Battling the Internet Water Army: Detection of Hidden Paid Posters

by

Cheng Chen
B.Sc., Beijing University of Posts and Telecommunications, 2010

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in the Department of Computer Science

Battling the Internet Water Army: Detection of Hidden Paid Posters

by

Cheng Chen
B.Sc., Beijing University of Posts and Telecommunications, 2010

Supervisory Committee

Dr. Kui Wu, Co-Supervisor
(Department of Computer Science)

Dr. Venkatesh Srinivasan, Co-Supervisor
(Department of Computer Science)

Supervisory Committee

Dr. Kui Wu, Co-Supervisor
(Department of Computer Science)

Dr. Venkatesh Srinivasan, Co-Supervisor
(Department of Computer Science)

ABSTRACT

Online social media, such as news websites and community question answering (CQA) portals, have made useful information accessible to more people. However, many of online comment areas and communities are flooded with fraudulent information. These messages come from a special group of online users, called online paid posters, or termed "Internet water army" in China, represents a new type of online job opportunities.

Online paid posters get paid for posting comments or articles on different online communities and websites for hidden purpose, e.g., to influence the opinion of other people towards certain social events or business markets. Though an interesting strategy in business marketing, paid posters may create a significant negative effect on the online communities, since the information from paid posters is usually not trustworthy.

We thoroughly investigate the behavioral pattern of online paid posters based on a real-world trace data from the social comments of a business conflict. We design and validate a new detection mechanism, including both non-semantic analysis and semantic analysis, to identify potential online paid posters. Using supervised and unsupervised approaches, our test results with real-world datasets show a very promising performance.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

The two-year graduate study in the University of Victoria is a fascinating experience for me. It is a great pleasure to express my gratitude to many people who made this thesis possible. I would like to thank:

My Co-Supervisors, Dr. Kui Wu and Dr. Venkatesh Srinivasan, for providing me with this valuable opportunity to work with you. I'm deeply inspired by your scientific spirit of research. I can hardly find words that are able to fully express my gratitude for you. This thesis would not have been possible without your considerable tutoring, constant encouragement and patient correction along the past two years.

PANDA Research Group, for the discussion we undertook. I'm appreciate all the excellent suggestions you offered. Without your help, this thesis may not appear as complete as it does now. It is a quite unforgettable memory for me to do research with you.

My Parents, for supporting me to study abroad. My mother has always been there whenever I feel discouraging and depressed. My father often discusses research with me which gives me extra insights how to further improve my work. Thank you both for always encouraging me to overcome the difficulties in the daily life.

Chapter 1

Introduction

1.1 Who are the Online Paid Posters

According to China Internet Network Information Center (CNNIC) [4], there are currently around 457 million Internet users in China, which is approximately 35% of its total population. In addition, the number of active websites in China is over 1.91 million. The unprecedented development of the Internet in China has encouraged people and companies to take advantage of the unique opportunities it offers. One substantial issue is how to make use of the huge online human resource to make the information diffusion process more efficient. Among the many approaches to e-marketing [2], we focus on *online paid posters* used extensively in practice.

Working as an online paid poster is a rapidly growing job opportunity for many online users, mainly college students and the unemployed people. These paid posters are referred to as the "Internet water army" in China because of the large number of people who are well organized to "flood" the Internet with purposeful comments and articles. This new type of occupation originates from Internet marketing, and it has become popular with the fast expansion of the Internet. Often hired by public relationship (PR) companies, online paid posters earn money by posting comments and articles on different online communities and websites. Companies are always interested in effective strategies to attract public attention towards their products. The idea of online paid posters is similar to word-of-mouth advertisement. If a company hires enough online users, it would be able to create hot and trending topics designed to gain popularity. Furthermore, the articles or comments from a group of paid posters are also likely to capture the attention of common users and influence their decision. In this way, online paid posters present a powerful and efficient strategy for companies. To give one example, before a new TV show is broadcast, the host company might hire paid posters to initiate many discussions on the actors or actresses of the show. The content could be either positive or negative, since the main goal is to attract attention and trigger curiosity.

1.2 Typical Examples

To better understand the behavior and the social impact of online paid posters, we investigated several social events, which are likely to be boosted by online paid posters. We introduce two typical cases to illustrate how online paid posters could be an effective marketing strategy, in either a positive or a negative manner.

1.2.1 The Event of Junpeng Jia

On July 16, 2009, someone posted a thread with blank content and a title of "*Junpeng Jia, your mother asked you to go back home for dinner!*" on a Baidu Post Community of World of Warcraft, a Chinese online community for a computer game [15]. In the following two days, this thread magically received up to $300,621$ replies and more than 7 million clicks. Nobody knew why this meaningless thread would get so much attention. Several days later, a PR company in Beijing claimed that they were the people who designed the whole event, with an intention to maintain the popularity of this online computer game during its temporary system maintenance. They employed more than 800 online paid posters using nearly $20,000$ different user IDs. In the end, they achieved their goal– even if the online game was not temporarily available, the website remained popular during that time and it encouraged more normal users to join. This case not only shows the existence of online paid posters, but also reveals the efficiency and effectiveness of such an online activity.

1.2.2 A Business Conflict between Two Companies

On July 17, 2009, a Chinese IT company *Qihu 360*, also known as *360* for short, released a free anti-virus software and claimed that they would provide permanent anti-virus service for free. This immediately made *360* a super star in anti-virus software market in China. Nevertheless, on July 29 an article titled "*Confessions from* *a retired employee of 360*" appeared in different websites. This article revealed some inside information about *360* and claimed that this company was secretly collecting users' private data. The links to this post on different websites quickly attracted hundreds of thousands of views and replies. Though *360* claimed that this article was fabricated by its competitors, it was sufficient to raise serious concerns about the privacy of normal users. Even worse, in late October, similar articles became popular again in several online communities. *360* wondered how the articles could be spread so quickly to hundreds of online forums in a few days. It was also incredible that all these articles attracted a huge amount of replies in such a short time period.

In 2010, *360* and Tencent, two main IT companies in China, were involved in a bigger conflict. On September 27, *360* claimed that Tencent secretly scans user's hard disk when its instant message client, QQ, is used. It thus released a user privacy protector that could be used to detect hidden operations of other software installed on the computer, especially QQ. In response, Tencent decided that users could no longer use their service if the computer had *360*'s software installed. This event led to great controversy among the hundreds of thousands of the Internet users. They posted their comments on all kinds of online communities and news websites. Although both *360* and Tencent claimed that they did not hire online paid posters, we now have strong evidence suggesting the opposite. Some special patterns are definitely unusual, e.g., many negative comments or replies came from newly registered user IDs but these user IDs were seldom used afterwards. This clearly indicates the use of online paid posters.

Since a large amount of comments/articles regarding this conflict is still available in different popular websites, we in this thesis focus on this event as the case study.

1.3 Importance of Detection of Online Paid Posters

Overall, the consequences of using online paid posters are yet to be seriously investigated. While online paid posters can be used as an efficient business strategy in marketing, they can also act in some malicious ways.

As the rapid development of online social network services and E-Business, such as Facebook, Twitter and Amazon, it becomes more convenient for people to be connected with each other for feedback and advice. Before making an decision to purchase some goods, people are more likely to find relevant online reviews of the products. Comments can lead to many different results; positive comments often attract more customers while negative ones harm the sales of a product. Compared to advertisements broadcast on televisions or printed on newspapers, hiring online paid posters to post fake comments is a much cheaper approach [35]. Disguised as normal and professional comments, fake reviews have the ability to mislead people's evaluation of products. One such example is that some companies claim that they have the marketing strategy to boost the ranking of the App Store applications. Most of those companies would use so-called *bot farm* to automatically download the applications and post comments so as to hype the popularity of applications. Some of these companies also hire real humans to do this job. It has been a problem that makes Apple to continuously change the ranking algorithms.

Moreover, as shown in the second example, the malicious online activities are harmful to the order of market. Companies have the desire to hire online paid posters to spread negative information towards competitors across the Internet. Normal people, however, would be disappointed when they realize that they can hardly trust what they see and what they read.

We would like to remark here that the use of paid posters extends well beyond China. According to a news report in the Guardian [7], the US military and a private corporation are developing a specific software that can be used to post information on social media websites using fake online identifications. The objective is to speed up the distribution of pro-American propaganda. We believe that it would encourage other companies and organizations to take the same strategy to disseminate information on the Internet, leading to a serious problem of spamming.

Since the laws and supervision mechanisms for Internet marketing are still not mature in many countries, it is possible to spread wrong, negative information about competitors without any penalties. For example, two competitive companies or campaigning parties might hire paid posters to post fake, negative news or information about each other. Obviously, ordinary online users may be misled, and it is painful for the website administrators to differentiate paid posters from the legitimate ones. Hence, it is necessary to design schemes to help normal users, administrators, or even law enforcers quickly identify potential paid posters.

1.4 Related Work

In this thesis, we focus on paid posters who post comments online to influence people's thoughts regarding popular social and business events. We characterize the basic organizational structure of paid posters as well as their online posting patterns.

1.4.1 The Goal of this Thesis

This thesis conducts comprehensive study of evaluation of a real business battle between two Chinese companies. Our work can shed light on different websites of other languages because we reveal effective features of hidden online paid posters which were not covered in previous work. Our goal is to identify the potential paid posters through learning their behavioural patterns.

According to [35], China has the world's largest Internet population (485M) and two largest and most representative crowd-sourcing systems are host on Chinese networks. The work of spreading rumors and malicious advertisements are accomplished by the large amount of crowd-sourced labor. These malicious activities are also called crowd-turfing. The researchers in [35] conducted a survey on the crowd-turfing market in other countries. They found that only 12% of all the campaigns on the Amazon Mechanical Turk, a US-based crowd-sourcing website, were crowd-turfing type, decreasing tremendously from 41% spam-related tasks according to a report [14] in 2010. This confluence of factors suggest to us that dataset collected from the Chinese websites will provide us varieties of spam so as to make our research robust to different types of spam attack.

1.4.2 Trustworthy Content in Social Media

As Web 2.0 social websites play an increasingly important role on the Internet, social media has become a popular research topic. The social platforms are shown to have the ability to take advantage of the wisdom of crowds [34]. With easy-access and large number of users, those websites offer numerous information. Using the online information, scientists attempt to make use of the social networks to solve varieties of problems. For example, Sakaki *et al.* [28] find clues to predict the earthquake and other events from the user-generated contents. Kwak *et al.* [17] discuss the role and the impact of Twitter in today's society. Our life can benefit from these interesting research. However, the social networks are also known to suffer from uneven user-generated content, from professional to poor or even fake messages. Retrieving high-quality information from the Internet has become a significant task.

Previous work focused on forum and blog spammers who posted advertisements or malicious URLs on the websites. The spammers in those scenarios used software to post malicious comments on their forums and blogs to change the results of search engine or to make their sites popular. However, the definition of spam has been extended to a much wider concept. Basically, any user whose behavior interferes with normal communication or aids the spread of misleading information is specified as a spammer. Examples include comment spammers and review spammers in social media and online shopping stores.

Yin *et al.* [38] studied so-called online harassment, in which a user intentionally annoyed other users in a web community. They investigated the characteristics of harassment using local features, sentimental features and contextual features.

Gao *et al.* [8] conducted a broad analysis on spam campaigns that occurred in Facebook network. From the dataset, they noticed that the majority of malicious accounts were compromised accounts, instead of "fake" ones created for spamming. Such compromised accounts can be obtained through trading over a hidden online platform, according to [33].

Huang *et al.* [12] developed a regression model with features suggesting quality-biased short text in Microblogging service, Twitter. They judged the quality of tweets based on relevance, informativeness, readability, and politeness of the short content and assigned different scores from 1 to 5. However, they didn't explicitly present how they define a spam-like tweet. Huang [11] conducted a similar study on commercial spam on blogging sites. They showed that the propaganda of some products in the comment of a blog post was crucial in detecting the malicious comments. The propaganda appeared in the form of URL, phone number, E-mail address, MSN numbers and etc.

Morris *et al.* [22] conducted a survey about assessing the credibility of tweets on Twitter. They examined several features, *message topic*, *user name*, *user image* and et al., to find out how they influence people's perception towards the credibility of tweets. According to their findings, it is often quite difficult for people to judge the credibility of tweets base on the content of tweets. People tend to trust users who have real photos and messages which are retweeted for several times. However, this findings would encourage the paid posters to better hide themselves since creating online identities with photos can be easily implemented and multiple user accounts can be controlled to retweet the malicious messages.

Moturu *et al.* [23] studied the problem of trustworthiness on health content from two social media applications. They developed a two-step framework to get trustworthy information. In the feature identification step, they proposed features indicating the credibility of the content. In the qualification step, they tested four different scoring models based on the features.

Willemsen *et al.* [37] collected review data from Amazon.com and conducted a research on which content characteristics have effect on the perceived usefulness.

