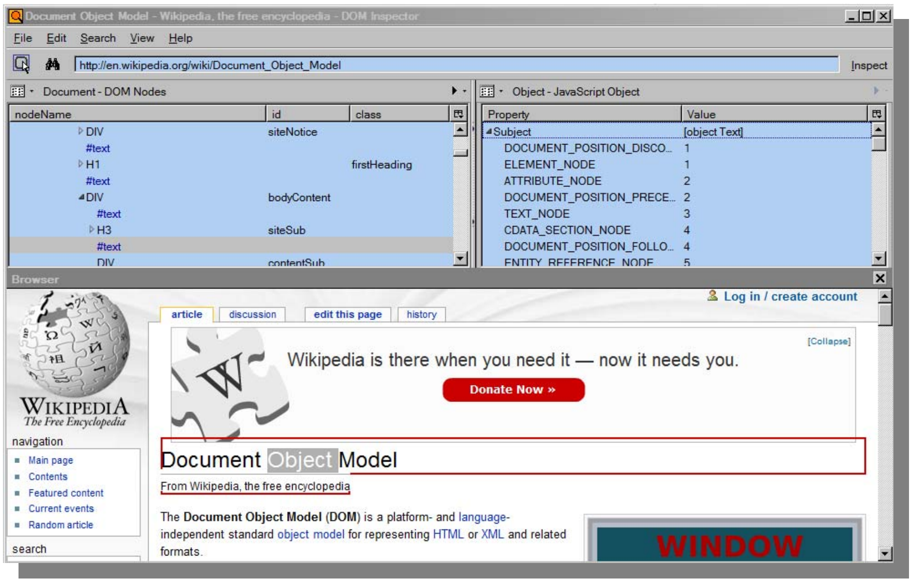
A stemmer is one kind of algorithms which restore a given word into a simplified form. Formal definition of stemmer in linguistics is that it is a morphological analyzer that associates variants of the same term with its root form [82]. A case in point is that if you pass words ‘watching’, ‘watched’, ‘watches’ and ‘watch’ into stemmer, it yields ‘watches’ as the result for four times. There are two kinds of morphological processes: inflectional and derivational [82], inflectional morphology expresses syntactic relations between words of the same part of speech, while derivational morphology expresses lexical relations between words that can be different parts of speech. In ANGEL, we employed a stemmer based on Porter stemming algorithm [83], a very popular tool for English stemming. For every words appears in the suspicious paragraph, we pass them into stemmer first and thereby normalize them before feeding them into the machine learning module. Besides the standard process, we also perform replacements such as substituting symbol ‘@’ with letter ‘a’, symbol ‘|’ or number ‘1’ with letter ‘l’, number ‘0’ with letter ‘o’ and only using lower case form of the word. This stemming process, helps the machine learning module to better identify words with their origin meanings and hence tolerant the variation of word to some extent.

A ‘stop word’ list helps the machine learning algorithm to select more valuable words or features and reduce the number of words or feature in consideration. Stop words are sometimes called function words [82] by linguistics, consisting mostly of a relatively small class of articles (‘the’,’a’,’an’,’this’,’that’,etc.) , prepositions (‘at’,’by’,’for’,’from’,’of’,etc.) pronouns (‘he’,’she’,’it’,’them’,etc.) and verbs and verb particles (‘am’,’is’,’be’,’was’,etc) . However, in some cases, using stop word list might have bad impact on the learning process, as some abbreviations of organizations or technical terms share the same form of these stop words. For example, information tech technology (IT) vs it, World Health Organization (WHO) vs who, etc.) In ANGEL, we define the contents to be filtered are more often conversations and non-technical terms used in social networking environment, therefore, using stop word list should have benign influence in our cease. For that matter, ANGEL employs a stop word list of 319 entries which contains most of common stop words in literature.

5.3.2 Grease Monkey JavaScript Injector

ANGEL's feedback module relies on Grease Monkey [80] to manipulate the webpage to be present to users. As a popular extension for Firefox browser, Grease Monkey allows you to customize the way a webpage displays on-the-fly by injecting JavaScripts into the original page. With pre-defined primitives specially defined in Grease Monkey, user scripts are not only allowed to access the resource of Firefox browser, e.g. user's preference setting, but they also can change the appearance of web-pages by adding, modifying or deleting html elements in a tree-like Document Object Model (DOM) of current page. Once the scripts have been saved on local file system, these changes made to the web pages are executed every time the pages are opened. In addition, as each user script is a separate file independent from the Grease Monkey, the injector itself, therefore, scripts could be easily released, deployed, and shared among different users as long as Grease Monkey has been preloaded.

In ANGEL, once a suspicious URL or offensive content has been identified, either through heuristic rules, web service checking or machine learning based module, ANGEL's filtering module, implemented as user script of Grease Monkey, is called and makes changes on current page or DOM according to the filtering configurations. As defined by W3C [88], DOM is a platform- and language-independent standard object model for representing HTML or XML and related formats. As showed in Figure **5-9**, it is usually in the form of a tree. DOM is the key structure that user scripts, written in JavaScript, rely on to either inspect or modify a web page.



http://en.wikipedia.org/wiki/Document_Object_Model

nodeName	id	class
DIV	siteNotice	
#text		
H1		firstHeading
#text		
DIV	bodyContent	
#text		
H3	siteSub	
#text		
DIV	contentSub	

Property	Value
Subject	[object Text]
DOCUMENT_POSITION_DISCO...	1
ELEMENT_NODE	1
ATTRIBUTE_NODE	2
DOCUMENT_POSITION_PRECE...	2
TEXT_NODE	3
CDATA_SECTION_NODE	4
DOCUMENT_POSITION_FOLLO...	4
ENTITY_REFERENCE_NODE	5

Wikipedia is there when you need it — now it needs you.

