CSIS0801 Final Year Project

[Industry Project] Social Media Content Analysis System

Interim Report

Supervisor

Dr. Anthony T.C. Tam

Project Group

LEUNG Ka Wai	3035050250
LEUNG Kar Chun Victor	3035050676
NG Chun Ming	3035049794

CONTENT PAGE

Project Objective 3
Sustainable Data Extraction 3
Automatic Evaluation 3
Consistent Evaluation 3
Project Background 4
The Church of United Brethren in Christ Social Service Division 4
Trend of Severe Cyberbullying and “PeaceNet” 4
Current Situation 4
Problem Identification 5
Our Solution 6
Project Methodology 7
System Architecture 7
Design Methodology 7
Tools 8
Implementation Details 9
Overview 9
Crawler 11
Evaluator 12
Database 14
Interface 15
Progress 15
Testing 16
Effectiveness of First Page Monitor 16
Precision and recall of Evaluator 17
Future Tasks 18
Mood Analysis 18
Custom Set 18
Other Future Tasks 19
Project Schedule And Milestones 20
Milestones 20
Schedule 20
Gantt chart 21

PROJECT OBJECTIVE

SUSTAINABLE DATA EXTRACTION

The Social Media Content Analysis System will be able to gather the details of posts including author, user to be addressed, post content and last modified time. Our system will imitate a normal user's access pattern in order to extract the forum data by opening a legitimate browser session to crawl the data page by page and post by post with a random time interval. From the server's perspective, there will be no difference from our system compared to a normal user and thus no automated blocking would be issued hopefully.

AUTOMATIC EVALUATION

To relief the strict requirements on manpower, our system will provide a one-stop integrated service with regards to crawling, storage, display and content analysis of forum posts. The processes are automatic and run continuously day and night in backend server. Users need not to bother and pay attention to the insignificant details such as how to obtain and view the posts but can rather concentrate resources on identifying the key victims from the potential negative posts that the system has screened out. As such, user effort is minimized as well as time and labor demands on looking for targets is greatly reduced and allows better resource allocation on the client side.

CONSISTENT EVALUATION

Using multi-facets of analysis, our system would generate a consistent and representable evaluation on forum posts which can provide a constant and consistent reference for social workers to work on. Keyword list analysis based on frequency and severity and other types of analysis give rise to an overall credible analysis.

PROJECT BACKGROUND

THE CHURCH OF UNITED BRETHREN IN CHRIST SOCIAL SERVICE DIVISION

The Church of United Brethren in Christ Social Service Division (the Division) has been actively exploring the needs of the community especially children and youth and providing diversified services for them since 1984. The Division encourages whole person development physically, psychologically, intellectually and spiritually for teenagers and aims to react to the ever changing society to create a better environment which benefits the development of families and teenagers.

TREND OF SEVERE CYBERBULLYING AND "PEACENET"

With the increasing popularity and penetration of the internet in the recent years, cyberbullying through the web is becoming more and more severe. To address such issue, one of the major services provided by the organization, PeaceNet-Internet Outreaching Scheme, aims at seeking out potential victims or those being cyber-bullied in different kinds of social media platforms by utilizing online data mining tools such as KMatrix. Social workers are then able to identify and reach the victims to provide timely support to reduce the distress of the teenagers and prevent further disastrous tragedies.

CURRENT SITUATION

The division adopted some approaches to figure out the latest victims of cyberbullying but failed. The division used data mining tool (later on it used self-developed software) to extract large amount of data from the social media. After that, it discovered the potential threads by keywords which are believed to be frequently appeared in cyberbullying messages. Looking at the results one by one, it filtered the results manually. Once it found the targeted message threads, it contacts the victims.

Data mining tools
(Monthly Subscription)/
Self-created Software

Find the potential threads by
keyword &
Filter the results manually

Contact the victims

Data mining tools such as KMatrix requires a monthly subscription. In other words, it involves a regular amount of capital investment which is not economically sustainable in the long run. In order to alleviate the costing issue, the Church of United Brethren in Christ Social Service Division tried to implement a similar application themselves to handle the web crawling and content analysis of the social media platforms. However, their data extraction and analyzing approach was rather straightforward and thus unsustainable being blocked by the servers and platform administrators. As a result, the Social Service Division started to narrow down the scope from various social media platforms to a specific forum, the HKGolden forum. The staff of the Division switched their approach to manually scanning through the forum and subjectively analyzing the content of each post seeking for negative behavior online.

PROBLEM IDENTIFICATION

UNSUSTAINABLE AUTOMATED DATA EXTRACTION

The Division's previous application did not take the social media's defense mechanisms and server access analytics into consideration. Thus, abnormal patterns in terms of frequent access from the same domain and extremely rapid refresh rates can be easily traced and automatically blocked by the server or administrators. Therefore, web crawling was not sustainable due to the automated extraction method being identified easily and preventive measures can be allocated from the social media's perspective with ease.