1.4.3 Crowd-sourcing Spam

For the crowd-sourcing spam, such as online paid posters, Jindal *et al.* [16] and Ott *et al.* [26] worked on detecting review or opinion spam in the online shopping stores, like Amazon's online store.

The work by Jindal and Liu [16] had similarities to our research. They studied a dataset crawled from Amazon.com, the online shopping store, and tried to detect "opinion spam" or "review spam". By their definition, review spammers posted undeserving positive opinion or malicious negative opinion on a product. In [16], the authors assumed the review spammer acts individually. In a recent work [25, 24], the authors focused on detecting groups of spammers on Amazon. According to our study, however, paid posters have their own posting patterns and do not exhibit the features proposed in [16, 25, 24]. In our work, the data is not reviews for products, but any social comments regarding the components of the two companies, the chairman, products and marketing activities. As a result, the features used in our work are different from those in [16, 25, 24]. In addition, our dataset comes from a real business conflict and thus we have high confidence in the involvement of online paid posters. Furthermore, our semantic analysis method to improve detection performance is based on the identification of common content words and is different from those in [16, 25, 24].

The problem addressed by Ott et al.[26] is different from ours. The main task of Ott et al. is to detect fictitious opinions that are deliberately (and intelligently) written to be authentic. To emphasize this point, the authors set strict quality control on the fictitious posts, that is, any submission found to be of insufficient quality, e.g., written for the wrong hotel, unintelligible, unreasonably short, plagiarized, etc., will be rejected. And they also require that "the review needs to sound realistic and portray the hotel in a positive light". In other words, they try to tackle a long-time known problem of detecting professional "ghost writers", who are skilful in writing and are capable of writing very persuasive articles. In contrast, we do not intentionally focus only on the deceptive opinions, but instead we aim at detecting disruptive comments, which are not hard to determine if a person has enough resource and time, i.e., she/he has collected a large pool of comments from different sites, a large pool of user IDs, and she/he has enough patience as us to read all comments and compare the comments (which may distribute over different web sites) from a same user.

1.5 Contribution of This Thesis

Despite the broad use of paid posters and the damage they have already caused, it is unfortunate that there is currently no systematic study to solve the problem. This is largely because online paid posters mostly work "underground" and no public data is available to study their behavior. This thesis is within the first few works that tackle the challenges of detecting potential paid posters. We make the following contributions.

1. By working as a paid poster and following the instructions given from the hiring company, we identify and confirm the organizational structure of online paid posters similar to what has been disclosed before [20].

2. We collect real-world data from popular websites regarding a famous social event, in which we believe there are potentially many hidden online paid posters.

3. We statistically analyze the behavioral patterns of potential online paid posters and identify several key features that are useful in their detection.

4. We integrate semantic analysis with the behavioral patterns of potential online paid posters to further improve the accuracy of our detection.

The rest of the thesis is organized as follows. We introduce more background information and identify the organizational structure of online paid posters in Chapter 2. We describe the way of data collection and labeling in Chapter 3. We present the statistical and semantic analysis in Chapter 4 and Chapter 5 respectively. After that, we apply supervised learning and unsupervised learning methods to classify the online users. We then provide a design of real-time detection system in Chapter 7. The thesis is concluded in Chapter 8.

Chapter 2

How Do Online Paid Posters Work?

2.1 Basic Overview

These days, some websites, such as *shuijunwang.com* [29], offer the Internet users the chance of becoming online paid posters. To better understand how online paid posters work, I registered on such a website and worked as a paid poster. My experience was then summarized to illustrate the basic activities of an online paid poster.

Once online users register on the website with their Internet banking accounts, they are provided with a mission list maintained by the webmaster. These missions include posting articles and video clips for ads, posting comments, carrying out Q&A sessions, etc., over other popular websites. Normally, the video clips are pre-prepared and the instructions for writing the articles/comments are given. There are project managers and other staff members who are responsible for validating the accomplishment of each poster's mission. Paid posters are rewarded only after their assignments pass the validation. An assignment is considered a "fail" if, for example, the posted articles or contents are deleted by other websites' administrators. In addition, there are some regular rules for the paid posters. For example, articles should be posted at different forums or at different sections of the same forum; Comments should not be copied and pasted from other users' replies; The mission should be finished on time (normally within 3 hours), and so on.

Although the mission publisher has regulations for paid posters, they may not strictly follow the rules while completing their assignments, since they are usually rewarded based on the number of posts. That is why we can find some special behavioral patterns of potential paid posters through statistical analysis.

2.2 Management of Paid Posters

Occasionally, PR companies may hire many people and have a well-organized structure for some special events. Due to the large number of user IDs and different post missions, such an online activity needs to be well orchestrated to fulfill the goal. Being hired as a paid poster, we got a chance to read internal guidelines and observe the organizational structure of online paid posters. The observation was consistent with the findings in [20]. We remark here that other types of organizational structures different from the one described here are also possible.

2.2.1 An Open and Loose Structure

Figure 2.1 show an open and loose structure.

As we mentioned earlier, many crowd-sourcing websites provide a platform on which their customers can publish varieties of missions and transfer payment to the "workers". These websites help their customers find enough "workers". The "workers" consist of different people who would like to make money in their spare time. Once the workers register online, they can read the requirements of missions and begin to work. They will get paid when the administrators of these websites confirm the accomplishment of the tasks. The structure of this system is loose and sparse. There are little connections between the mission publishers and the workers. Other than the requirements of missions, the workers don't need any other specific instructions from the publishers. In addition, the administrators of these websites are not responsible for the preparation of the requirements. They only maintain the websites and judge the accomplishment of a task.

2.2.2 A Hidden and Tight Structure

Except for those public websites, there are undercover organizations which have closer connections within their members. These organizations help some companies hype their products in a hidden approach. When a mission is released, an organization structure as shown in Figure 2.2 is usually formed. In [35], the authors have a similar

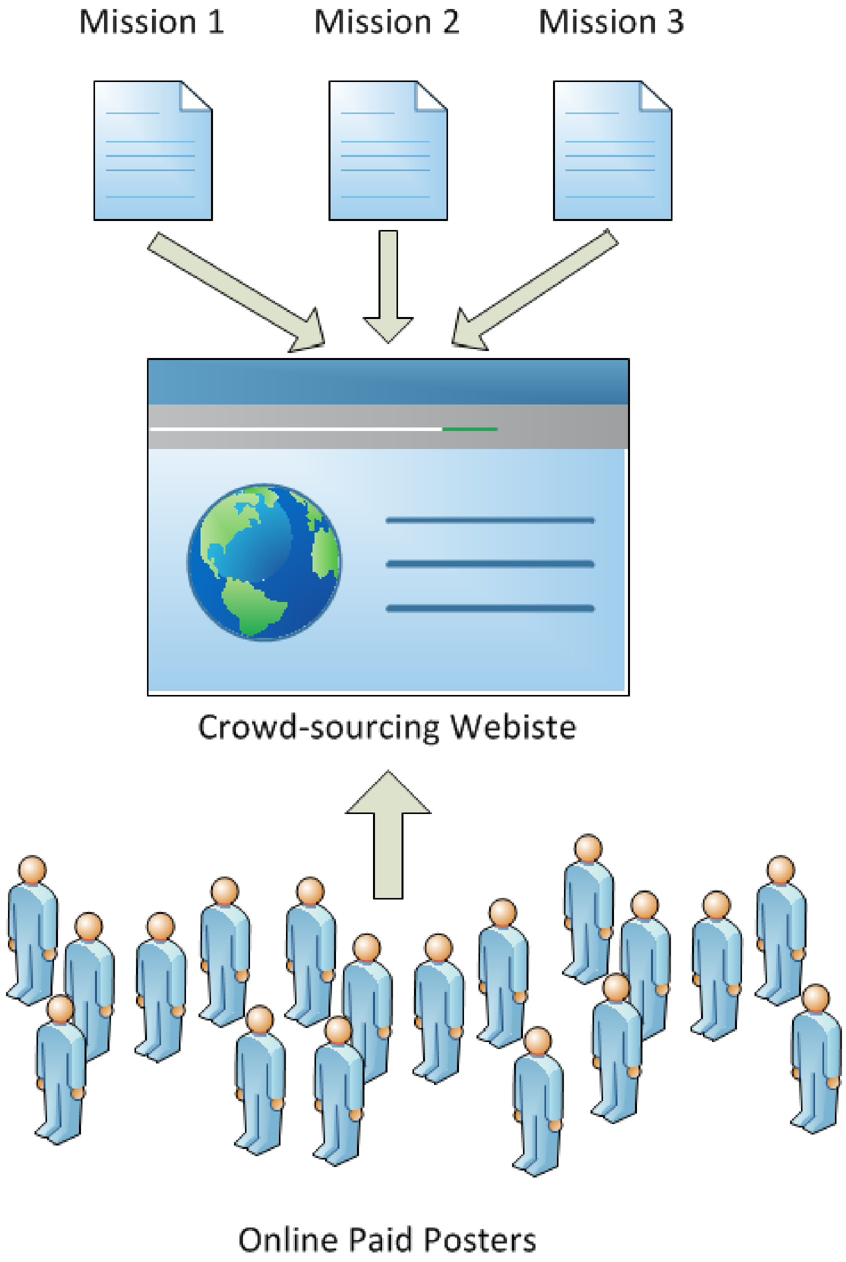

Figure 2.1: Sparse management structure of online paid posters

discovery of the structures of online paid posters (they use the word "*crowdturfing*" in their work).

Below, we describe the function of each role in the structure.

- ***Mission*** represents a potential online event to be accomplished by online paid

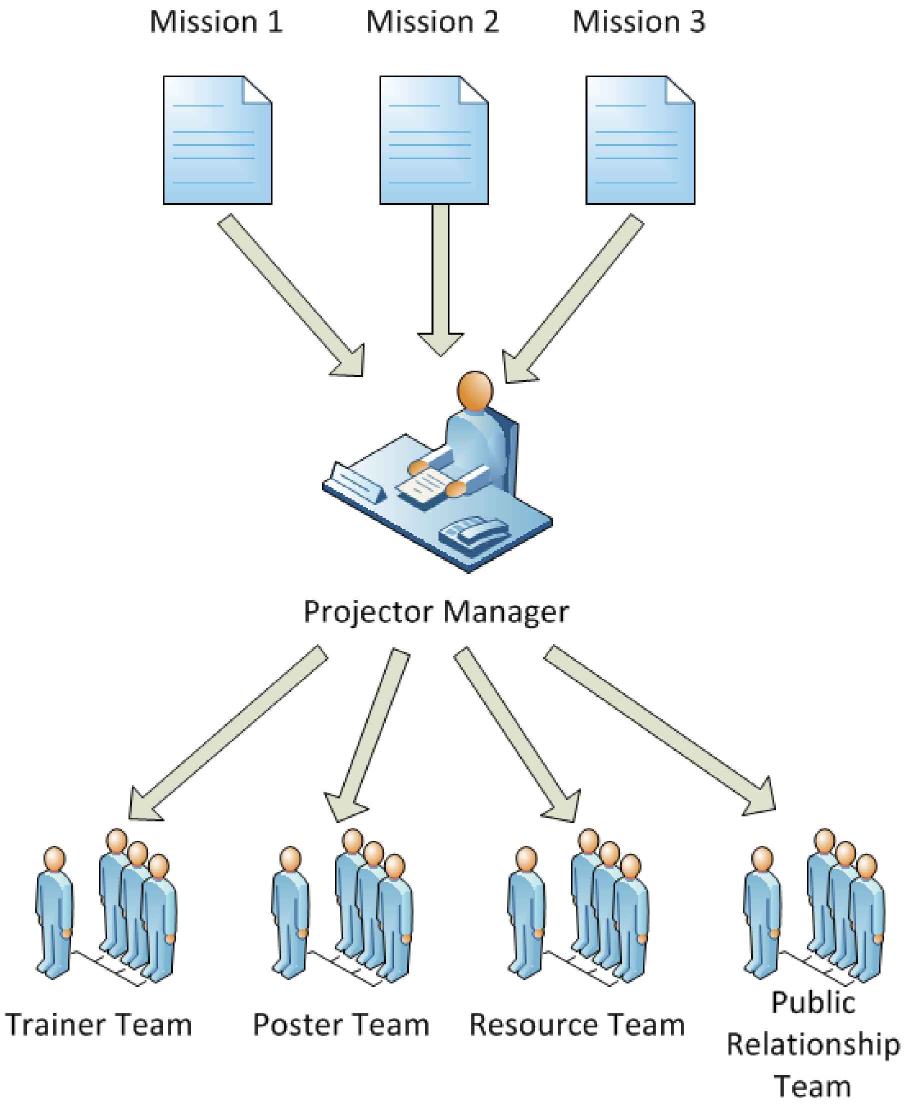

Figure 2.2: Tight management structure of online paid posters

posters. Usually, 1 project manager and 4 teams, namely the trainer team, the poster team, the public relationship team, and the resource team are assigned to a mission. All of them are employed by PR companies.