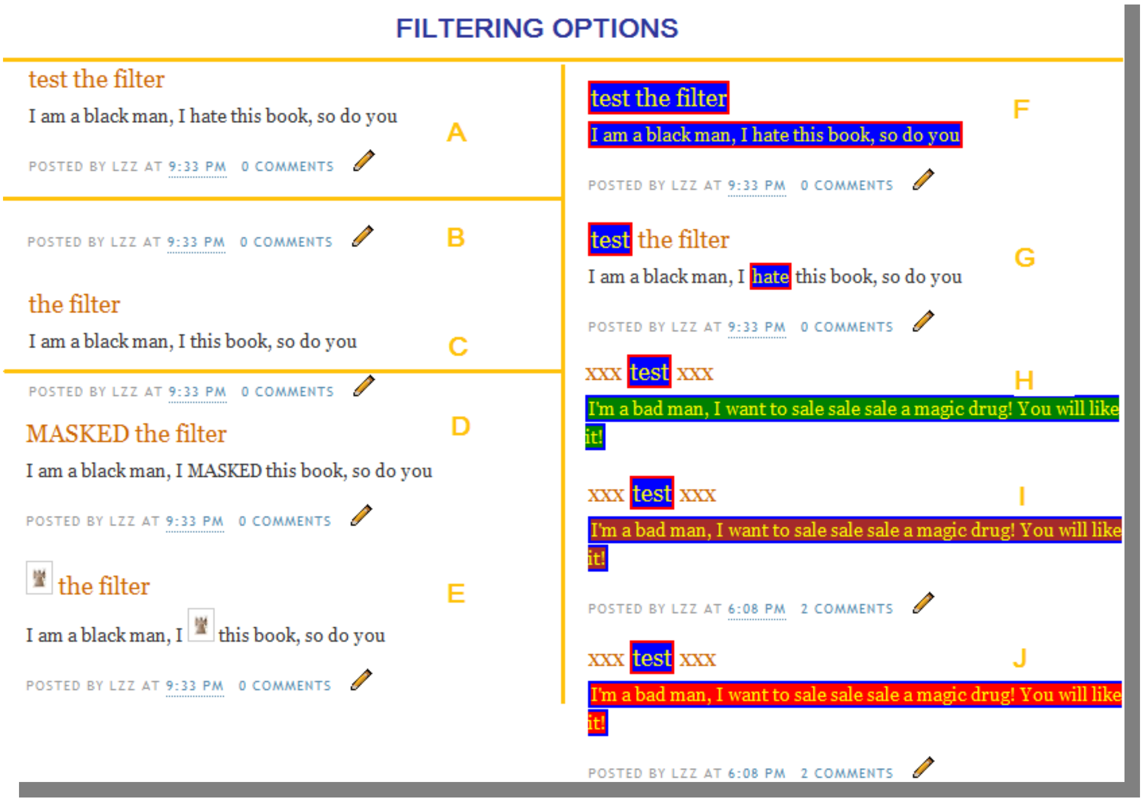
Document Object Model

From Wikipedia, the free encyclopedia

The **Document Object Model** (**DOM**) is a platform- and language-independent standard object model for representing HTML or XML and related formats.

Figure **5-9**: An Example of DOM Tree

In ANGEL, as we need to inspect both the visible contents, e.g. text paragraphs and URL strings and invisible contents, e.g. hyperlinks, so we use different XPath [88] expressions to locate appropriate nodes in the DOM tree. For truncated options, ANGEL either simply deletes the DOM node containing the suspicious content as a whole or only deletes the suspicious word while at same time creating two more DOM nodes to hold the contents before and after that word. For masking options, ANGEL substitutes the suspicious content with the word or phrase defined in user's preference setting or with a pre-loaded image of baby ANGEL. As discussed in Section 5.2.2.1, different from previous two options, the highlighting options do not remove or mask the suspicious content. Instead it provides the viewers additional information to help make the judgment himself. To this end, once locates the DOM node containing the value we intend to highlight, ANGEL just change styles of its appearance with pre-defined color schemes or choosing different color schemes according to the valve scores set for the machine learning based filtering.



FILTERING OPTIONS

test the filter

I am a black man, I hate this book, so do you

POSTED BY LZZ AT 9:33 PM 0 COMMENTS

POSTED BY LZZ AT 9:33 PM 0 COMMENTS

the filter

I am a black man, I this book, so do you

POSTED BY LZZ AT 9:33 PM 0 COMMENTS

MASKED the filter

I am a black man, I MASKED this book, so do you

POSTED BY LZZ AT 9:33 PM 0 COMMENTS

the filter

I am a black man, I this book, so do you

POSTED BY LZZ AT 9:33 PM 0 COMMENTS

test the filter

I am a black man, I hate this book, so do you

POSTED BY LZZ AT 9:33 PM 0 COMMENTS

test the filter

I am a black man, I hate this book, so do you

POSTED BY LZZ AT 9:33 PM 0 COMMENTS

xxx test xxx

I'm a bad man, I want to sale sale sale a magic drug! You will like it!

xxx test xxx

I'm a bad man, I want to sale sale sale a magic drug! You will like it!

POSTED BY LZZ AT 6:08 PM 2 COMMENTS

xxx test xxx

I'm a bad man, I want to sale sale sale a magic drug! You will like it!

POSTED BY LZZ AT 6:08 PM 2 COMMENTS

Figure **5-10**: Comparison of different Filtering Options

Figure **5-10** illustrates effect of seven available filtering options provided by ANGEL in different regions. Image A is the original texts appear in the webpage with ANGEL disabled. Image B and C show the effect of ANGEL with truncated option being chosen. In B’s situation, the whole suspicious paragraph is deleted, while in case of C, only suspicious words (‘test’ and ‘hate’) are removed. D and E depict the scenario while masking option has been chosen. Image D shows the suspicious words have been replaced with user defined phrase for substation, “MASKED” in this case, while in image E , suspicious words are deleted and images of angel are placed in the empty positions. Text paragraphs in image F , G ,H ,I, J indicate the consequence of different highlighting options. Dyed texts use high contrast background and foreground colors to provide warnings to users. These three images are results of user's choice among highlighting the whole paragraph (Image F), highlighting only the suspicious words (Image G) or highlighting suspicious contents with different color schemes (Image H, I, J) according to the score of badness computed by the machine learning module. Three color schemes are pre-loaded, with a green background representing the content is least likely to be offensive, the brown background representing the medium condition and the red background stands for the maximum likelihood of being offensive. Thresholds of score could be further tuned in the administrator mode under "Filtering configuration" panel.

5.3.3 Phish Tank Web Service

Web service based checking relies on public online services to help validate the content of suspicious paragraph. While collecting data usually takes long time or involves intensive computation resources, verification process are much faster. These online services are usually run by well-recognized organizations or trusted third parties. In the Section 4.2.3, we have mentioned that most anti-phishing tool bars have already applied such techniques in their implementations, and services they are referring to are either run by Phish tank , APWG, or the vendors themselves. The accuracy of the checking is determined by both the size of the phishing repository and the quality of data as most submitted phishing links should be validated with human assistance. In our case, we prefer the service provided by phish tank [78] to APWG [79], as the latter one does not provide free access to its phishing repository unless you pay a certain fee annually and maintain the membership. As indicated in their statistic page, phish tank database currently holds 384,184 phishing URLs. Since the phishing websites are always set up by malicious party for a specific campaign, most of these identified URLs become invalid shortly and up till now there are more than 8000 active phishing pages still online.