HUGE DEMAND OF MANPOWER

Due to the immense pool of data including the enormous number of posts to be extracted and analyzed through pure manpower, the Division needs many dedicated staff to scan and analyze in order to do the job efficiently. In reality, this is not possible as there is not enough human resources. Thus, the process would be time consuming and rather inefficient which definitely diminishes the possibility of finding victims.

INCONSISTENCY OF CONTENT EVALUATION

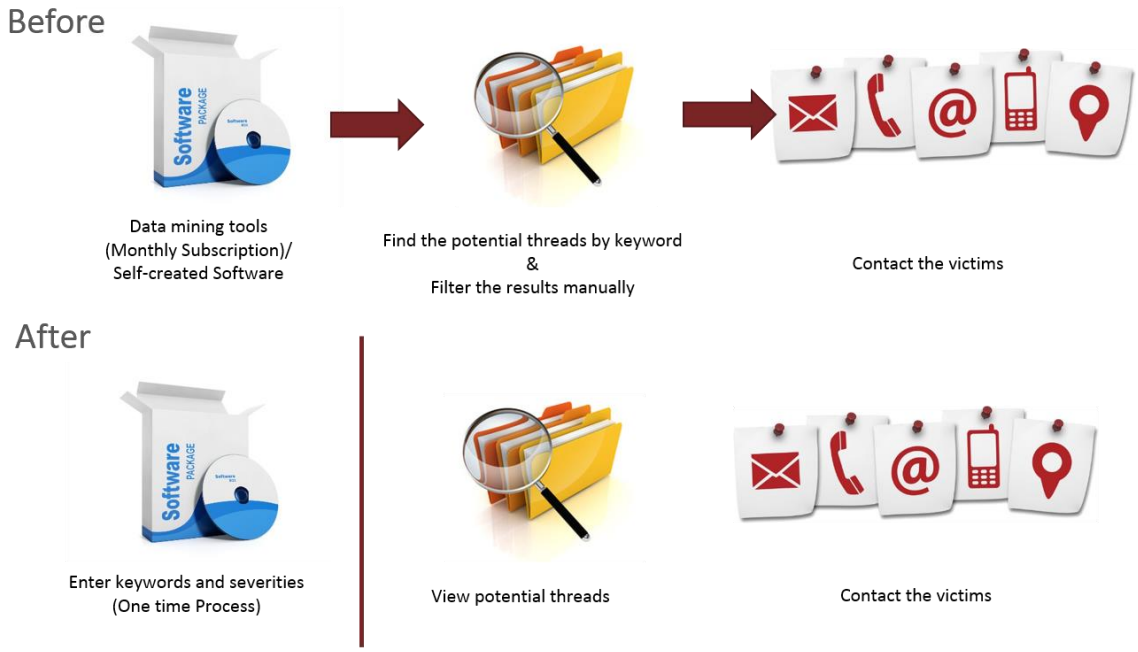
As the content analysis is done by different staff over time and each have subjective opinions on the degree and severity of cyberbullying, the overall result would be rather inconsistent. Some may see a certain case as offensive while others might think it is acceptable. This confuses the social workers whether or not to contact the potential victim as different views are raised.

OUR SOLUTION

Social Media Content Analysis System is dedicated to improve both efficiency and effectiveness of the data extraction of forums such as HKGolden. It should be capable of extracting data sustainably as well as allowing efficient identification of potential victims or those being cyber-bullied. It will crawl posts and threads of HKGolden on a real time basis and rank them based on severity and other algorithms. Quoting of the negative expressions and wordings will be displayed for preview for the worker's convenience, for each thread, their corresponding posts will also be arranged in a ranked order so that social workers can track the most severe cases easily.

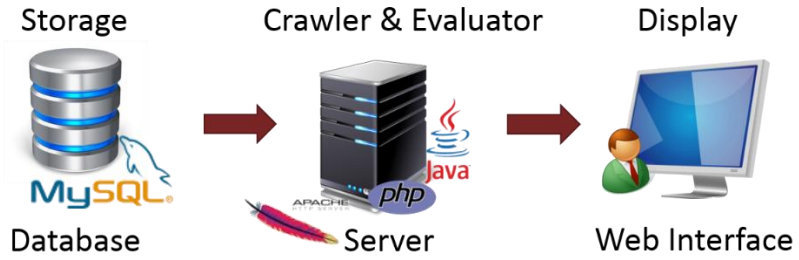
NEW USER FLOW

Users will be only required to enter keywords and severities at first use of the new system. After that, the users simply view the potential threads ranked by the system and contact the victims if necessary.



PROJECT METHODOLOGY

SYSTEM ARCHITECTURE



The system consists of 3 components – database, server, and web interface.