- ***Project manager*** coordinates the activities of the four teams throughout the whole process.

- ***Trainer team*** plans schedule for paid posters, such as when and where to post and the distribution of shared user IDs. Sometimes, they also accept feedback from paid posters.

- ***Posters team*** includes those who are paid to post information. They are often college students and unemployed people. For each validated post, they get 30 cents or 50 cents. The posters can be grouped according to different target websites or online communities. They often have their own online communities for sharing experience and discussing missions.

- ***Public relationship team*** is responsible for contacting and maintaining good relationship with other webmasters to prevent the posted messages from being deleted. Possibly, with some bonus incentives, these webmasters may even highlight the posts to attract more attention. In this sense, those webmasters are actually working for the PR companies. It is worthwhile to mention here that webmasters may not be willing to cooperate, and even if so, they cannot replace the role of online paid posters.

- ***Resources team*** is responsible for collecting/creating a large amount of online user IDs and other registration information used by the paid posters. Besides, they employ good writers to prepare specific post templates for posters.

Examples of this structure are the companies who claim they can hype the popularity of the applications on the App Store. A typical procedure of a successful undercover marketing activity is as follows.

Suppose that a company which develops a new game would like to deploy it on the App Store. The manager of the hidden organization works out a detailed marketing plan with the developers. The plan is based on the characteristics of the game offered by this company.

According to this plan, leaders of the four teams distribute corresponding tasks among their team members. To remain undercover, the members in these hidden organizations will not publish the missions online. They contact each other by using instant message clients or sending emails. In order to increase the popularity of the game, the basic strategy is to increase the number of downloads of the application. For resources, the promotion company often has thousands of Apple IDs to be used to download the application and write comments. Since the Apple also has its own monitor system, the trainer team needs to carefully select user IDs and choose the appropriate time interval to download the application. In this situation, the poster team are responsible to write comments for the application. This work can be completed either by robots or by real humans. The posters log in the App Store using different user IDs and keep changing their IP address to avoid being blocked by Apple. The public relationship team inspects the ranking of the game on the App Store. They will notify the trainer team to adjust the plan based on the real-time ranking. This is also a strategy to bypass the Apple's censorship.

Once the game achieves the target ranking on App Store, the hidden organization needs to take some actions to maintain its ranking. If the game's ranking is high enough, normal users might be attracted to download this game. Therefore, the hidden organization can reduce the artificially downloading and commenting.

Chapter 3

Data Collection and Labeling

3.1 Data Collection

In this chapter, we use the second typical case introduced in Chapter 2, the conflict between *360* and Tencent, as the case study. We collected news reports and relevant comments regarding this special social event. While the number of websites hosting relevant content is large, most posts could be found at two famous Chinese news websites: Sina.com [30] and Sohu.com [31], from which we collected enough data for our study. We call the data collected from Sina.com *Sina dataset* and will use it as the training data for our detection model. The data collected from Sohu.com is called *Sohu dataset* and it will be used as the test data for our detection method.

We searched all news reports and comments from Sina.com and Sohu.com over the time period from September 10, 2010 to November 21, 2010. As a result, we found 22 news reports in Sina.com and 24 news reports in Sohu.com. For each news report, there were many comments. For each comment, we recorded the following relevant information: *Report ID*, *Sequence No.*, *Post Time*, *Post Location*, *User ID*, *Content*, and *Response Indicator*, the meanings of which are explained in Table 3.1.

We were faced with several hurdles during the data collection phase. At the outset, we had to tackle the difficulty of collecting data from dynamic web pages. Due to the application of AJAX [27] on most websites, comments are often displayed on web pages generated on the fly, and thus it was impossible to retrieve the data from the source code of the web page. To be specific, after the client Internet explorer successfully downloads a HTML page, it needs to send further requests to the server to get the comments, which should be shown in the comment section. Most of the web

Table 3.1: Recorded information for each comment

Field	Meaning
Report ID	The ID of news report that the comment belongs to
Sequence No.	The order of the comment w.r.t. the corresponding news report
Post Time	The time when the comment is posted
Post Location	The location from where the comment was posted
User ID	The user ID used by the poster
Content	The content of the comment
Response Indicator	Whether the comment is a new comment or a reply to another comment

crawlers that retrieve the source code do not support such a functionality to obtain the dynamically generated data. To avoid this problem, we adopted Gooseeker [9], a powerful and easy-to-use software suitable for the above task. It allows us to indicate which part of the page should be stored in the disk and then it automatically goes through all the comment information page by page. In our case study, due to the popularity and the broad impact of this social event, some news reports ended up with more than 100 pages of comments, with each page having 15 to 20 comments. We stored all the comments of one web page in a XML file. We then wrote a program in Python to parse all files to get rid of the HTML tags. We finally stored all the required information in the format described in Table 3.1 into two separate files depending on whether the comments were from Sina.com or from Sohu.com.

We then needed to clean up the datasets caused by some bugs on the server side of Sina and Sohu. We noticed that the server occasionally sent duplicate pages of comments, resulting in duplicate data in our dataset. For example, for a certain report, we recorded more than 10, 000 comments, with nearly 5, 000 duplicate comments. After removing the duplicate comments, we got 53, 723 records in Sina and 115, 491 records in Sohu. There was a special type of comments sent by mobile users with mobile phones. The user IDs of mobile users, no matter where they came from, were all marked as “Mobile User” on the website, which meant this ID, “Mobile User”, represented a group of users who were using mobile phones to send their comments. There was no way to tell how many users were actually behind this unique user ID. For this reason, we had to remove all comments from “Mobile User”. We also needed to remove users who only posted very few comments, since it was difficult to tell whether they were normal users or paid posters, even with manual investigation. To this end, we removed those users who only posted less than 4 comments. Finally, due to the fact that Sohu allowed anonymous posts (i.e., a user can post comments without a user ID), we could no longer keep track of individual user’s record in the dataset. Since the real number of users behind the anonymous posts was unknown, we excluded these anonymous posts from our dataset. So far, we are unclear on how to deal with the situation that behind a user ID (i.e., “Mobile User” or “Anonymous”), there are many normal users and potentially some paid posters. We leave it as an open challenge.

After the above steps, our Sina dataset includes 552 users and 20, 738 comments, and our Sohu dataset has 223 users and 1, 220 comments. It is very interesting to see that the two datasets seem to have largely different statistical features, e.g., the average number of comments per user in the Sina dataset is about 37.6 while that in the Sohu dataset is only 5.5. One main reason is that Sohu allows anonymous posts, while Sina does not. Therefore, many anonymous comments have been removed.

3.2 Manual Identification

In order to analyse the behavioral pattern and classify potential paid posters and normal users, we need to find out the ground truth in the two datasets. We used the following (intelligent) criteria in our manual identification of potential paid posters:

1. Users who post meaningless or contradicting comments. For example, the comments are not even slightly related to the topic in discussion. Also, a user may post multiple comments showing completely different opinions.

2. Users who post many short comments without any support. For example, short comments like "I like 360" and "360 is good" are less likely from reasonable users involved in serious discussion.

3. Users who post negative and irrational comments to attack other persons.

4. Users who post multiple duplicate or near duplicate comments. Unlike the above three behaviors, we do not consider it as a critical criterion in labeling the datasets because both potential paid posters and normal users can have this behavior. Before making final decision, users with this behavior are carefully considered together with other criteria.

Note that the criteria above are based on the real experience working as a paid poster and also based on human intelligence[1]. As a result, we manually selected 70 potential paid posters from the Sina dataset and identified 82 potential paid posters from the Sohu dataset.

We use the word *potential* to avoid the non-technical argument about whether a manually selected paid poster is really a paid poster. Any absolute claim is not possible unless a paid poster admits to it or his employer discloses it, both of which are unlikely to happen.

[1] To avoid biased judgement from one person, we spent a significant amount of time to independently identify the paid posters and then worked together for the final decision.

Nevertheless, we are confident about our labels, as we believe any reasonable person will agree that a user who posts seven “I hate 360” within 2 minutes should be a potential paid poster; and any reasonable person will also agree that a user who posts both “I really like 360 because it protects my computer so well” and “It is really bad that 360 steals my private information. I hate 360” seems a potential paid poster. We stress that when we manually labeled our datasets, we read the contents of a user’s comments. The meaning can be understood by human but is hard to use in machine learning based classification and clustering.

Finally, we note that substantial efforts from the research community have been made to find out the ground truth. Many of them use cross-checking among multiple annotators, as what we have done in this work. One extreme way is to hire paid posters to post fake comments and collect the corresponding texts. This method was used by Ott et al., who worked on a related (but different) problem and obtained “gold-standard” labels by using Amazon Mechanical Turk (AMT) to hire turkers to post fictitious hotel reviews [26]. Nevertheless, even with such a costly method, it is difficult to obtain “gold-standard” labels. For example, in [26], it may be helpful that Ott et al. have collected articles from the paid posters (turkers), but this does not necessarily mean they have obtained the ground truth, because they have no guarantee that posts not from their hired tuckers are truthful.

Chapter 4

Analysis of Non-Semantic Features

After manually identifying the potential paid posters, we perform statistical analysis to investigate objective features that are useful in capturing the potential paid posters' special behavior. We use Sina dataset as our training data and thus we only perform statistical analysis on this dataset. We mainly test the following four features: percentage of replies, average interval time of posts, the number of days the user remains active and the number of news reports that the user comments on. In the following figures, we use "pp" and "nu" to denote potential paid posters and normal users, respectively.

4.1 Percentage of Replies

In this feature, we calculate the probability whether a user tends to post new comments or reply to others' comments. We conjecture that potential paid posters may not have enough patience to read others' comments and reply. Therefore, they may create more new comments.

Figure 4.1 and Figure 4.2 show a rough comparison of percentage of replies, where p represents the ratio of number of replies over the number of total comments from the same user. Based on the results, 84.3% potential paid posters have less than 50% of posts being replies. In contrast, most normal users (73.2%) posted more replies than new comments.

Figure 4.3 shows the statistical results, with respect to the density and cumulative density function of reply ratio. Our results show that potential paid posters tend to have smaller reply ratio. From the figure of cumulative density function, over 70%

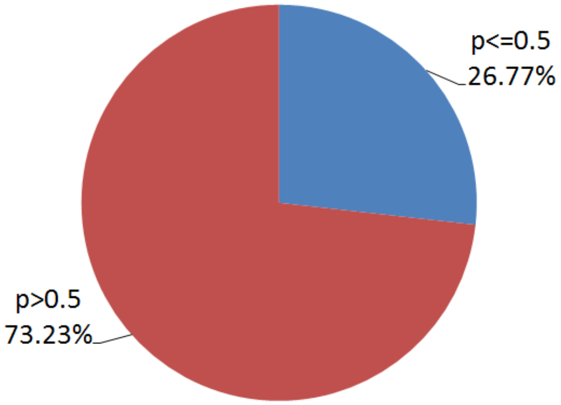

Figure 4.1: The percentage of replies from normal users

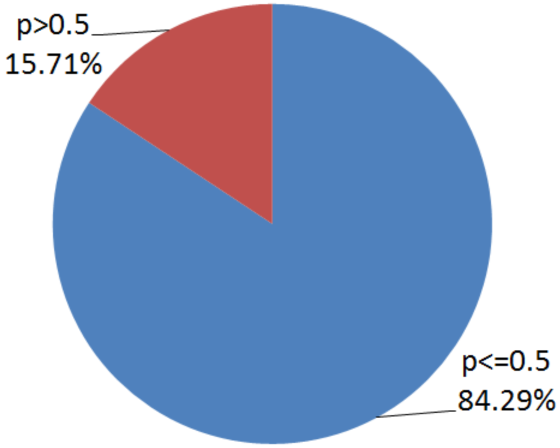

Figure 4.2: The percentage of replies from potential paid posters

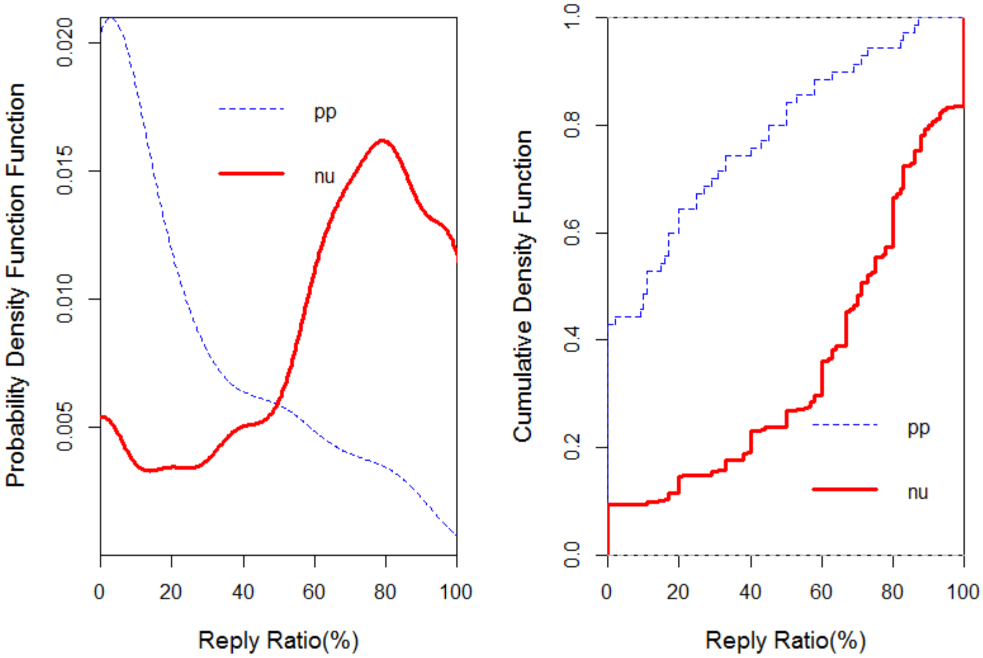

Figure 4.3: The PDF and CDF of reply ratio

potential paid users have reply ratio smaller than 40%. In contrast, over 70% normal users have reply ratio up to 90%, indicating that most normal users post more replies than creating new comments. This observation confirms our conjecture that potential paid posters are more likely to post new comments instead of reading and replying to others' comments.