ANGEL use the standard Phish Tank API to access its service, to do this, it has to follow a series of user authentication procedures and send all http requests over encrypted channel (through port 443). A signature hashed with MD5 is required to be appended at the end of every http request so as to guarantee the integrity of the message.

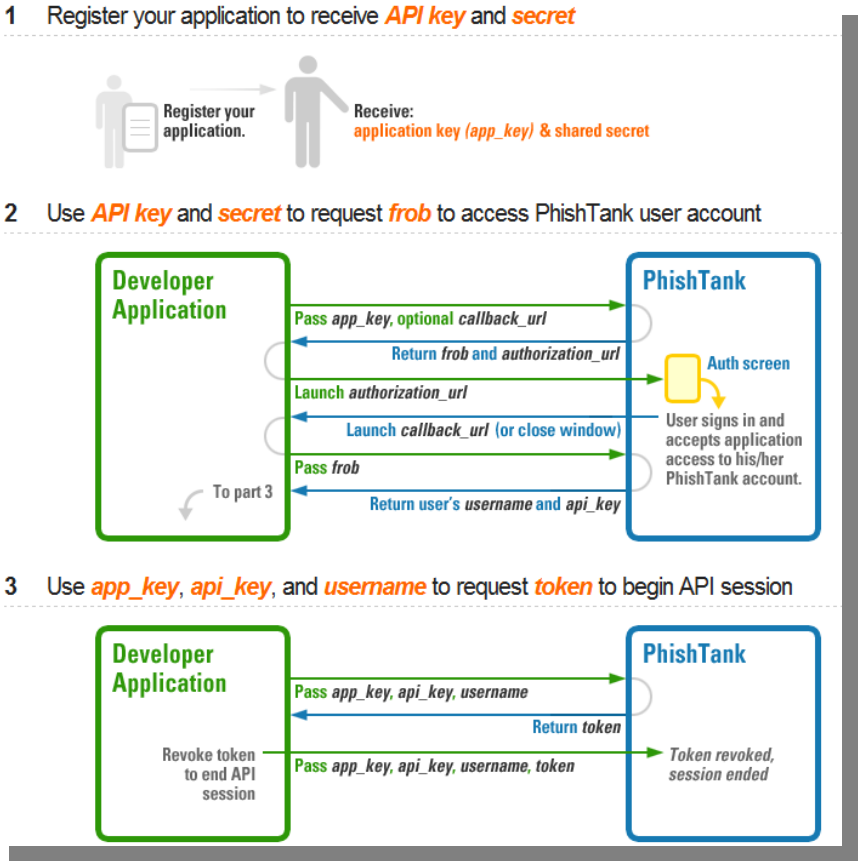

1 Register your application to receive API key and secret
Register your application.
Receive: application key (app_key) & shared secret
2 Use API key and secret to request frob to access PhishTank user account
Developer Application
PhishTank
Pass app_key, optional callback_url
Return frob and authorization_url
Auth screen
Launch authorization_url
Launch callback_url (or close window)
User signs in and accepts application access to his/her PhishTank account.
Pass frob
To part 3
Return user's username and api_key
3 Use app_key, api_key, and username to request token to begin API session
Developer Application
PhishTank
Pass app_key, api_key, username
Return token
Revoke token to end API session
Pass app_key, api_key, username, token
Token revoked, session ended

Figure 5-11: Phish tank Authentication procedure [78]

As illustrated in Figure 5-8, ANGEL first registers itself as a third party application with phish tank, and thereby obtain the application key and a shared secret. With these two credentials, it could further get the API key, which is one of the three required parameters to get the session tokens in the third step.

Once get the session token, ANGEL launches a URL validation request by concatenating application key, session token, the URL text under examination. A successful transaction always depends on a valid session token, which sometimes might expire and leads to a failed request. To counter this problem, ANGEL automatically restarts the token requesting process once receive the first error feedback and temporarily holds the checking process and save all unchecked URLs in memory. URL checking process restarts once a new valid token is obtained. Upon receiving the request string from ANGEL, the web service responses with one of following four results.

A) URL is not in the database

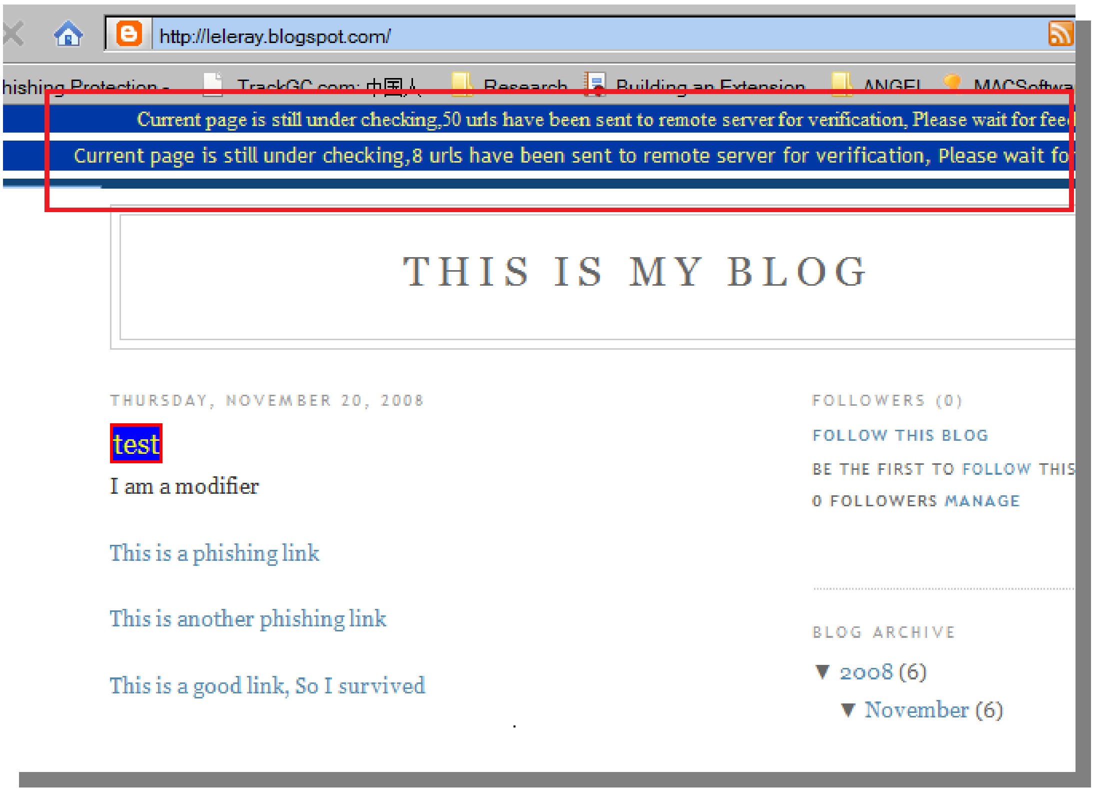
B) URL is in database but has not been verified yet

C) URL is in database and is verified as a false phishing link

D) URL is in database and is verified as a true phishing link

ANGEL marks a suspicious URL as phishing URL only if it receives a response in category D from the Phish Tank server, since in this case, it guarantees that current URL has been reported by someone to Phish Tank website as a suspicious phishing link, and then be verified by human staff and identified as an actual phishing URL.