Database storage is a relational database operating under MySQL Version 5 or above to store the necessary information for the system to run such as forum threads, posts, and keywords.

Server contains forum crawler and evaluator. It runs an operating system which supports web browsers and have a stable version of Chrome/Firefox/Opera/Safari installed to facilitate web crawling to get raw data and also do analysis based on predefined algorithms.

Any devices with browser support and Internet access can access the web interface provided by the server.

DESIGN METHODOLOGY

For the Social Media Content Analysis System, our team believes adhering to a Rapid Application Development (RAD) based method can achieve efficiency while also clarifying and meeting user requirements comprehensively by delivering partial parts of the system to the user for trial. Users will be able to understand their underlying requirements better and suggest enhancements that shall make the system more all rounded.

We shall continue to split the implementation into phases as it provides an excellent and well-defined schedule so that we can ensure our project progress is on track. Time is a crucial and limiting factor for our project as we have to strictly adhere to the 8 months limitation for FYP grading purposes. The remaining phases will include more advanced algorithms that will enhance the content analysis accuracy. Meeting with client shall be arranged and users can try out and give feedback for each phase to ensure their needs are satisfied and give suggestions and improvements whenever applicable. Basic

functions are also ensured to work thoroughly before we commit to the next phase as that basic functionalities can be guaranteed for the system.

TOOLS

JAVA

We use Java as the major implementation language for our system in server as Java is designed to be interoperable among different operating systems, Windows and Linux alike, so that our system will function on cross platforms.

The system executable will also be more portable compared to most other compiled alternatives in the sense that no additional modules will be required to be installed first.

LIBRARIES

To facilitate our project, libraries that are under non-commercial licenses may be used. At the current stage, we have identified 2 libraries that will be utilized.

1. Selenium
    - test automation tool which allows web crawling automation using multiple web browsers
2. MySQL connector/J
    - JDBC connector for MySQL

ENVIRONMENT

WAMP (Windows, Apache, MySQL, Php)

DEVELOPMENT IDE

Eclipse

VERSION CONTROL

Git is being used for version control. All codes are committed to BitBucket in https://kwleung@bitbucket.org/kwleung/finalyearproject.git

IMPLEMENTATION DETAILS

OVERVIEW

The overview of all classes involved in the system can be found in Figure 1.

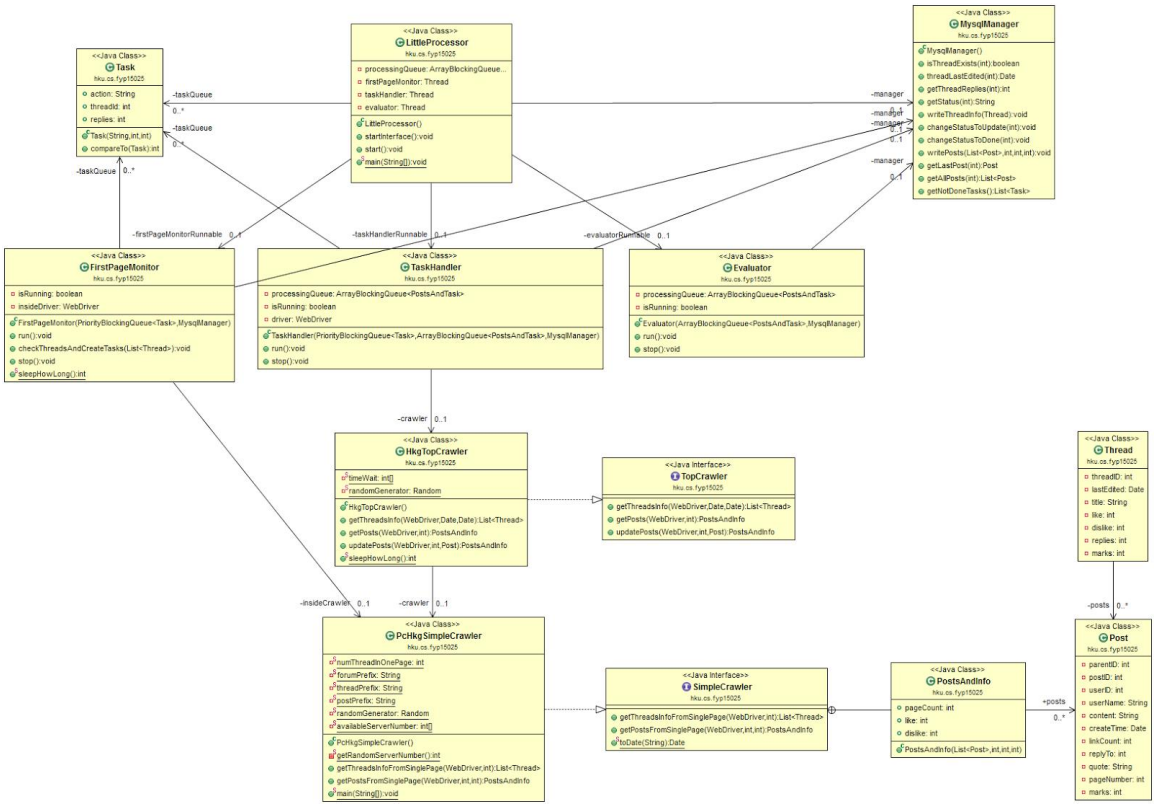


Figure 1: Class Diagram

From a high level perspective, there are three major Java classes running on the server: First Page Monitor, Task Handler, and Evaluator.