4.2 Average Interval Time of Posts

We calculate the average interval time between two consecutive comments from the same user. Note that it is possible for a user to take a long break (e.g., several days) before posting messages again. To alleviate the impact of long break times, for each user, we divide his/her active online time into epochs. Within each epoch, the interval time between any two consecutive comments cannot be larger than 24 hours. We calculate the average interval time of posts within each epoch, and then take the average again over all the epochs.

Intuitively, normal users are considered to be less aggressive when posting comments while paid posters care more about finishing their jobs as soon as possible. This implies that the average interval time of posts from paid posters should be smaller. Figure 4.4 and Figure 4.5 show the corresponding proportion graphs. Figure 4.6 shows the statistical results for the probability distribution of interval posting time.

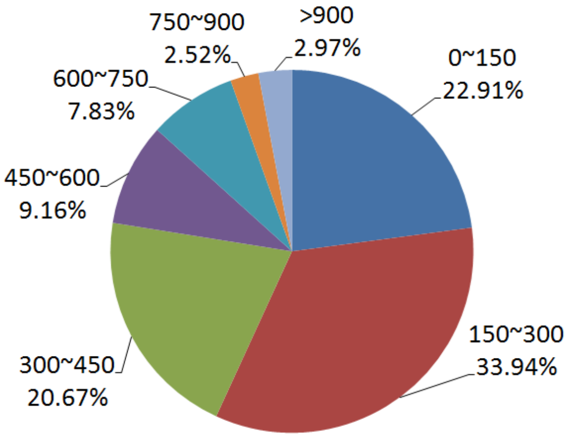

Figure 4.4: The average interval time of posts from normal users

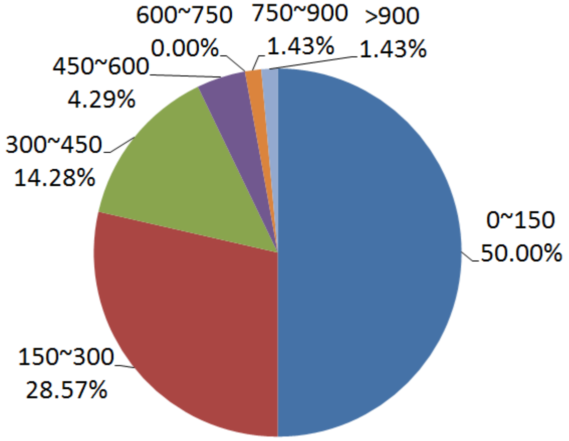

Figure 4.5: The average interval time of posts from potential paid posters

Based on the figures, 60% potential paid posters post comments within interval time of 200 seconds while only 40% of normal users post at such speed. The difference can be easily detected in the cumulative probability plot. This is consistent with our intuition that paid posters only care about finishing their jobs as soon as possible and do not have enough interest to get involved in the online discussion.

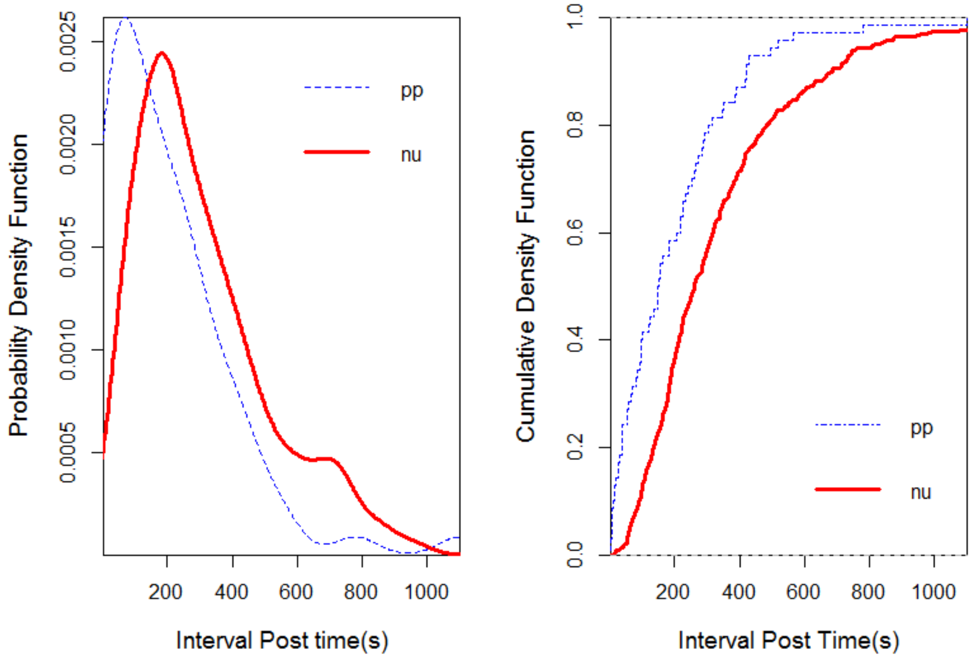

Figure 4.6: The PDF and CDF of average interval time

We observed that some potential paid posters also post messages in a relatively slow speed (the interval time is larger than 1000 seconds). There is one main explanation for the existence of these “outliers”. As mentioned earlier, the *trainer team* may enforce rules that the paid posters need to follow. For example, identical replies should not appear more than twice in a same news report or within a short time period. Such rules are made to keep the paid posters from being detected easily. If a paid poster follows these tactics, he/she may have a statistical feature similar to that of a normal user. Nevertheless, it seems that the majority of potential paid posters did not follow the rules strictly.

4.3 Active Days

We analyze the number of days that a user remains active online. This information can be extracted from the timestamp of their comments. We divide the users into 7 groups based on whether they stay online for $1, 2, 3, 4, 5, 6$ days and more than 6 days, respectively. Potential paid posters usually do not stay online using the same user ID for a long time. Once a mission is finished, a paid poster normally discards the user ID and never uses it again. When a new mission starts, a paid poster usually uses a different user ID, which may be newly created or assigned by the *resource team.* Figure 4.7 and Figure 4.8 show the corresponding proportion graphs. Figure 4.9 shows the statistical result. In the figures, "7" at the x-axis is the number of active days for 7 days or more.

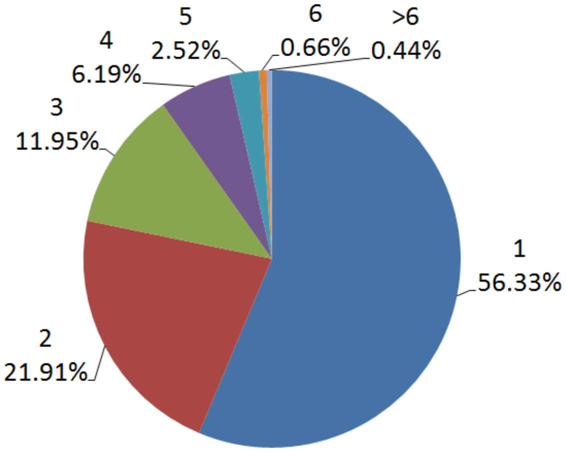

Figure 4.7: The number of active days of normal users

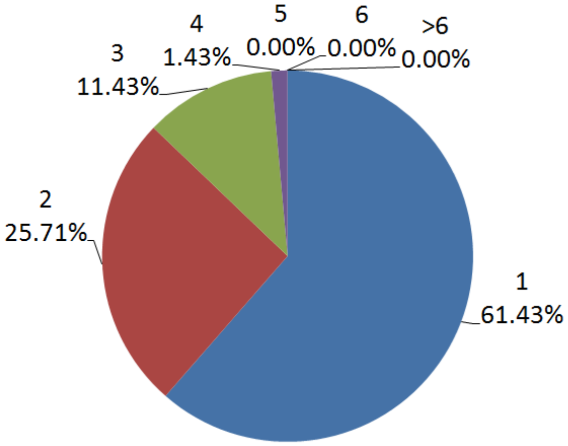

Figure 4.8: The number of active days of potential paid posters

According to statistical result, the percentage of potential paid posters and the percentage of normal users are almost the same in the groups that remain active for $1, 2, 3$ and 4 days. Nevertheless, the solid line representing normal users has a longer

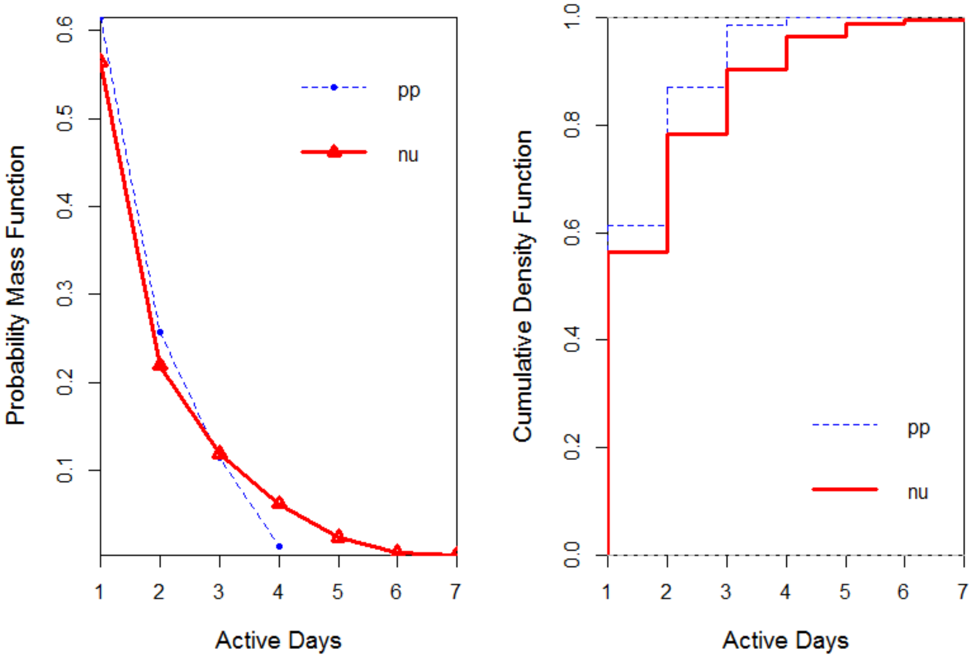

Figure 4.9: The PMF and CDF of number of active days

tail which indicates that some normal users keep taking part in the discussion for 5 or more days, while we find no potential paid posters stay for more than 4 days. This evidence suggests that potential paid posters are not willing to stay for a long time. They instead tend to accomplish their assignments quickly and once it is done, they would not visit the same website again.

4.4 The Number of News Reports

We study the number of news reports for which a user has posted comments. Both Sina and Sohu have nearly 20 news reports. Figure 4.10 and Figure 4.11 show the corresponding proportion graphs. Figure 4.12 shows the corresponding graphs.