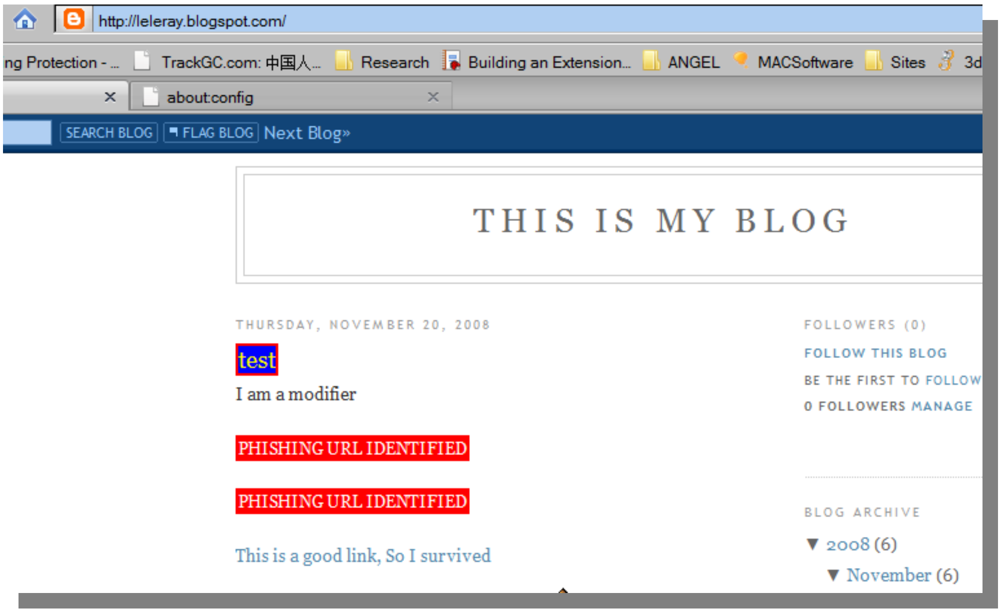
In most cases, filtering a webpage with local resources only takes few seconds, however, while applying web service based verification, network latency and other compulsory delays should be taken into account. First of all, as every time URL checking should always be followed by the user authentication process this cold start latency is inevitable. Second, token re-applying delay happens if current token expires. Finally, as there are normally more than 1 hyperlink (20 to 50 on average during evaluation) appears in one webpage, each hyperlink corresponds to a separate http request. After sending check-URL requests in sequential, ANGEL has to handle feedbacks in sequential as well. Since current webpage is not said to be sanitized until all feedbacks are correctly handled, it seems that we should postpone the display of current page. However, to do so, it would make user wait another 15 to 30 seconds on average. To counter such dilemma, a specialized banner, stating the number of hyperlinks being sent for verification, is inserted at the top of each webpage under examination. (Figure **5-12**). As shown in Figure **5-12**, sometime more than one banner would appear in one webpage, that is because other than the main frame, there is still another iframe embedded in current page, hence, Grease Monkey script is actually called twice, and both frames are under examination concurrently. Once all feedbacks have been handled and identified phishing URLs are removed this banner will be removed (Figure **5-13**).



Current page is still under checking,50 urls have been sent to remote server for verification, Please wait for fee

Current page is still under checking,8 urls have been sent to remote server for verification, Please wait fo

Figure 5-12: Waiting for feedback from remote server



THIS IS MY BLOG

THURSDAY, NOVEMBER 20, 2008

test

I am a modifier

PHISHING URL IDENTIFIED

PHISHING URL IDENTIFIED

This is a good link, So I survived

Figure 5-13: Handle feedbacks from remote server

5.4 Priorities and Checking Sequence

As deliberated in the above few sections, ANGEL could examine a target page with more than one criterion. That is, for any given webpage, the content of the pages could be checked against different rules in the case of heuristic module or examined with different statistical algorithm in the case of machine learning based module. Which order should we use when appliy these rules and algorithms? In general, we prefer heuristic rules to machine learning algorithms, and prioritize URL based rules to text based rules, besides, white list will be given the highest priority. The underlying reason of this ordering is that, admittedly, machine learning based rules could generally catch more subtleties of suspicious content in compared to their heuristic counterparts , however, their instinctive statistics based core also somewhat reduce the confidence of their judgment, as the result of these algorithm are best effort predictions with pre-known samples. On the contrary, heuristic rules are more often derived from known patterns of offensive instances or other reliable sources either from user’s personal experience or other trust authorities, for example, a list of well-recognized offensive words, thus, they are in certain should be given higher priorities.   In terms of URL checking, in ANGEL, all suspicious URLs will be firstly sent to a web service run by Phish Tank [78], after that, they are fed into the heuristic module if they pass the remote validation process successfully.

Chapter 6

Evaluation

6.1 Data Collection

Since machine learning module in ANGEL need to be initialized before using for the first time, we need to collect appropriate data corpus. Most data corpuses we could find are related to variants of spams. WEBSPAM-UK2007 corpus contains 6479 labeled web hosts (URLs) with 344 spam hosts and 5709 ham hosts, and remaining 426 are classified as undecided. TREC 2005 spam corpus is one of the most frequent used public corpuses in the realm of anti-spam research, which includes 39399 ham samples, and 52790 spam samples; the last one is a small spam blog corpus released by a group of researchers in University of Maryland Baltimore City. It has 1400 labeled blog homepages with 700 marked as spammy blogs and 700 marked as ham blogs. In ANGEL, we applied feature selection on later two corpuses and then use result feature vectors to initiate the machine learning module.