The communications between classes are implemented using producer consumer approach (Figure 2) for maximized web crawling efficiency and easier maintenance. First Page Monitor pushes tasks into a PriorityBlockingQueue. Task Handler takes the tasks from the queue, executes the tasks to get lists of posts, and put the lists into another blocking queue. Finally, Evaluator takes the lists from the queue, makes analysis on the posts, and put the information into database.

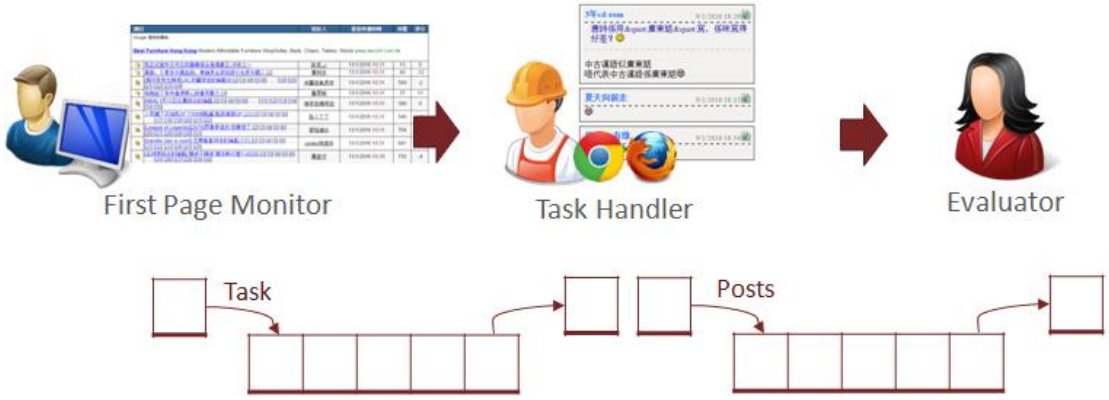


Figure 2: The system uses blocking queues to handle the interactions between classes

The detailed description of the three major classes are as follows:

FIRST PAGE MONITOR

First Page Monitor looks for new message threads and threads that require further updating by browsing the first page of the forum thread list page. It creates tasks and orders Task Handler to get or update the message threads.

TASK HANDLER

Task Handler receives tasks from First Page Monitor and executes the tasks. Tasks can be divided into two types: “New” and “Update”. For “New” task, task handler browses and keeps all the posts of a message thread from the beginning page to the last page. For “Update” task, Task Handler browses and keeps new posts by comparing each post with the last post that is stored in the database. After the successful execution of tasks, it passes the list of posts to Evaluator for evaluation.

EVALUATOR

Evaluator, as the name suggests, evaluates the posts and message threads using an algorithm involving keyword searches, hyperlink counts, likes, dislikes.

CRAWLER

The crawler is implemented using layered structure (Figure 3).

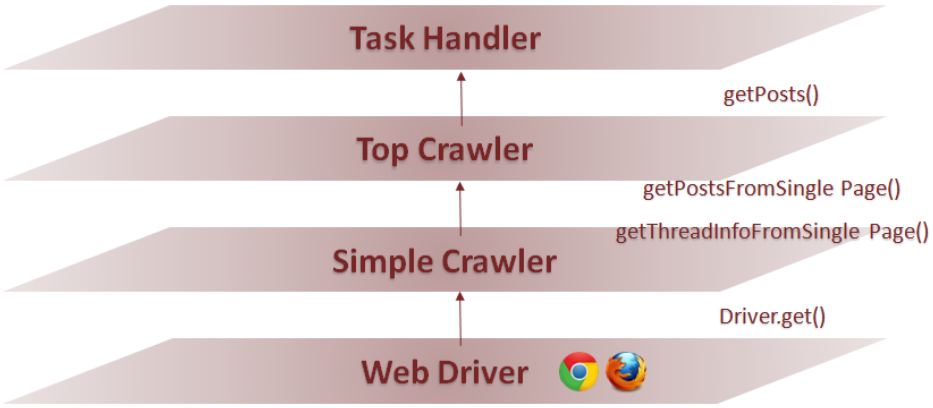


Figure 3: Layered structure of web crawling

TASK HANDLER

Being one of the three major Java classes in the system, Task Handler executes tasks from First Page Monitor utilizing the service provided by the top crawler. For the task ordering inside the PriorityBlockingQueue, “New” tasks have higher order than “Update” tasks. Within each category of task, the ordering is by shortest job first (smallest number of replies) to achieve higher efficiency.