According to the result, the potential paid posters and normal users have similar distribution with respect to the number of commented news reports. We originally conjectured that paid posters might have a larger number of news reports that they comment to. While normal users may not be interested in reports that are not well

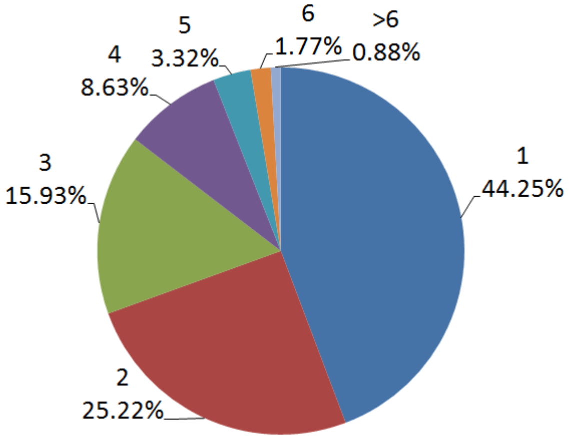

Figure 4.10: The number of news reports that a normal user has commented

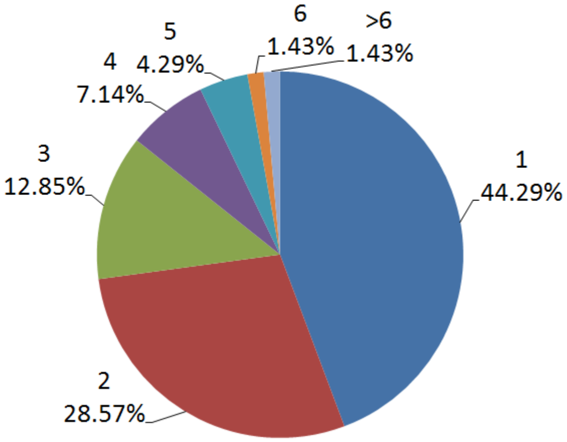

Figure 4.11: The number of news reports that a potential paid posters has commented

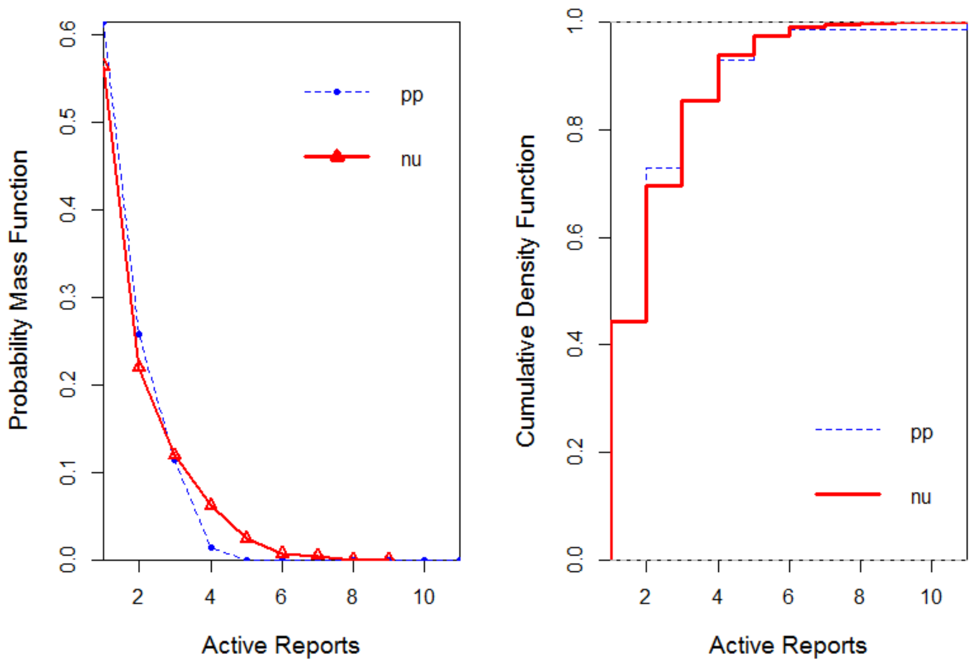

Figure 4.12: The PMF and CDF of number of active news reports

written or not interesting, paid posters care less about the contents of the news. Nevertheless, we did not find strong evidence to support this conjecture in the Sina dataset. This indicates that the number of commented news reports alone may not be a good feature for the detection of potential paid posters.

4.5 Other Observations

We also discuss other possible features of potential paid posters. These observations come from our working experience as a paid poster. Although we cannot find sufficient evidence in the Sina dataset, we discuss these features as they can be beneficial for future research on this topic.

First, there may be some pattern in geographic distribution of online paid posters. We performed statistical study on the Sina dataset, but found that both normal users and potential paid posters are mainly located in the center and the south regions of China. While the two companies involved in the event, Tencent and *360*, are located in the province of Guang'dong and Beijing, respectively, we found no relationship between the locations of potential paid posters and the locations of the two companies. Figure 4.13 shows the geographic distribution of normal users and potential paid posters, with the darker color representing more users.

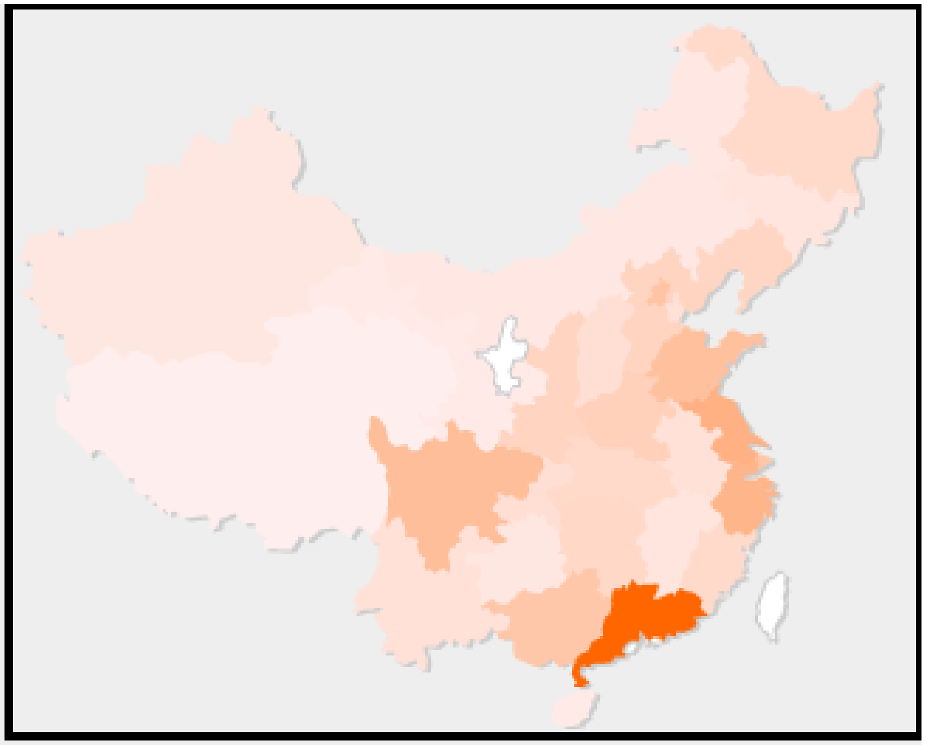

Figure 4.13: The geographical distribution of normal users

According to Figure 4.13 and Figure 4.14, provinces of Guang'dong and Jiang'su have the majority of potential paid posters while Guang'dong has the most normal users. Unfortunately, we didn't see clear relationship between location and paid posters except Guang'dong and Jiang'su.

Second, the same user ID appears at different geographical locations within a very short time period. This is a clear indication of paid poster. Normal users are not able to move to a different city in a few minutes or hours, but paid posters can because their user IDs may be assigned dynamically by the *resource team*. We have identified this possible feature during analysis but could not find sufficient evidence in the Sina dataset.

Third, there might be contradicting comments from paid posters. The reason is that they are paid to post without any personal emotion. It is their job. Sometimes, they just post comments without carefully checking their content. Nevertheless, this

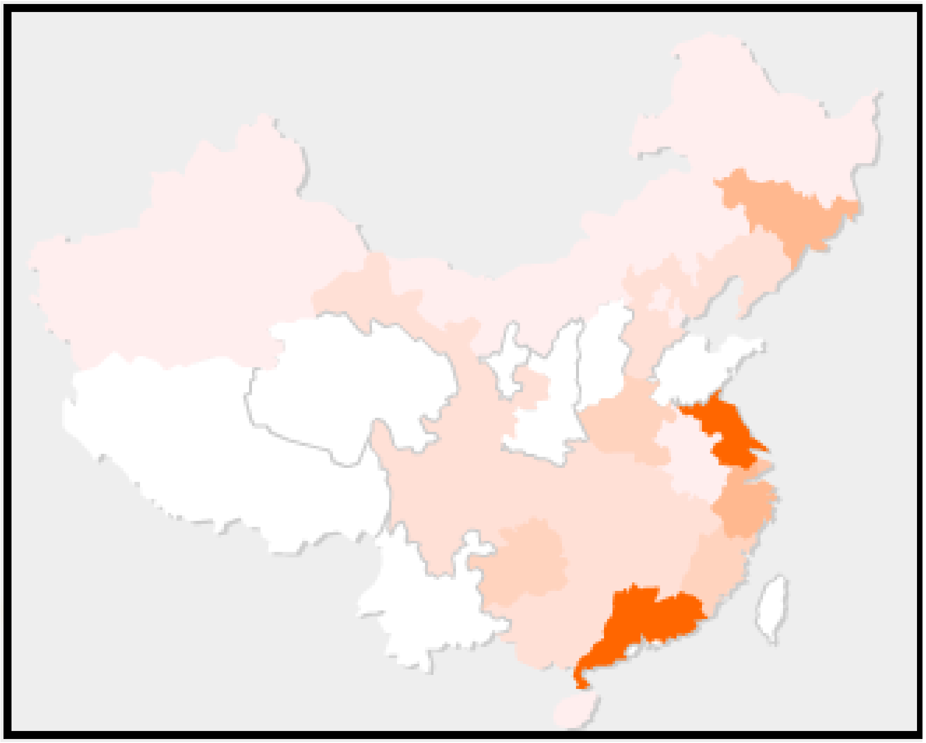

Figure 4.14: The geographical distribution of potential paid posters

feature requires the detection system to have enough intelligence to understand the meaning of the comments. Incorporating this feature into the system is challenging.

Fourth, paid posters may post replies that have nothing to do with the original message. To earn more money, some paid posters just copy and paste existing posts and simply click the *reply* button to increase the total number of posts. They do not really read the news reports or others' comments. Again, this feature is hard to implement since it requires high intelligence for the detection system.

Chapter 5

Analysis of Semantic Features

5.1 Overview

An important criterion in our manual identification of a potential paid poster is to read his/her comments and make a choice based on common sense and online experience. For example, if a user posts meaningless messages or messages contradicting each other, the user is very likely to be a paid poster. Nevertheless, it is very hard to integrate such human intelligence into a detection system. In this section, we propose a simplified semantic analysis method that is demonstrated to be very effective in detecting potential paid posters.

While it is hard to design a detection system that understands the meaning of a comment, we observed that potential paid posters tend to post similar comments on the web. In many cases, a potential paid poster may copy and paste existing comments with slight changes. This provides the intuition for our semantic analysis technique. Our basic idea is to search for similarity between comments.

5.2 Word Splitting

To do this, we first need to overcome the special difficulty in splitting a Chinese sentence into words and phrases. Unlike English sentences that have a space between words, many languages in Asia such as Chinese and Japanese depend on context to determine words. They do not have space between words and how to split a sentence is left to the readers. We used a famous Chinese splitting software, called ICTCLAS2011 [13], to cut a sentence into words. For a given sentence, the software outputs its content words and stop words [6]. Simply put, content words are words that have an independent meaning, such as noun, verb, or adjective. They have a stable lexical meaning and should express the main idea of a sentence. Stop words are words that do not have a specific meaning but have syntactic function in the sentence to make it grammatically correct. Stop words thus should be filtered out from further processing.

5.3 Similarity Calculation

The above step translates a sentence into a list of content words. For a given pair of comments, we compare the two lists of content words. As mentioned before, a paid poster may make slight changes before posting two similar comments. Therefore, we may not be able to find an exact match between the two lists. We first find their common content words, and if the ratio of the number of common content words over the length of the shorter content word list is above a threshold value (e.g., 80% in our later test), we conclude that the two comments are similar. If a user has multiple pairs of similar comments, the user is considered a potential paid poster. Note that similarity of comments is not transitive in our method.

We found that a normal user might occasionally have two *identical* comments. This may be caused by the slow Internet access, due to which the user presses the *submit* button twice before his/her post is displayed. Our manual check of these users confirmed that they are normal users, based on the content they posted. To reduce the impact of the "unusual behavior of normal users", we set the threshold of similar pairs of comments to 3. This threshold value is demonstrated to be effective in addressing the above problem.

While there are many other complex semantic analysis methods to represent the similarity between two texts [18, 10, 19], we believe that comments are much shorter than articles and therefore a simple method as above would be good enough. The performance is demonstrated later in Chapter 6.

5.4 Comparison of Results

We performed the semantic analysis over the Sina dataset. Figure 5.1 and Figure 5.2 are the pie graphs showing the proportion. Figure 5.3 shows the statistical results for the probability distribution of similar pairs of comments. In the figure, "6" on the x-axis means the number of similar pairs is larger than or equal to 6. The two groups of users obviously show different patterns. Normal users have much higher probability to post different comments. In the opposite, the potential paid posters have many similar pairs of comments in their profiles. Therefore, it is important to monitor the number of similar pairs of comments in a user's profile as it is a significant indication of malicious behavior.