6.2 Evaluation Process

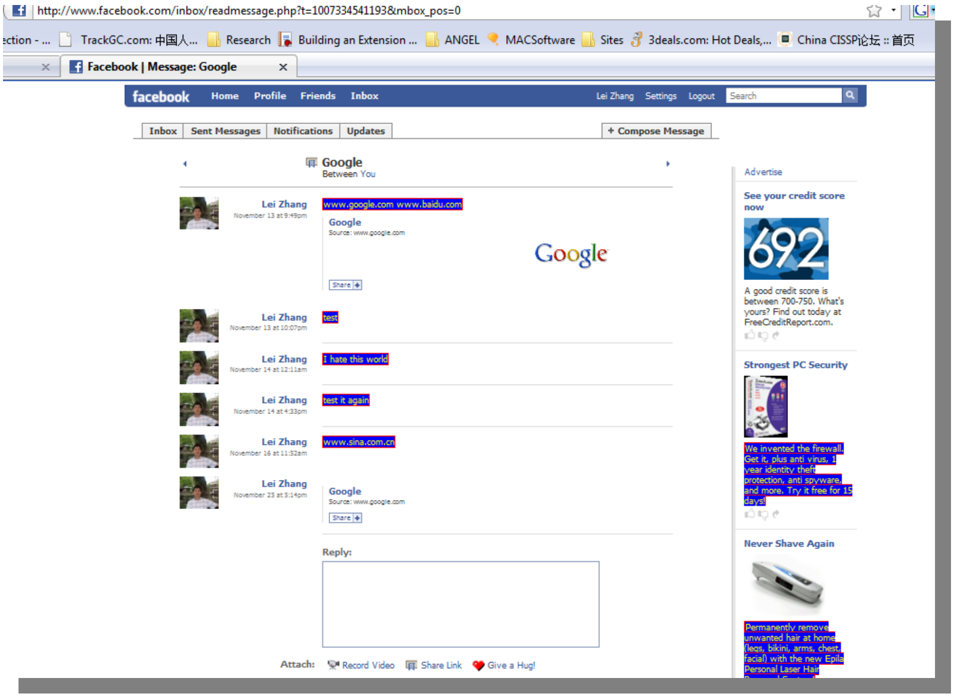
We first examine ANGEL with improvements we proposed to counter the weak points of state-of-art techniques. We list them again as following:

1. Black/white listing alone is not sufficient to filter inappropriate contents.
2. Labeling websites are not sufficient as well.
3. Inline Phishing URL checking is required.

4. Key words based heuristic rules are not sufficient.
5. Spamming messages should also be considered.
6. Simply warning or blocking the whole webpage is not effective.
7. Interaction is required between the protector and the protected.

To solve and improve item 1,2 and 5, in ANGEL, we implement hybrid filtering techniques which combine enhanced heuristic rules and machine learning based classification algorithms , which help us to look into the webpage contents, and thereby provide a more accurate assessment of webpage based on the classification result of each text paragraph it consists. As for item 3, by integrating Phish tank web service and auxiliary heuristic based rules, ANGEL possesses inline URL checking ability and also tolerates potential denial of service attack to the remote service. For item 4, ANGEL improves its heuristic module with both regular expression based rules and much powerful function based rules. User could now choose between performing training on the new samples or composing a new heuristic rule for the specific kind of threat. Plenty of options provided in ANGEL's filtering configuration panel make item 6 no longer a problem, users now could just choose a most preferable option for his children and make switches among different options on-the-fly. Finally, for the item 7, the white list submission interface in feedback configuration panel provides an easy way for teens and parents to communicate and work together to make ANGEL a better tool that relied and trusted both by the parent and their children.

Next we evaluate ANGEL with design principals we have proposed in Section 5.1. which are usability, extensibility and flexibility, effectiveness and performance. When designing the user interface, we keep in mind that the end user of ANGEL will be someone who has little knowledge of the PC or other filtering technology, therefore, although complicated techniques such as web service and machine learning module have been applied in ANGEL, users do not need to spend extra time studying how to use them or configure them, all user interfaces are simplified and made as intuitive as possible. For the second principle, ANGEL provides most flexible options among all available tools to present the filtering result, besides, its lose-coupling architecture allows it to be easily updated or install additional modules for further improvements. For the third principle, we perform tests on the popular social network website, Facebook (**Figure 6-1**) , and ANGEL successfully catches all suspicious contents on the page, besides, it also correctly identifies advertisements on the webpage with machine learning module, as these ads are somewhat similar to the spammy messages it have seen in its training sample. Compared to those ad-hoc ads removal software designed for specific websites, we believe that, with machine learning based classifier, ANGEL could also become excellent ads remover once fed sufficient training samples.



Lei Zhang November 13 at 9:49pm
www.google.com www.baidu.com
Google
Source: www.google.com

Lei Zhang November 13 at 10:07pm
test

Lei Zhang November 14 at 12:11am
I hate this world

Lei Zhang November 14 at 4:33pm
test it again

Lei Zhang November 16 at 11:52am
www.sina.com.cn

Lei Zhang November 25 at 5:14pm
Google
Source: www.google.com

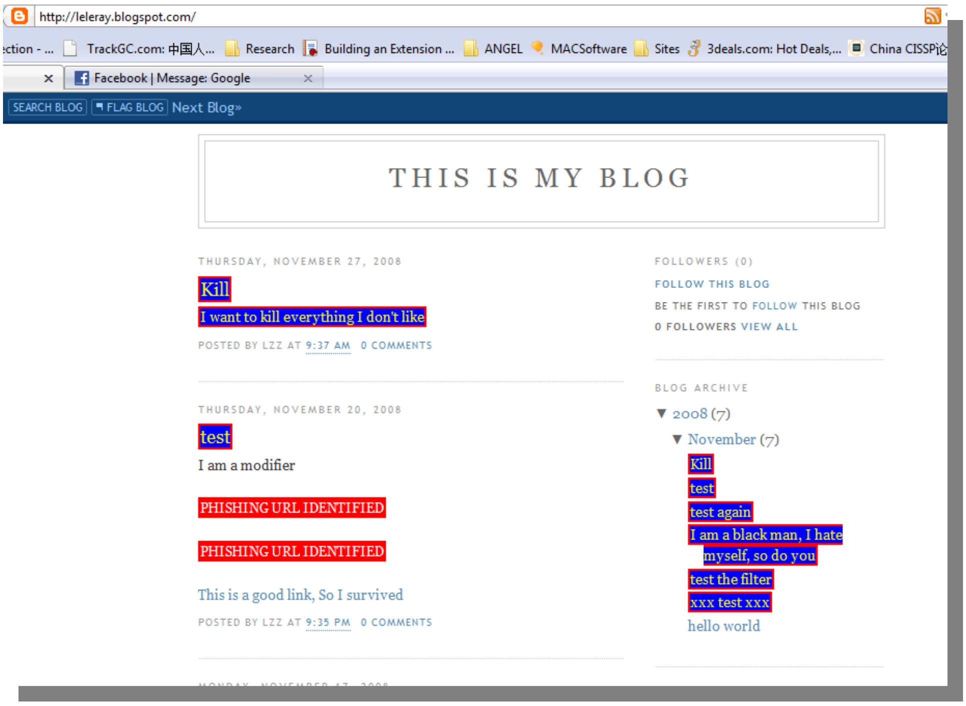
See your credit score now
A good credit score is between 700-750. What's yours? Find out today at FreeCreditReport.com.