TOP CRAWLER

Top crawler is responsible for the inter-page crawling. It does not know how to get the info within one page but it utilizes the service provided by simple crawler to get all the info. It knows the logics to crawl multiple pages within one message thread, with the ability to simulate user activities using Thread.sleep(time) method and following human web access pattern.

SIMPLE CRAWLER

Simple crawler is responsible for the intra-page crawling. It utilizes the service provided by web driver to get the specified web elements and retrieves relevant information. It is the main layer to be changed by programmers when the layout of the web pages are

overhauled. PC and mobile site crawlers can be implemented in this layer. For load distribution, simple crawler shall access different HKGolden forum servers randomly.

WEB DRIVER

Web driver is a part of test automation tools. It controls browser activities (e.g. Chrome, Firefox and IE). Therefore, the HTTP request headers are exactly the same as that from normal browsers. Each web driver uses a random port to support simultaneous operation of multiple browser sessions.

EVALUATOR

Evaluator is responsible for evaluating posts and threads immediately after crawling. Posts and threads are given corresponding score that represent their relevance to potential targets. The higher the score, the more relevant it is. After evaluation, data of the posts as well as corresponding results would be inserted to the database.

POST EVALUATION

Within one message thread, 2 main components of its posts, which are wordings and receivers, would be considered respectively to generate the scores.

1. Wordings

Wordings of each post would be evaluated based on 4 sets of keywords. Score of each keyword set is given according to the severity as well as the popularity, i.e. the frequency of existence of the keywords in normal posts.

Rather than purely manually analyzing the forum to discover the keyword score based on severity, we determined the popularity of the keywords using a relatively more scientific method. We have trained a Stanford NLP Segmenter (presuming that can be reused in the latter stages of the project when we touch upon machine learning) which is able to segment post content into meaningful Cantonese vocabularies. We used the gathered data as input for the segmenter to find common vocabularies used in normal conversations inside the forum and then we did the same for the threads that we have identified having cyberbullying elements. From the comparison of these two sets of vocabulary term frequencies, we were able to determine which vocabularies actually represent negative emotions and give scores accordingly.

The score is directly proportional to the severity while the popularity is inversely proportional to the score. Table 1 shows scores of these keyword sets and their respective severity and popularity with examples.

Keyword Score	Severity	Popularity	Examples
10	High	Low	起底、推上報
5	High	High	岑
2	Low	Low	廢青
1	Low	High	負皮、樣衰

Table 1: Keyword examples and their scores

Scores would be given to each post according to the occurrences of keywords. However, multiple occurrences of keywords from same set would only be considered once in the current implementation. Therefore, the scores of a post would be ranged from 0 to 18 (10 + 5 + 2 + 1).

2. Receivers

After evaluation on wordings, the results would be further evaluated based on receivers of posts. If a user receive posts with keywords, i.e. posts with scores > 0, from 3 distinct users in one single thread, scores of those posts would be doubled.

THREAD EVALUATION

A thread would be evaluated based on 4 categories: posts, titles, hyperlinks as well as likes and dislikes. Posts and title of a thread would act as major factors of thread score while hyperlinks and likes and dislikes only contribute a small portion of the thread score.

$$\begin{aligned} Thread\ Mark = &\left(If\ no.of\ posts\ \leq 35\ then \sum Post\ Marks\ else\ AVG(Post\ Marks) \times 100\right) \\ &+ (If\ title\ contain\ keywords\ then - 80\ else\ 0) \\ &+ \left(If\ thread\ contains\ post\ with > 3\ hyperlinks\ then \left(\frac{no.of posts\ with > 3\ hyperlinks}{10} + 10\right) else\ 0\right) \\ &+ (If\ |Like - Dislike| > 30\ then\ 10\ else\ 0) \end{aligned}$$

1. Posts

Posts of each thread would be first considered. The thread score is evaluated by summarizing the scores of all corresponding posts. However, thread with more than 35

posts would have its score calculated by average post scores times 100 so that threads with lots of posts will not dominate the ranking.

2. Title

Title of a thread is recognized as a useful indicator to eliminate false positive threads, such as threads discussing politics, TV programs and games. Threads with titles containing specific keywords would have their scores being diminished by 80.

3. Hyperlinks

After that, number of hyperlinks in each post of a thread are assessed. With existence of posts with more than 3 hyperlinks in a thread, the evaluator appends the corresponding thread scores by 10. Moreover, a score of value equaling the number of posts with more than 3 hyperlinks in that thread divided by 10 would also be added to the overall thread score.