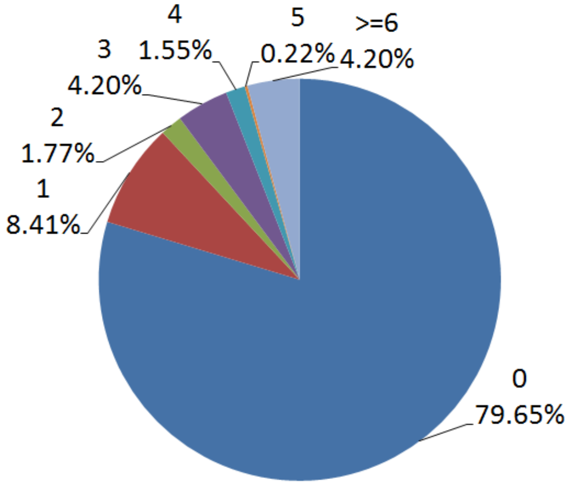

Figure 5.1: The number of similar pairs of comments posted by normal users

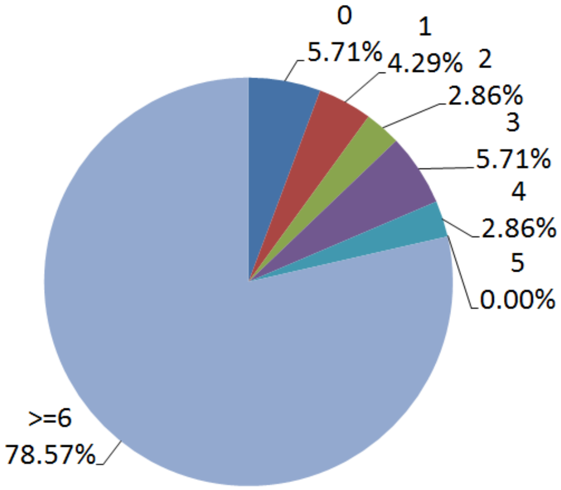

Figure 5.2: The number of similar pairs of comments posted by potential paid posters

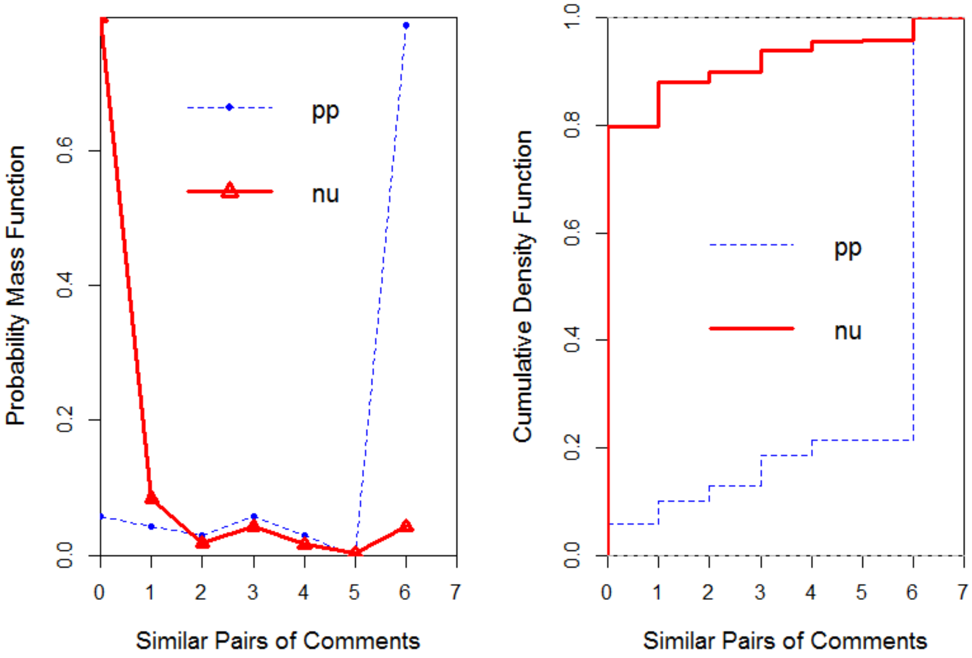

Figure 5.3: The PMF and CDF of the number of similar pairs of comments

Chapter 6

Detection Method and Detection Results

6.1 Classification

The objective of our classification system is to classify each user as a potential paid poster or a normal user using the features investigated in Chapter 4 and Chapter 5. According to the statistical and semantic analysis results, we found that any single feature is not sufficient to locate potential paid posters. Therefore, we use and compare the performance of different combinations of the five features discussed in the previous two sections in our classification system. We model the detection of potential paid posters as a binary classification problem and solve the problem using a support vector machine (SVM) [5].

SVM is a one of the effective supervised machine learning methods that are used for both classification and regression problems. SVM can output predication based on training datasets. To have it do the task, the first step is to collect datasets of enough samples. The samples including their corresponding feature vectors and labels. The labels can be either discrete numbers in classification problems or continuous numbers in regression problems. In our case, the labels are discrete numbers indicating different categories of users. The SVM takes the datasets and attempt to find appropriate parameters for a model that can divide the samples as wide as possible. After that, the model can be used to predicate the labels of new samples.

We used LIBSVM [3] as the tool for training and testing. By default, LIBSVM adopts a radial basis function [5] and a 10-fold cross-validation method to train the data and obtain a classifier. The Sina dataset is divided into 10 subsets of equal size. Then the model is trained on the 9 subsets and tested on the remaining subset. The process returns a model with the highest cross-validation accuracy. After training the classifier with the Sina dataset, we used the classifier to test the Sohu dataset.

We evaluate the performance of the classifier using the four metrics: *precision*, *recall*, *F-measure* and *accuracy*, which are defined in Table 6.1. Note that these four metrics are well known and broadly used measures in the evaluation of a classification system [32]. In the table, *benchmark result* means the result obtained with manual identification of potential paid posters.

Finally, we also use the K-means algorithm to perform unsupervised learning on the merged dataset of Sina and Sohu. This algorithm groups the samples according to their feature vectors.

Table 6.1: Metrics to evaluate the performance of a classification system

		Classified Result	
		Normal User	Paid Poster
Benchmark Result	Normal User	True Negative	False Positive
	Paid Poster	False Negative	True Positive

$$
\begin{aligned}
Precision &= \frac{TruePositive}{TruePositive + FalsePositive} \\
Recall &= \frac{TruePositive}{TruePositive + FalseNegative} \\
F - measure &= 2 * \frac{Precision * Recall}{Precision + Recall} \\
Accuracy &= \frac{TrueNegative + TruePositive}{TotalNumberofUsers}
\end{aligned}
$$

6.1.1 Classification without Semantic Analysis

To simplify the notation, the five features, *reply ratio*, *average interval posing time*, *active days*, *active reports* and *degree of similarity* are labeled as features “1”, “2”, “3”, “4” and “5”, respectively. The first four features are statistical ones while the last is a semantic feature.

We firstly focus on the classification only using statistical analysis results based on the four statistical features. Different combinations are applied to test their performance for identification. We train the SVM model using the Sina dataset with different combinations of the features. Then we test the model with the Sohu dataset to see the performance. Note that combinations that result in 0 true positive or 0 false positive are not considered.

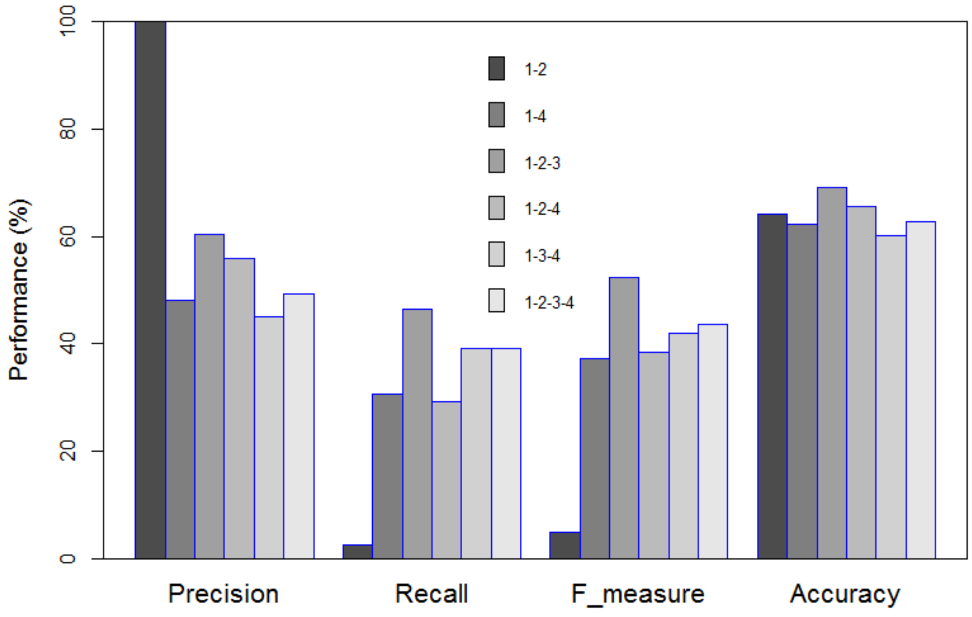

Figure 6.1: The performance of different combinations of statistical features

The results are shown in Figure 6.1. Although the (1-2)-feature test has the highest precision, its recall and f-measure are very low, showing that the (1-2)-feature can hardly separate different classes of users. This result suggests that the first two features lead to significant bias and we need to add more features to our classifier. With features 3 and 4 considered, we observe better performance. For example, the (1-2-3)-feature test has better performance over all the metrics, except precision.

Nevertheless, we notice when we use only non-semantic features to train the SVM model, the overall performance on the four metrics is not good enough to claim acceptable performance. Particularly, the low precision and accuracy results indicate that the SVM classifier using the four non-semantic features as its vector set is unreliable and needs to be improved further. We achieve this by adding the semantic analysis to our classifier.

6.1.2 Classification with Semantic Analysis

As described in Chapter 5, we have observed that online paid posters tend to post a larger number of similar comments on the web. Based on this observation we have designed a simple method for semantic analysis. We test the performance of all the five features. After integrating this semantic analysis method into our SVM model, we observed the much improved performance results as shown in Figure 6.2.

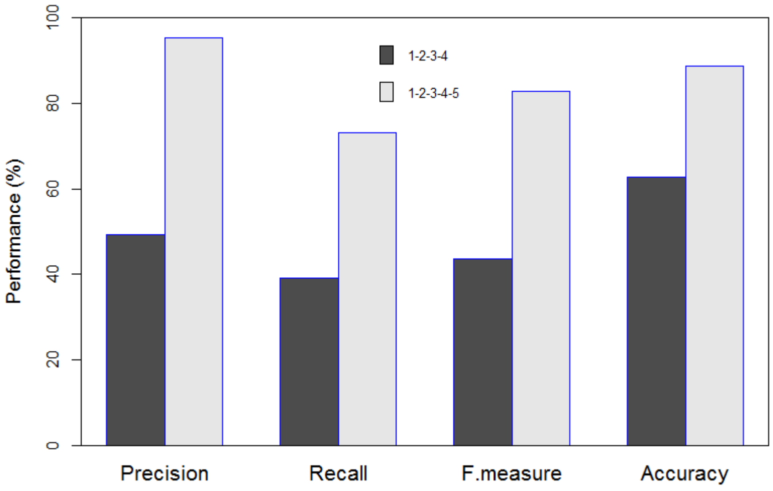

Figure 6.2: The performance of statistical and semantic features

The results clearly demonstrate the benefit of using semantic analysis in the detection of online paid posters. The precision, recall, F-measure and accuracy have been improved by 95.24%, 73.17%, 82.76% and 88.79%, respectively. Based on these improved results, the semantic feature can be considered as a useful and important supplement to other features. The reason why the semantic analysis improves performance is that online paid posters often try to post many comments with some minor changes on each post, leading to similar sentences. This helps the paid posters post many comments and complete their assignments quickly, but also helps our classifier to detect them.

6.2 Test with Unsupervised Learning

SVM, as one of the supervised machine learning techniques, depends on manually labeled datasets. However, manual checking is time consuming. Thus we perform unsupervised clustering on our datasets based on the five features. Unsupervised clustering further validates the effectiveness of the five proposed features and also provides us with a possible approach to developing more efficient detection scheme.

For unsupervised learning, we merged Sina dataset and Sohu dataset and applied K-means clustering algorithm [36] to obtain K clusters. If the five features have the ability to distinguish paid posters from normal users, we expect that paid posters should be grouped into a cluster. In our work, we only need two clusters, one for paid posters and one for normal users. Furthermore, to check the reliability of our features, we studied two more cases, corresponding to $K = 3$ and $K = 4$.

Figure 6.3, Figure 6.4 and Figure 6.5 show the size of each cluster as well as the number of potential paid posters and normal users in each cluster. From the figures, we notice that when $K = 2$, a large proportion (approximately 85%) of potential paid posters is assigned to a particular cluster (cluster 1). When $K = 3$ and $K = 4$, cluster 1 (the group of paid posters) remains stable. Nevertheless, the other cluster (the group of normal users) is further divided into smaller clusters. This phenomenon suggests that although normal users might have different behavioral patterns, they in general behave much different from potential paid posters.