Strongest PC Security
We invented the firewall. Get it, plus anti virus, 1 year identity theft protection, anti spyware, and more. Try it free for 15 days!

Never Shave Again
Permanently remove unwanted hair at home (legs, bikini, arms, chest, facial) with the new Epila Personal Laser Hair

Figure **6-1**: Evaluate ANGEL on Facebook

As Facebook owner is aware of the threats bought by phishing links, they use redirecting to counter such problem, therefore, hyperlinks shared by Facebook users are no longer the true URLs that points to the destination but a link to the Facebook server. To test the effectiveness of ANGEL towards phishing URLs, we try it on another well-known blog site, www.blogger.com , ranks 9th in the world in terms of traffic. ANGEL analyzes over 50 URLs in the test webpage, and correctly identifies and replaces the phishing URLs with highlighted warning phrases within 30 seconds (Figure 6-2).



THIS IS MY BLOG

THURSDAY, NOVEMBER 27, 2008

Kill

I want to kill everything I don't like

POSTED BY LZZ AT 9:37 AM 0 COMMENTS

THURSDAY, NOVEMBER 20, 2008

test

I am a modifier

PHISHING URL IDENTIFIED

PHISHING URL IDENTIFIED

This is a good link, So I survived

POSTED BY LZZ AT 9:35 PM 0 COMMENTS

FOLLOWERS (0)

FOLLOW THIS BLOG

BE THE FIRST TO FOLLOW THIS BLOG

0 FOLLOWERS VIEW ALL

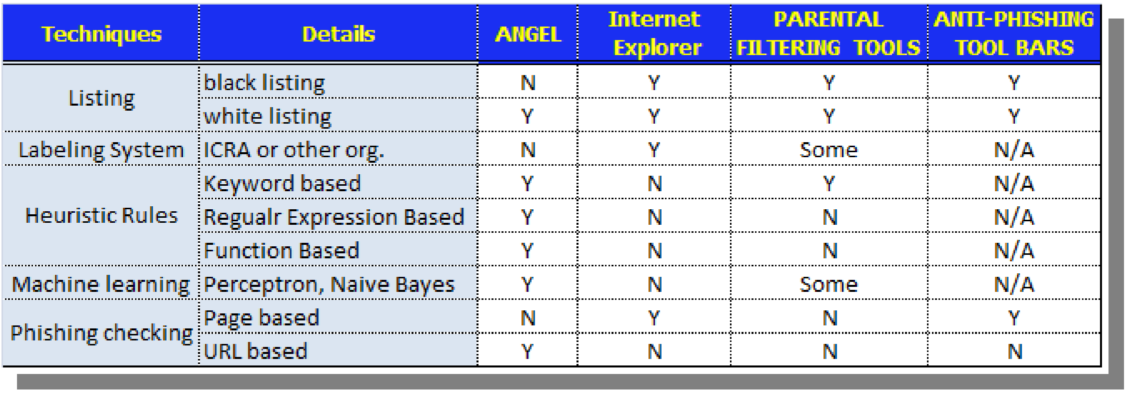
BLOG ARCHIVE

- ▼ 2008 (7)
  - ▼ November (7)
    - Kill
    - test
    - test again
    - I am a black man, I hate myself, so do you
    - test the filter
    - xxx test xxx
    - hello world

Figure **6-2**: Evaluate ANGEL on Blogger

The last item on the design principle list is the performance. When testing ANGEL on Facebook and Blogger.com, it takes just seconds for it to apply all URL or textual rules on each URL and paragraphs extracted from the current page, and surprisingly, the machine learning algorithm also behaves well. We believe that it is due to the relatively small size of each text paragraph and another reason is that the ML module is only called if the suspicious unit has passed heuristic - rule module. However, compared to checking with merely local resources, which are the rules in terms of heuristic module and feature vectors in the case of machine learning algorithms, checking with remote service usually takes much longer time to get the feedback. We have stated this issue in Section 5.3.3 and performed optimization on the check-URL process with reusable token and quick restart strategy once current token expires. With all these improvements, it takes ANGEL less than 20 seconds to validate 20 URLs and less than 30 seconds to verify 60 URLs.

Finally, we summarize key features of ANGEL make it different from other state-of-art tools in Table **5-2**.



Techniques	Details	ANGEL	Internet Explorer	PARENTAL FILTERING TOOLS	ANTI-PHISHING TOOL BARS
Listing	black listing	N	Y	Y	Y
	white listing	Y	Y	Y	Y
Labeling System	ICRA or other org.	N	Y	Some	N/A
Heuristic Rules	Keyword based	Y	N	Y	N/A
	Regualr Expression Based	Y	N	N	N/A
	Function Based	Y	N	N	N/A
Machine learning	Perceptron, Naive Bayes	Y	N	Some	N/A
Phishing checking	Page based	N	Y	N	Y
	URL based	Y	N	N	N

Table **6-1** Feature Comparison Table

6.3 Security Evaluations

We have proved in the last section that, ANGEL fulfills all goals we want to achieve and outperforms all other state-of-art tools in terms of providing better protection to children's online social networking against the threats from spamming, phishing, offensive words and web scams. Now we will ask such question, if ANGEL is vulnerable to any kinds of attacks launched by malicious party? What would happen if our children, the person we want to protect, want to attack the tool either due to curiosity or rebelliousness**?** Formally, we define the former attack model as outside attacks and the latter one as insider attacks.

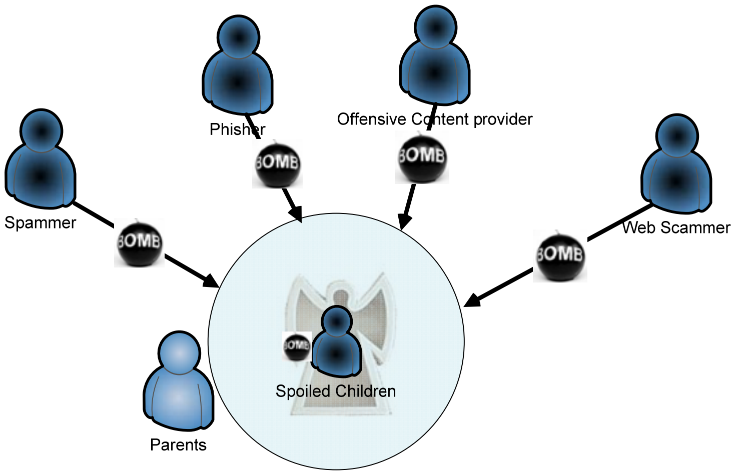


Figure **6-3**: Attack Model

6.3.1 Outsider Attack

We assume that external attackers in this mode have little knowledge of our system and could only influence the performance of ANGEL by either **(a)** disable some key functions of the tool, **(b)** fool the tool to create unbearable number of false positives so that parents, who are administrators of ANGEL, have no choice but to shut it down, or **(c)** bypass the guard by changing the words or URLs shown in the webpage.