4. Like & Dislike

Eventually, number of likes and dislikes of a thread would be processed. Any abnormality of the number of likes and dislikes, i.e. difference between likes and dislikes is more than 30, would trigger the evaluator to append the thread score by 10.

DATABASE

After crawling and evaluating the threads and posts in real time, the gathered data are stored into a database for efficient writing and reading operations handling and to prevent concurrency issues as the crawler and evaluator runs in real time. The schema and relations of the database tables can be seen in the below ER Diagram (Figure 4). Besides from obvious attributes such as content and userId to represent the original posts, other attributes such as pageNumber, linkCount, replyTo, _like, etc. are used to improve crawling efficiency and act as inputs for the evaluator and display interface.

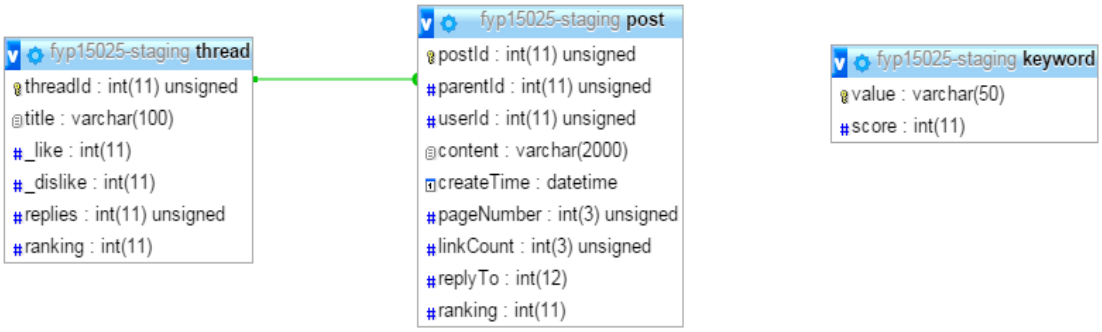


Figure 4: Entity Relationship Diagram

INTERFACE

Web interface has been chosen as GUI for the system to support multiple operating systems as it is cross-platform in nature. Also, it relaxes the requirements on resources of the client computers to act as thin clients to connect to one central server to handle all the logics and support multiple connections simultaneously. Access to the server is written in PHP to output HTML so that CSS can be used as styling. The current version supports 3 main features including dashboard to display summary of the server, view results to display the problematic threads with ranking and keyword revision to modify keyword values and severity to alter the evaluation according to client preferences.

PROGRESS

Phase 1 and 2 were to do basic content analysis and preliminary implementation. They are completed quite successfully and the project is on schedule.

TESTING

EFFECTIVENESS OF FIRST PAGE MONITOR

MISS RATE

Miss rate is the rate of missing thread updates during the idle time of the First Page Monitor. First Page Monitor keeps on retrieving the first page of thread lists to discover the new messages at random intervals. On the other hand, there are users that the user generated contents pop up rapidly. Therefore, it is very likely to miss some of the updates.

Currently, the system uses the number of unchanged threads in the first page to calculate the waiting time for the next Thread.sleep(time). If the system find that contents updates rapidly, first page monitor will poll the page more frequently. Figure 5 shows the number of unchanged threads against time for one day. Among 3089 polling, the first page missed 530 times. The miss rate is 17.16%.