We also notice that a small number of normal users are assigned to cluster 1. This is because our manual labeling uses human intelligence (refer to Chapter 3), which cannot be completely captured by the five features. This poses the challenge of developing more intelligent detection mechanism for our future work.

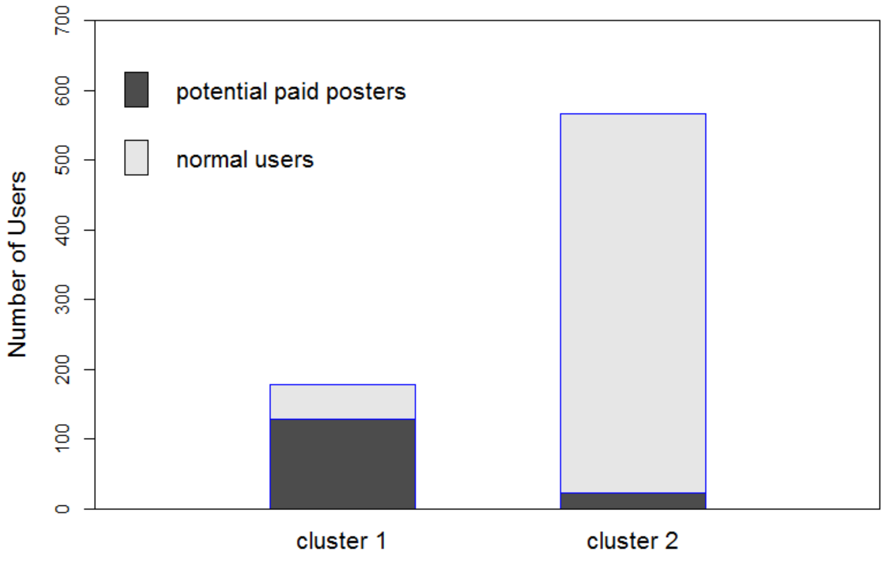

Figure 6.3: Clustering: $K = 2$

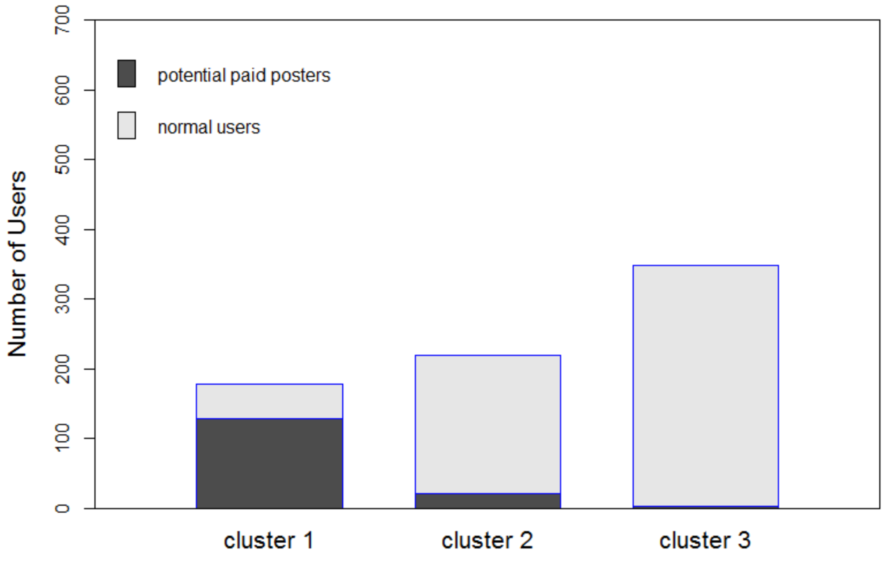

Figure 6.4: Clustering: $K = 3$

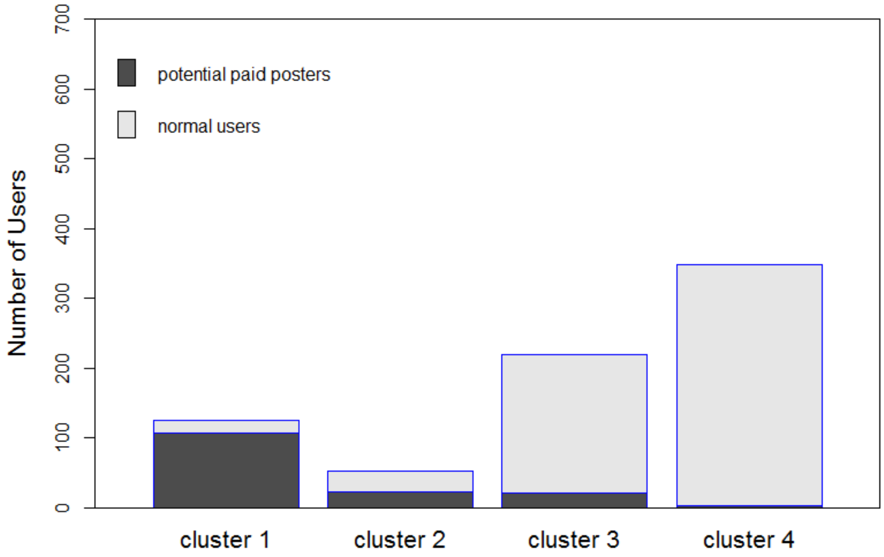

Figure 6.5: Clustering: $K = 4$

Chapter 7

Real-Time Detection System Prototype Design

7.1 Introduction

In this section, we discuss fundamental architecture and design of a detection system that can identify malicious behavior and potential paid posters in real time.

We notice that there are plenty of information retrieval software available. They can help companies to collect online comments, articles and relevant discussion of their products and services. However, most of the software do not have the ability to remove malicious content which we studied in this thesis. Consequently, analysts have to spend large amount of time dealing with the unreliable information. For example, in order to hype a company's popularity, the manager of the company might hire hidden paid posters to publish numerous positive messages in a wide range. These fake customers' online opinions are artificially created with the intent to spam people's opinion. Ordinary people can hardly figure out the truth. Detecting such behavior is often time-consuming and not efficient. On the contrary, the goal of our system is to identify potential paid posters and locate their user IDs during the process of collecting information. This system will automatic collect data from different resources/websites and generate reports of the behavior of potential paid posters. Our system will provide valuable information for the analysts and online users to differentiate on various aspects.

There is some difference between the methods used in the thesis and those to be applied in the real-time system. Recall that we have collected users' comments regarding a fight between two companies. The statistical and semantic analysis conducted in this thesis are based on the large amount of training data. In reality, it often takes several months to get such datasets. It would be less useful if we can only begin to detect the potential paid posters after the event ends for a long time. A popular question on the online social networks is a how to track each user and identify malicious behavior in real time. In other words, a real-time detection system that can analyze people's online behavior using data collection will be much more useful to the public, companies and the administrators of websites.

To be specific, suppose that a popular event happens and many websites publish news reports about it and online users begin to join the discussion on those portals. Suppose we are interested in collecting and analysing online users' comments regarding this popular social event. In the following sections, we discuss how to implement such a system based on this situation.

7.2 Software Architecture and Design

7.2.1 Architecture Overview

The system consists of four major components, data crawler, scheduler, data analyser and database system. Table 7.1 gives a brief introduction of each component of the system. Figure shows the structure of the system.

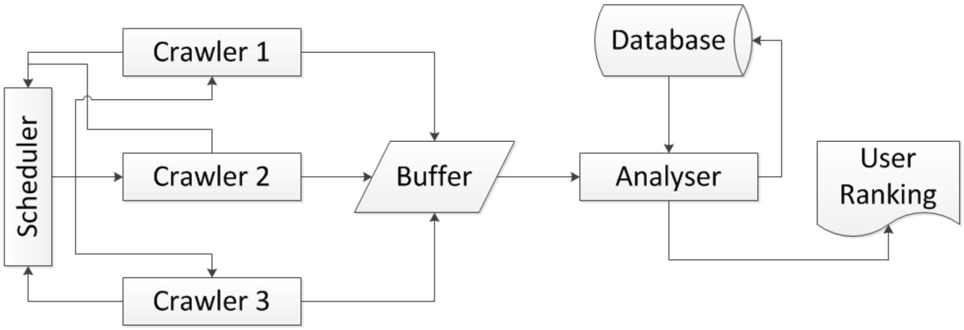

Figure 7.1: Flow chart of the detection system

As shown in the figure, the scheduler module controls multiple crawlers. The crawlers save the collected information to a buffer shared with analyser module. The analyser module gets data from the buffer, updates the users' records and then saves

Table 7.1: Brief introduction of each component

Component	Functionality
Data Crawler	It visits the target webpages and extracts information according to specific format.
Scheduler	It controls the data crawler and distribute the target webpages among different crawler objects.
Analyser	It receives the update information from data crawler and calculate the scores for each user according to the real-time information.
Database	It stores all the information collected from the Internet.

the data into database. It also generates reports regarding the information of each malicious users. We then describe each component in details.

7.2.2 Crawler Module

The crawler module collects raw webpages and extracts formatted data. The scheduler module assigns target websites to the crawler so that the crawler doesn't need to worry about which website it should visit. Figure 7.2 shows the UML class diagram of this module.

Data Crawling

The most important concern for the crawler is how to adjust itself according to the target websites with different techniques and formats. Websites often have different regions and formats for the comments. The attempt to design a generic framework that can handle all the situations is not practical. One way to solve this problem is to provide the crawler with different configuration files for different websites. Before diving into the crawling process, the crawler needs to read the configuration file and adjust itself to successfully locate the corresponding regions of comments and retrieve comments from the target website. In this system, the configuration files will be maintained by the scheduler module.

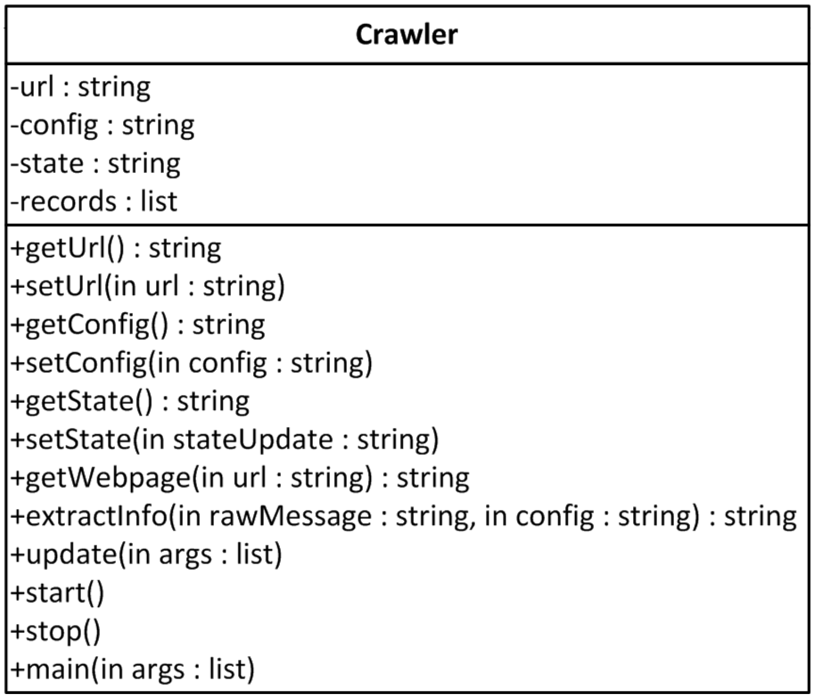

Figure 7.2: The UML class diagram of crawler module

Another instruction which should be included in the configuration files is how to deal with comments placed on multiple webpages. Since websites often have many pages for users' comments, the crawler should visit each of the pages and collect comments on them. If the website assigns a unique url to the "*Next Page*" button, the crawler can simply visit that url. However, if the "*Next Page*" button is associated with a Javascript command, the crawler should be able to recognize the Javascript command and send the request to the server and then retrieve the information from the response data.

Data Extraction

The crawler firstly collects raw data having HTML/XML tags mixed with the desired information. The function, *dataExtraction()*, gets rid of the unrelated signs and tags. It returns each comment in a format described in Table 3.1. When the information of a batch of comments are extracted, *update()* will be used to transfer the comments into a buffer space which is shared with *Analyser module*.

7.2.3 Scheduler Module

The scheduler component starts and monitors the crawlers. It distributes the target websites and URLs among different objects of the crawler. If some of the crawlers get trapped, it should restart the crawlers and report the problem. Figure 7.3 shows the UML class diagram of this module.