Let us examine case (a) first. Start with the simpler case, in which both browser itself and operation system ANGEL resides in have not been compromised or modified. Under such condition, attacker could only influence ANGEL by restricting external resources it relies on. After reviewing the system architecture, blacklist resides on Phish Tank server is the only possible weak point they could exploit. A DOS attack to the Phish Tank server might make such attack succeed. Phish Tank filter [77], did fail in such scenario. While we can’t guarantee the performance and availability of the Phish Tank server, we claim that, even under such extreme case, ANGEL could still differentiate a large amount of phishing URLs by either using multiple blacklisting resources, making local copy of blacklist periodically, and using pre-defined heuristic rules. Now what if the computer is compromised and the attacker could access the files on the local system. Such condition is equivalent to the insider attack which we will cover in section 6.4.2.

Now let us examine case (b) and (c), and evaluate the possibility that attackers can influence ANGEL's performance by causing high false positive or trying to get around ANGEL's filtering. To make this happen, there are general two potential ways. First one is that, the attacker could take advantage of the white listing function applied in ANGEL, that is, alter the content of the webpage once the URL of that page has been approved by the administrator. Although it seems to be quick and effective at first glance, in real case, it is quite un-realistic for the webpage owner to known if the URL is listed on white list or not. The other way is to get around or fool the ANGEL in which case attackers have to know either the content of heuristic rules or the feature information of machine learning based classifier, which is an impossible mission for attacker as ANGEL residing at client's end and providing no feedback of filtering result to the server that hosts the webpage.

6.3.2 Insider Attack

Now we come to a more interesting situation, in which the children we want to protect is really spoiled and want to get access to some contents which are not allowed by the administrator or parents. We discuss such case under the presumption that Firefox is the only browser child could use and he has not administrator privilege to install or uninstall software.

As the child has no idea of the password, and the password is stored as hashed token on the local file system, therefore he could not get access to the configuration panel to disable the tool or make changes to existing heuristic rules. Plus, as all the additional training samples are also encrypted with MD5 and using checksum to guarantee its integrity, it is reasonable to believe that ANGEL could defeat such attacks.

What if the shared secret or the password of configuration UI has been compromised as well, we estimate the degradation of performance in such worst scenario, where all settings might be manipulated. Recap that ANGEL's classification module composes of two sub-modules, heuristic -based and machine-learning-based, we claim that both of them are subjective to attacks under current condition. For the heuristic-based module, attackers can either disable or modify existing rules. As for the machine-learning-based module, attacker can change the behavior of learning by supplying fabricated samples. Compared to the heuristic sub-module, whose power attenuates quickly with the number of rules being modified or deleted, the performance of machine-learning-based module is more stable therefore takes attacker much more time to make a significant change. In [84], researchers have studied potential attacks to machine learning based applications, and they also proposed various types of defense against such attacks.

Chapter 7

Conclusion and Future work

Just like Danah boyd said [85] , today's teenagers are being socialized into a society complicated by shifts in the public and private. Emergence of new social technologies like social networking websites has dramatically altered the underlying architecture of social interaction and information distribution. The fact we have to admit is that, nowadays, besides valuable knowledge and infinite resources that are accessible online, there are also many malicious parties who create and propagate spam, phishing, offensive contents and web scams to threat and impair teens' experience of using such new technologies. In this work, we analyze the severity of such problem and pointed out the insufficiency of state-of-art approaches, we also propose and implement ANGEL to serve as our first attempt to solve this problem. There are many possible improvements could be done in future to further enhance the capability of ANGEL. Resorting to more phishing repositories could definitely reduce the miss rate. As some words have different meaning in different scenarios, thus a better way to find offensive words for heuristic rules is to analyze the text paragraphs with more semantic approaches or identify the topic model of webpage as a whole before making a judgment on the classification of a single word, Turney [86] has already conducted some exploration experiments in this direction. Besides, collaboration among different ANGEL users and providing them an easy and efficient way to share heuristic rules or feature vectors could also be a promising direction. With its extensible architecture, we believe a more powerful ANGEL will be created someday and serve as a loyal guard to protect teens' safety in online social network.

Appendix A

System Configurations

- ✓ Firefox 3.0.4
- ✓ Internet Explorer 7.0.6001.18000
- ✓ Grease Monkey 0.8.20080609.0
- ✓ Ruby 1.8.6
- ✓ JavaScript 1.7

Reference list

1) Danah m. boyd, Nicole B. Ellison Social Network Sites: Definition, History, and Scholarship 2007

2) Myspace www.myspace.com

3) Cloudmark, securing communications for social networking providers

4) Blocked Sites and Offensive Videos - The Challenges of Teen Computer Use

5) Lawrence Bartlett, Bullies In Cyberspace Spark Growing Concern

6) Gisle Hannemyr, Web Scams. http://hannemyr.com/links/webscams.html

7) Marcus A.Maloof Machine learning and data mining for computer security

8) Rapleaf, Study of Social Network Users vs. Age 2008

9) Jeremiah Owyang, Social Network Stats: Facebook, MySpace, Reunion , Jan, 2008

10) Alexa, www.alexa.com ,Nov.7 2008

11) Cybertip , Internet Safety (age-Specific Tips) http://www.cybertip.ca

12) Amanda lenhart, Mary Madden, Teen, Privacy & Online Social Networks. PEW/INTERNET

13) The MySpace Generation. Bussiness week December 2005.

14) MyYearbook, www.MyYearbook.com

15) Stanley Wasserman and Katherine Faust, Social network analysis : methods and applications

16) www.facebook.com

17) http://www.rockyou.com/corp/facebook/dev.php

18) Data harvesting http://en.wikipedia.org/wiki/Data_harvesting

19) David Meyer, ZDNet UK, Facebook admits it's the in thing for spammers, April 2008

20) Myspace API http://developer.myspace.com/community/

21) OpenSocial API, Google, http://code.google.com/apis/opensocial/

22) Commtouch,Email Threats Trend Report,Quarter 1 2008

23) W.Gansterer,M.Ilger,P.Lechner,R.Neumayer,J.Straub, Anti - Spam Methods State-of-the-Art.

24) Detecting Spam Web Pages through Content Analysis,May 23-26, 2006.