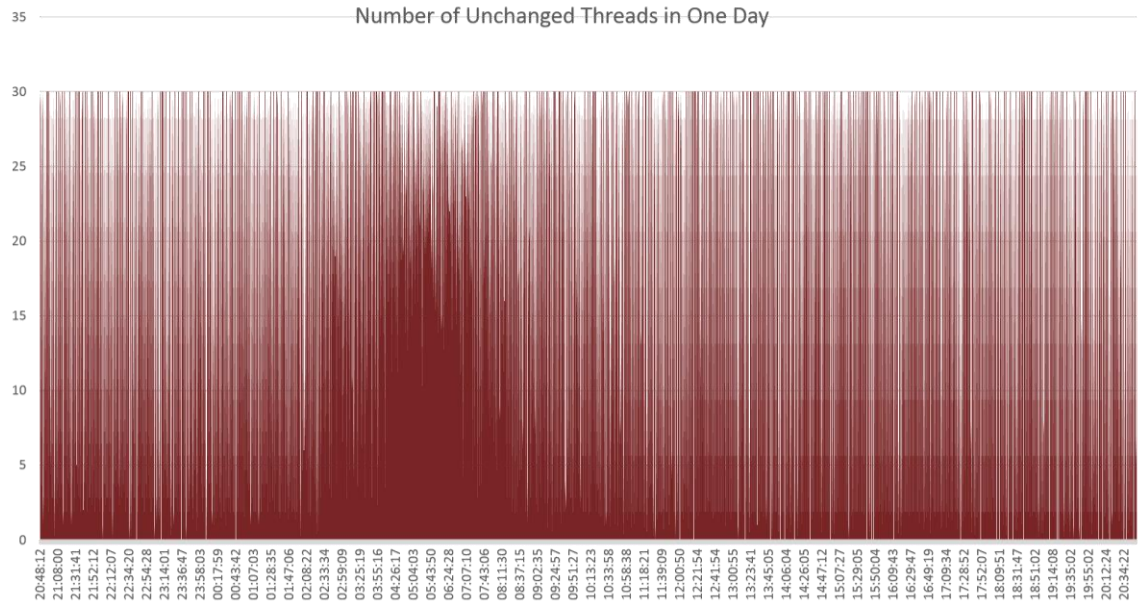


Figure 5: Number of unchanged threads in one day

It can tell from the graph density of the lines that users are inactive from 3 a.m. to 8 a.m. Adjustments can be made to relieve the loading from system server to forum server. One effective way to reduce the miss rate is to increase the poll rate. However, this may jeopardize the sustainable crawling as one of the project objectives.

It is worth mentioning that the project team had tried using online user count from forum page to set the First Page Monitor's sleeping time. However, it is believed that the online user count is a trick after team's observation for several days.

PRECISION AND RECALL OF EVALUATOR

Test cases are selected manually. Positive cases not only consist of the well-known old cyberbullying cases but also brand new cases identified by the project team. For negative cases, they are selected randomly with manual screening to filter out the positives.

Test cases are feed into the system with no entries in database. Precision and recall are calculated after the message thread evaluations. Table 2 shows the evaluation result by taking 80 as the threshold thread score. Precision is 82% and recall is 85%. While the precision and recall of client approach are 38% and 98% respectively. Although there is a slight decrease in recall for the new approach, precision is increased significantly, which, in no doubt, reduces time needed for clients to screen false positive cases from positive ones. Moreover, recall can be enhanced by lowering the threshold thread score with only minor influence on precision. Project team will also keep on improving the precision and recall by refining the preliminary evaluation algorithm and using mood analysis to adjust the rankings.

	True	False
Positives	58	13
Negatives	102	10

Table 2: Evaluation result (new approach), Precision: 81.69%, Recall: 85.29%

	True	False
Positives	67	111
Negatives	4	1

Table 3: Evaluation result (client approach), Precision: 37.64%, Recall: 98.53%

FUTURE TASKS

MOOD ANALYSIS

Our system would extend to support multiple types of mood classification rather than solely cyberbullying content such as including classification of anger, joy and sadness. The underlying working principle would be similar to that of cyberbullying identification - after evaluation results certain marks which if exceeding a certain threshold would be considered as fulfilling the type. The results of different classification of moods can further be leveraged as input to increase the accuracy of finding cyberbullying victims. For instance, a thread being classified as “anger” may further increase the rankings of that content in cyberbullying.

We plan to implement this feature in 2 alternative approaches and measure which is more feasible and suitable in our project scope based on recall and precision. The first approach is more aggressive and we are not fully inclined that it would work flawlessly so we deem to use a second approach as backup if the former fails.

1. Machine Learning

Supervised learning would be carried out to train a classifier to classify threads into different type or classes in terms of moods. We will use the data we have crawled so far (which is still crawling in real time) as training set to feed the AI and hopefully have enough data to support a functional classifier. At this stage we have not determined which AI algorithm would be preferable in terms of Cantonese wording classification and therefore plan to test on different available AI algorithms to seek the best performing one.

2. Keyword Definition for Different Moods

Using the exact same implementation of the keyword analysis of cyberbullying, different keyword sets would be discovered and defined for each mood. Evaluation of a certain mood would be based on its respective set of keywords. Mood binding would be possible such that a keyword can represent multiple moods with respective severities. Thus, in the grand scheme of things, a certain thread may be classified as having multiple moods.

CUSTOM SET

We also expect the system to be able to cater custom set of keywords. That is, if the user wants to identify a certain type or mood of thread content, the user can define its own custom keyword set which accommodates the characteristic of that certain type or mood. Our system will evaluate the content based on his or her inputs and give ranking results

correspondingly. This allows support of the system to evolve to cater future cases even when we pass the system to the client after project completion.

OTHER FUTURE TASKS

KEYWORD SET CUSTOMIZATION

There are frequent change of trendy words used by netizens and forum users. The ability for the system to adapt to the changes of keywords is the key to achieve long run usage. On the other hand, client might want to adjust the scores of keywords with time. Therefore, in the future, the system shall provide an interface for client users to add/drop a keyword and adjust scores.

The preliminary interface is already in use but the logics after the change in keyword set (cron job, real time evaluator) are not yet implemented. This becomes one of the future tasks and shall be done by the project team in the second semester.