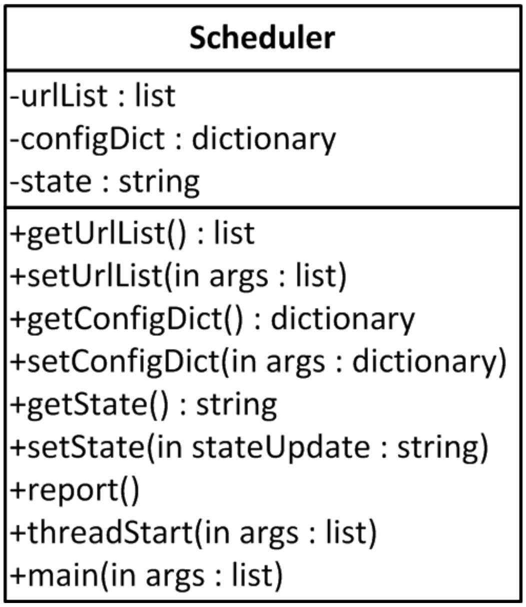
Scheduler
-urlList : list
-configDict : dictionary
-state : string
+getUrlList() : list
+setUrlList(in args : list)
+getConfigDict() : dictionary
+setConfigDict(in args : dictionary)
+getState() : string
+setState(in stateUpdate : string)
+report()
+threadStart(in args : list)
+main(in args : list)

Figure 7.3: The UML class diagram of scheduler module

Controlling the Crawler

The scheduler module initializes the parameters of the crawler, including target url and configuration files. Then it activates the thread of the crawler and monitors the status of the crawler. If it catches some errors returned by the crawler, the scheduler module should report the issue and might need to restart the crawler. If there are multiple news reports, the scheduler module can create multiple threads of the crawler, each of them assigned an unique url and configuration file.

Maintaining the Configuration Files

The configuration files are stored on the disk. The administrator of this system can access those files and update the rules and parameters. Since the websites may change their page formats frequently, the administrator needs to analyze the structure of the webpages and update the corresponding configuration files.

7.2.4 Analyser Module

The analyser module receives the list of newly collected information from a buffer space shared with the crawlers. These information are used to update the user information so that we can track each user's trail in real time. Figure 7.4 shows the UML class diagram of this module.

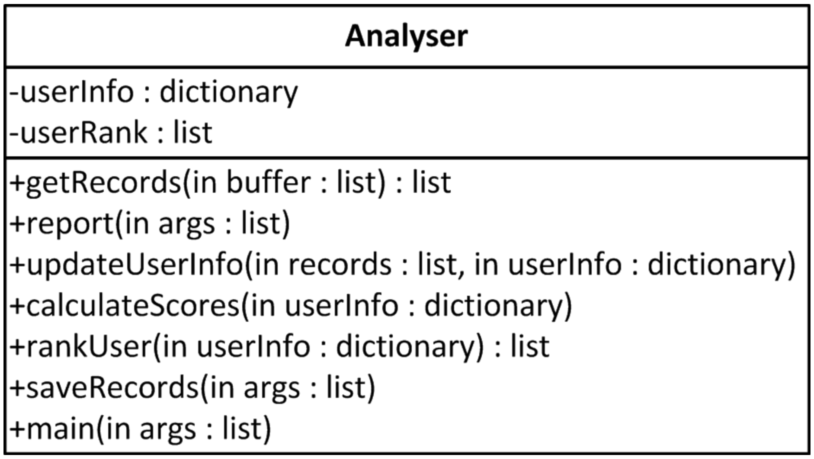
Analyser
-userInfo : dictionary
-userRank : list
+getRecords(in buffer : list) : list
+report(in args : list)
+updateUserInfo(in records : list, in userInfo : dictionary)
+calculateScores(in userInfo : dictionary)
+rankUser(in userInfo : dictionary) : list
+saveRecords(in args : list)
+main(in args : list)

Figure 7.4: The UML class diagram of analyser module

This system should be working in a real-time manner. It analyzes the data on the fly.

Maintaining the User List

The analyser module maintains a list of each user's information, including *reply ratio*, *average interval posting time*, *active days*, *active reports* and *average similarity between comments*. When this module receives newly collected information from the buffer space shared with the crawlers, it updates the value of those metrics. To reduce bias, the average interval posting time, active days and active reports will be regularized by the corresponding maximum value among all the numbers. Then we calculate an average number over the five metrics. We expect that higher value indicates that the user is more likely to be normal ones while lower value reveals malicious behavior.

Generating Reports and Visualization

Once the scores for each of the users are obtained, the analyser module ranks the users according to their scores and output a report. These information can be used to filter the malicious messages hidden in the comments. In addition, it can output graphics showing the development of users' behavior so that the administrator of this system can have better understanding of the comment trends.

7.2.5 Databases

It stores all the information collected by the crawler as well as the processed data by Analyser. The information includes both normal comments and malicious ones.

7.3 Other issues

Now, we describe several challenges when implementing this system.

1. AJAX is widely used to display the comments. Although the crawler can execute the corresponding Javascript command and communicate with the server, it is sometimes difficult to locate such commands. To bypass this difficulty, we can manipulate a web browser which has built-in Javascript engine. Examples are the IE-core browsers or the webkit-core browsers. When loading a webpage, the browser should wait until Javascript engine finishes generating the dynamic content. In this way, the crawler can control the browser to get the dynamically generated HTML content.

2. Before generating any interesting results, this system still needs a warming-up stage to accumulate enough data. Users with very small amount of data cannot be analyzed with confidence. The administrator of this system should monitor the change of the collected information and decide when to activate the analyser module.

Chapter 8

Conclusion

8.1 Conclusion of this Thesis

Detection of paid posters behind social events is an interesting research topic and deserves further investigation. In this thesis, we disclose the organizational structure of paid posters. We also collect real-world datasets that include abundant information about paid posters. We identify their special features and develop effective techniques to detect them. The performance of our classifier, with integrated semantic analysis, is quite promising on the real-world case study, as confirmed in both supervised learning and unsupervised learning techniques.

This work is our preliminary effort to battle online paid posters. It requires a prolonged and systematic effort to reach a complete solution, as the online paid posters evolve continuously and present new challenges to the detection mechanism. As the future work, we will further improve our detection system and evaluate the system in a broader and larger dataset.

8.2 Future Work

8.2.1 Implementation

Based on the description of the real-timbe detection system in Chapter 7, we plan to develop an intelligent system and deploy it for commercial use.

8.2.2 Detecting Other Types of Opinion Spam

This thesis mainly focused on social comments, from which we detected potential paid posters. In the future, we also plan to broaden our research to cover other types of opinion spam. One such direction is to evaluate the credibility of the community question and answer portals (CQA), like Yahoo! Answers [1] and Baidu Zhidao [39]. These websites offer a place where online users can ask and answer questions. However, online paid posters can artificially create Q&A sessions to promote products. They disguise themselves as normal users to ask and write suggestions. When reading those malicious Q&A sessions, others often cannot identify the fake information. So our future work will involve the research regarding the behavior of online paid posters who work on the CQA portals.

Bibliography

[1] Yahoo! Answers, Accessed March 2012.

[2] D. Chaffey. *Internet Marketing: Strategy, Implementation and Practice.* Pearson Education. Financial Times Prentice Hall, 2006.

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, Accessed May 2011.

[4] CNNIC, Accessed May 2011.

[5] N. Cristianini and J. Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods.* Cambridge University Press, 2006.

[6] E. Dragut, F. Fang, P. Sistla, C. Yu, and W. Meng. Stop word and related problems in web interface integration, 2009.

[7] Nick Fielding and Lan Cobain. Revealed: Us spy operation that manipulates social media, Accessed March 2011.

[8] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B.Y. Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th Annual Conference on Internet Measurement*, pages 35–47. ACM, 2010.

[9] Gooseeker, Accessed Apr. 2011.

[10] D. Higgins and J. Burstein. Sentence similarity measures for essay coherence. In *Proceedings of the 7th International Workshop on Computational Semantics (IWCS)*, 2007.

[11] Congrui Huang, Qiancheng Jiang, and Yan Zhang. Detecting comment spam through content analysis. In *Proceedings of the 2010 International Conference*

on Web-Age Information Management, WAIM'10, pages 222–233, Berlin, Heidelberg, 2010. Springer-Verlag.

[12] Minlie Huang, Yi Yang, and Xiaoyan Zhu. Quality-biased ranking of short texts in microblogging services. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 373–382, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.

[13] ICTCLAS2001, Accessed Jun. 2011.

[14] Panos Ipeirotis. Mechanical turk: Now with 40.92

[15] Junpeng Jia's Story, Accessed May 2011.

[16] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the International Conference on Web Search and Web Data Mining*, WSDM '08, pages 219–230, New York, NY, USA, 2008. ACM.

[17] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[18] Y. Li, D. McLean, Z.A. Bandar, J.D. O'Shea, and K. Crockett. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, pages 1138–1150, 2006.

[19] B.-Y. Liu, H.-F. Lin, , and J. Zhao. Chinese sentence similarity computing based on improved edit-distance and dependency grammar. *Computer Applications and Software*, 7, 2008.

[20] Qiong Luo. An undercover paid posters diary, Accessed Apr. 2011.

[21] MIT-Tech-Review, Accessed Nov. 2011.

[22] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 441–450, New York, NY, USA, 2012. ACM.

[23] Sai T. Moturu and Huan Liu. Quantifying the trustworthiness of social media content. *Distrib. Parallel Databases*, 29:239–260, June 2011.

[24] Arjun Mukherjee, Bing Liu, and Natalie Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 191–200, New York, NY, USA, 2012. ACM.

[25] Arjun Mukherjee, Bing Liu, Junhui Wang, Natalie Glance, and Nitin Jindal. Detecting group review spam. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 93–94, New York, NY, USA, 2011. ACM.

[26] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[27] L.D. Paulson. Building rich web applications with ajax. *Computer*, 38(10):14 – 17, oct. 2005.

[28] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[29] shuijunwang, Accessed May 2011.

[30] Sina.com, Accessed Apr. 2011.

[31] Sohu.com, Accessed Jun. 2011.

[32] M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. *AI 2006: Advances in Artificial Intelligence*, pages 1015–1021, 2006.

[33] E. Staff. Verisign: 1.5m facebook accounts for sale in web forum, April 2010.

[34] James Surowiecki. *The Wisdom of Crowds*. Anchor, August 2005.

[35] Gang Wang, Christo Wilson, Xiaohan Zhao, Yibo Zhu, Manish Mohanlal, Haitao Zheng, and Ben Y. Zhao. Serf and turf: crowdturfing for fun and profit. In *Proceedings of The 20th International World Wide Web Conference (WWW)*, 2012.

[36] Wikipedia, Accessed March 2012.

[37] Lotte M. Willemsen, Peter C. Neijens, Fred Bronner, and Jan A. de Ridder. Highly recommended! the content characteristics and perceived usefulness of online consumer reviews. *Journal of Computer-Mediated Communication*, 17(1):19–38, 2011.

[38] D. Yin, Z. Xue, L. Hong, B.D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the Web*, 2, 2009.

[39] Baidu Zhidao, Accessed March 2012.

Appendix A

Additional Information

Our work has attracted interests from various of research realms. It was first announced by *MIT Tech Review* in November, 2011[21]. The reviewers said that:

> "*That's an impressive piece of work and a good first step towards combating this problem, although they'll need to test it on a much wider range of datasets. Nevertheless, these guys have the basis of a software package that will weed out a significant fraction of paid posters, provided these people conform to the stereotype that Cheng and co have measured.*"

They also mentioned that the fight with paid spammers required a long-term exploration because the paid posters would adjust their behavior to bypass the inspection mechanism:

> "*And therein lies the rub. As soon as the first version of the software hits the market, paid posters will learn to modify their behaviour in a way that games the system. What Cheng and co have started is a cat and mouse game just like those that plague the antivirus and spam filtering industries.*"

This has been confirmed with one of our following up observations. Recently, we have an interesting finding on the fight between *360* and *Tencent*. The story between this two companies has not stopped. Recently, the fight between the two companies goes into trial. We read some of the news and find many potential paid posters. They often use very short sentences like "I support 360" or "I will definitely uninstall 360 on my computer". However, more than 90% of paid posters send the comments through "cell phones". So all of them have the same user ID. Since our detection method largely depends on the user IDs, it will be much difficult for us to locate the paid posters. For example, in a nearly 2000 users involved news report, we only find 57 users with their specific IDs. Most of the 57 users are actually normal ones. Others, who post comments through cell phones, look like malicious posters.

This phenomenon is very abnormal compared to the dataset collected one year ago. It seems like our first academic project taught the paid posters how to hide themselves. This encourages us to explore more techniques to identify those malicious users.

Below we list a few other media reports regarding our research.

1. Slashdot - “Internet Water Army On the March”
2. PC World - “Research: Paid Posters Poison the Internet”
3. The Atlantic Wire - “The Spam-Slinging Habits of Chinas Internet Water Army”
4. Wikipidia - “Internet Water Army”
5. ACM Tech News - “Undercover Researchers Expose Chinese Internet Water Army”
6. The Vancouver Sun - “Online product reviews skewed by paid posters”
7. Discovery News - “Beware: Product Reviews May Be Fake”