25) Carlos Castillo,Debora Donato,Luca Becchetti,Paolo Boldi,Stefano Leonardi, Massimo Santini and Sebastiano Vigna, A Reference Collection for Web Spam,2006.

26) Zoltan Gyongyi,Hector Garcia-Monlina, Web Spam Taxonomy.

27) Harris Interactive & Clouldmark. Survey Shows Rise in Social Network Spam http://www.harrisinteractive.com/NEWS/newsletters/clientnews/2008_Cloudmark.pdf

28) A list of bots for Social network website http://allbots.info/

29) BitDefender. Trojan Trojan Now Uses Hotmail, Gmail as Spam Hosts , www.bitdefender.com/world/News/pdfDescription/544.pdf

30) Kapaskey, Spam Report, www. Kapaskey.com, June 2008

31) Phishing Scams, http://www.irs.gov/, March 13, 2008

32) Min Wu, Robert C. Miller, Simson L.Garfinkel. Do Secuirty Toolbars Actually Prevent Phishing Attacks. MIT

33) Aditya K Sood aka Zeroknock, http://www.secniche.org, Phishing and Redirection Vulnerability in Yahoo Network

34) Adobe Product Security Incident Response Team, Verifying Installers

35) PC Magazine. Facebook Worm Spreads Rapidly. Watch Where You Get Your Flash Player , August 6, 2008

36) Nancy Willard, Cyberbullying and Cyberthreats Effectively Managing Internet Use Risks in Schools, http://cyberbully.org

37) Michele L. Ybarra, Kimberly J. Mitchell ,How Risky Are Social Networking Sites? A Comparison of Places Online Where Youth Internet Solutions for Kids, Inc, Santa Ana, CaliforniaSexual Solicitation and Harassment Occurs

38) Abbott Koloff,States push for cyberbully controls, USA TODAY 2008

39) National Consumers league www.frad.org

40) Government of Canada. The Canadian strategy to promote safe, wise and responsible internet use. www.cybertip.ca / www.connect.gc.ca/cyberwise

41) Cyber Bullying State Laws and Policies http://www.cyberbullyalert.compolicies/

42) Federal Bureau of Investigation, A Parent's Guide to Internet Safety, http://www.fbi.gov/publications/pguide/pguidee.htm

43) Parent/Child Online Agreement, http://life.familyeducation.com

44) Market Share, Browser Version Market Share October, 2008

45) ICRA. 2008 Vocabulary. http://www.icra.org/vocabulary/

46) Profil. Parental Filter. http://www.parentalfilter.eu/

47) TopChoice. Parental control software reviews, http://parental-control-software.topchoicereviews.com/

48) Lorrie Cranor, Serge Egelman, Jason hong, and Yue Zhang Phinding Phish: An Evaluation of Anti-Phishing Toolbars CMU 2007

49) Yue Zhang Jason Hong Lorrie Cranor, CANTINA: A Content-based Approach to Detecting Phishing Web Sites, WWW 2007

50) Lan Fette, Norman Sadeh, Anthony Tomasic, Learning to Detect Phishing Emails, WWW 2007

51) Sujata Garera, Niels Provos, Monica Chew, Aviel D. Rubin. A framework for detection and measurement of phishing attacks. 2007 ACM workshop on Recurring mal-code

52) Focus Adolescent Services. Connection, Monitoring, Autonomy Rules & Boundaries. http://www.focusas.com/Parenting.html

53) Machine Learning, Tom Mitchell, McGraw Hill, 1997]

54) Machine Learning, http://en.wikipedia.org/wiki/Machine_learning

55) Konstantin etc. Machine Learning Techniques in Spam Filtering 2004

56) Pranam Kolari, Akshay Java, Tim Finin,Tim Oates,Anupam Joshi, Detecting Spam Blogs: A Machine Learning Approach University of Maryland Baltimore County ,2006

57) Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, Suku Nair, A Comparison of Machine Learning Techniques for Phishing Detection, APWG e Crime Researsher Summit 2007

58) Anderson, J.P. Computer security threat monitoring and surveillance. 1980

59) Denning, D.E An intrusion-detection model. 1987

60) Learning to Detect and Classify Malicious Executables in the Wild

61) James P. Early and Carla E. Brodley Behavioral Features for Network Anomaly Detection, Springer London, 2006

62) S Eyheramendy, D.Lewis On the Naive Bayes model for text categorization

63) P.Graham. Better Bayesian Filtering

64) P.Graham. A Plan for Spam

65) M. Sahami,S.Dumais,.. A Bayesian approach to filtering junk e-mail

66) Vangelis Metsis , Ion Androutsopoulos, Spam filtering with Naive Bayes- which Naive Bayes?

67) Rosenblatt, Frank, The Perceptron: A Probabilistic Model for Information Storage and

Organization in the Brain 1958

68) Christopher M. Bishop, Pattern Recognition and Machine learning

69) D. Sculley, Gabriel M. Wachman, and Carla E. Brodley, Spam Filtering using Inexact String Matching in Explicit Feature Space with On-Line Linear Classifiers tufts.edu

70) s.a. bleha, j. knopp and m.s. obaidat performance of the perceptron algorithm for the classification of computer users, university of missouri – Columbia

71) Isabelle Guyon Andre Elisseeff, An Introduction to Variable and feature selection, Journal of Machine Learning Research

72) Marko Grobelnik , Feature selection for unbalanced class distribution and Naïve Bayes. ICML 1999.

73) R. Kohavi, G. H. John. Wrappers for feature selection. Artificial Intelligence, 1997.

74) David J. Hand, Classifier Technology and the Illusion of Progress , Statistical Science 2006

75) List of Bad Words http://www.noswearing.com/about.php

76) List_of_ethnic_slurs http://en.wikipedia.org/wiki/List_of_ethnic_slurs#E

77) Phish tank checker http://phishtanksitechecker.com/

78) Phish tank API. http://www.phishtank.com/api_documentation.php

79) APWG. Anti-Phishing Working Group http://www.apwg.org

80) Greasemonkey http://www.greasespot.net/

81) World Wide Web Consortium http://www.w3.org/

82) Peter Jackson and Isabelle Moulinier, Natural Language Processing for Online Applications. John Benjamins publishing company

83) M.F.Porter, An algorithm for suffix stripping, 1980 issue of the journal Program.

84) Marco Barreno,Blaine Nelson, Russell Sears ,Anthony D. Joseph, J. D. Tygar , Can Machine Learning Be Secure?, ASIACCS’06

85) Danah boyd , Social Network Sites: Public, Private, or What?, The Knowledge Tree : An e-Journal of Learning Innovation

86) Turney,P , Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Association for Computational Linguistics, 2002