CRAWLER IMPROVEMENT

It is found that the page number of a post does not change across different forum viewports (PC site, mobile site). Also, the project team had identified the way to locate the real post ID. Those two fields will be utilized to have better thread update efficiency. In addition, there several bugs need to be addressed. The system shall handle timeout exceptions properly using exponential backoff, instead of stop accessing that timeout server anymore.

USER ACCEPTANCE TEST

Early user acceptance test shall be conducted to collect client's feedbacks. Changes can be made to the system accordingly to have the best system implemented at final system delivery.

MOBILE CRAWLER IMPLEMENTATION

If crawling pc site was not efficient enough that it became the bottleneck, mobile site crawler shall also be implemented.

PROJECT SCHEDULE AND MILESTONES

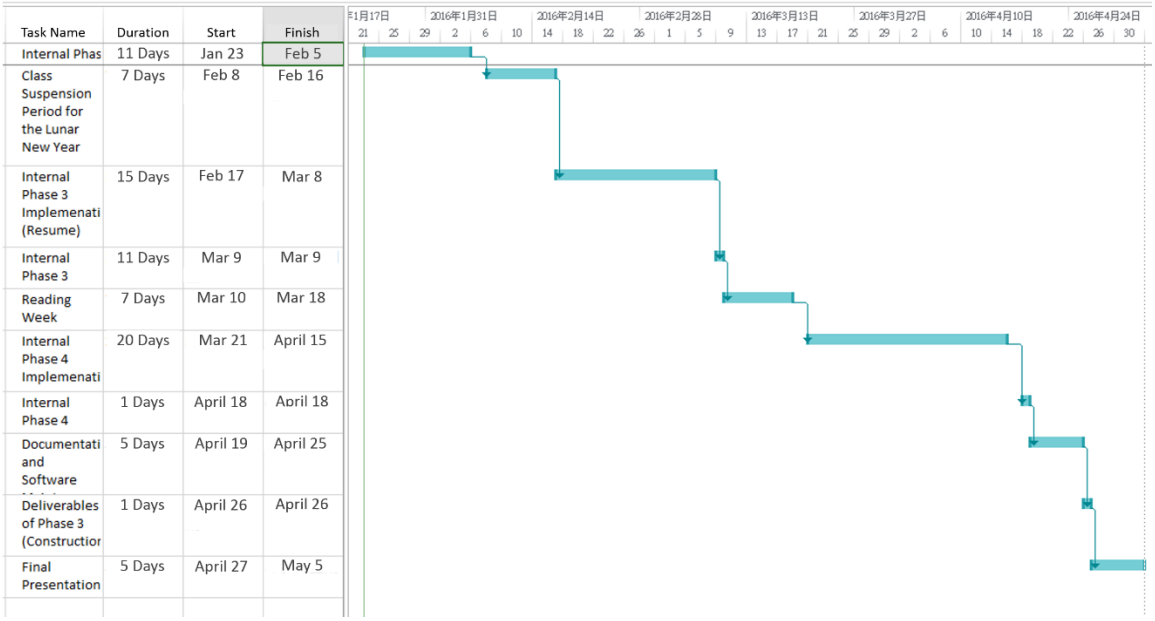
MILESTONES

Date	Milestone
5 March 2016	Internal phase 3 finish
9 April 2016	Internal phase 4 finish
17 April 2016	Deliverables of Phase 3 (Construction)
18 April 2016	Final presentation

SCHEDULE

Date	Tasks
25 Jan - 5 Feb	Internal phase 3 implementation • analysis enhancement • mood analysis
6 Feb - 14 Feb	Class Suspension Period for the Lunar New Year
15 Feb - 4 Mar	Internal phase 3 implementation (continued)
5 Mar	Internal phase 3 finish
6 Mar - 13 Mar	Reading week
14 Mar - 8 Apr	Internal phase 4 implementation • custom set
9 Apr	Internal phase 4 finish
11 Apr - 15 Apr	Documentation and software maintenance
17 April	Deliverables of Phase 3 (Construction) • Finalized tested implementation • Final report
18 - 22 April	Final presentation

GANTT CHART



Task Name	Duration	Start	Finish
Internal Phas	11 Days	Jan 23	Feb 5
Class Suspension Period for the Lunar New Year	7 Days	Feb 8	Feb 16
Internal Phase 3 Implemenati (Resume)	15 Days	Feb 17	Mar 8
Internal Phase 3	11 Days	Mar 9	Mar 9
Reading Week	7 Days	Mar 10	Mar 18
Internal Phase 4 Implemenati	20 Days	Mar 21	April 15
Internal Phase 4	1 Days	April 18	April 18
Documentati and Software	5 Days	April 19	April 25
Deliverables of Phase 3 (Constructior	1 Days	April 26	April 26
Final Presentation	5 Days	April 27	May 5