CSIS0801 Final Year Project

[Industry Project] Social Media Content Analysis System

Project Plan

Supervisor
Dr. Anthony T.C. Tam

Project Group
LEUNG Ka Wai **3035050250**
LEUNG Kar Chun Victor **3035050676**
NG Chun Ming **3035049794**

CONTENT PAGE

Project Background 3
The Church of United Brethren in Christ Social Service Division 3
Trend of Severe Cyberbullying and “PeaceNet” 3
Current Situation 3
Problems Identification 4
Our Solution 4
Project Objective 5
Sustainable Data Extraction 5
Accurate Content Analysis 5
Ranking 5
Project Methodology 6
Design Methodology 6
Tools 7
System Architecture 8
Project Schedule and Milestones 9
Milestones 9
Schedule 10
Gantt chart 12

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

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.

PROBLEMS IDENTIFICATION

1. 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.

2. 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.

3. 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 Forum. It should be capable of extracting data sustainably as well as allowing efficient identification of potential victims or those being cyber-bullied. It will rank threads of HKGolden Forum up to a certain time threshold and rank them based on 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.

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.

ACCURATE CONTENT ANALYSIS

To achieve efficiency and effectiveness of content analysis, different algorithms and language processing would be used by the system. However at this preliminary stage, only keyword list would be implemented. The system will evaluate the forum posts based on a keyword list provided by the Church of United Brethren in Christ Social Service Division. Based on calculations such as frequency and severity, each thread and its corresponding posts will be given a mark to show its negativeness. The accuracy will be further enhanced in the later phases using more advanced algorithms or processing methods. Based on this hopefully multi-facets of analysis, a consistent and representable content evaluation would be present for the social workers to use.

RANKING

To facilitate social workers' goal to contact victims, our system would rank the results from most severe to least severe so that the workers can better identify those who are in need the most and better allocate the Division's limited resources to help the needy netizens. This also provides consistent reference for workers to identify potential victims.

PROJECT METHODOLOGY

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.

	Structured Methodologies			RAD Methodologies		Agile Methodologies
Ability to Develop Systems	Waterfall	Parallel	Phased	Prototyping	Throwaway Prototyping	XP
with Unclear User Requirements	Poor	Poor	Good	Excellent	Excellent	Excellent
with Unfamiliar Technology	Poor	Poor	Good	Poor	Excellent	Poor
that are Complex	Good	Good	Good	Poor	Excellent	Poor
that are Reliable	Good	Good	Good	Poor	Excellent	Good
with a Short Time Schedule	Poor	Good	Excellent	Excellent	Good	Excellent
with Schedule Visibility	Poor	Poor	Excellent	Excellent	Good	Good

Figure 1 – Criteria for Selecting a Methodology

We shall use phased development under RAD 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. We plan to have 4 phases with the earlier phases only including the minimal and core functions while the latter phases will include more advanced algorithms that will enhance the content analysis accuracy. 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. Phased development also gives a more organized launching process so we can better quantify what is needed for our project milestones such as the presentations.

TOOLS

Although the full and detailed list of tools that will be involved and used is not necessarily defined at this preliminary stage. We are sure certain tools will be involved as follows.

1. Java

We plan to use Java as the programming interface for our system 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.

2. Libraries

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

(i) Selenium – allows automation of web browsers for testing and caching HTML data
(ii) SWT/Swing – allows designing and creation of the Graphical User Interface

3. MySQL

Relational databases will be existing in our framework. For the database format and language, MySQL shall be used due to its open source nature and being complementary with other web technologies such as PHP so the system may be scalable on a server structure if the need arises.

SYSTEM ARCHITECTURE

The system will consist of 3 components – Display, Storage and Crawler/Analyzer.

Display device shall be the major client PC running Windows 7 or above with Java Runtime Environment Version 6 or above installed to enable our Java written GUI which allows display and interactions with the underlying system.

Storage will be a relational database operating under MySQL Version 5 or above to store the necessary information for the system to run such as parts of the forum posts, keywords and the rankings.

Crawler/Analyzer shall be running on 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.

PROJECT SCHEDULE AND MILESTONES

MILESTONES

Date	Milestone
4 October 2015	Deliverables of Phase 1 (Inception)
31 October 2015	Internal Phase 1 Finish
30 November 2015	Internal Phase 2 Finish
11 January 2016	First Presentation
24 January 2016	Deliverables of Phase 2 (Elaboration)
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
1 Sept 2015	• First meeting with industry partner • Gather User Requirements
2 Sept - 14 Sept	Feasibility Analysis
15 Sept	• Second meeting with industry partner • Clarify User Requirements
16 Sept - 2 Oct	• Implementation of first prototype for sustainable forum crawling • Finalize User Requirements
4 Oct	**Deliverables of Phase 1 (Inception)** • Detailed Project Plan • Project Web Page
5 Oct - 9 Oct	Testing of First Prototype
10 Oct - 17 Oct	Reading Week
19 Oct - 30 Oct	Internal Phase 1 Implementation • Forum Crawler (for HKGolden Forum) • Basic Analysis
31 Oct	**Internal Phase 1 Finish**
2 Nov - 27 Nov	Internal Phase 2 Implementation • UI • Analysis Investigation
30 Nov	**Internal Phase 2 Finish**
1 Dec - 7 Dec	Revision Period

8 Dec - 23 Dec	Assessment Period
24 Dec 2015 - 10 Jan 2016	Semester Break
11 Jan - 15 Jan	**First Presentation**
18 Jan - 22 Jan	Documentation and Software Maintenance
24 Jan	**Deliverables of Phase 2 (Elaboration)** • Preliminary Implementation • Detailed Interim Report
25 Jan - 5 Feb	Internal Phase 3 Implementation • Analysis Enhancement
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 • Analysis Enhancement
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
1	First meeting with industry partner	0 days	Tue 9/1/15	Tue 9/1/15
2	Feasibility analysis	9 days	Wed 9/2/15	Mon 9/14/15
3	Second meeting with industry partner	0 days	Tue 9/15/15	Tue 9/15/15
4	Implementation of first prototype	13 days	Wed 9/16/15	Fri 10/2/15
5	Finalize user requirements	13 days	Wed 9/16/15	Fri 10/2/15
6	Deliverables of Phase 1 (Inception)	0 days	Sun 10/4/15	Sun 10/4/15
7	Testing of first prototype	5 days	Mon 10/5/15	Fri 10/9/15
8	Internal phase 1 implementation	10 days	Mon 10/19/15	Fri 10/30/15
9	Internal phase 1 finish	0 days	Sat 10/31/15	Sat 10/31/15
10	Internal phase 2 implementation	20 days	Mon 11/2/15	Fri 11/27/15
11	Internal phase 2 finish	0 days	Mon 11/30/15	Mon 11/30/15
12	First presentation	5 days	Mon 1/11/16	Fri 1/15/16
13	Documentation and software maintenance	5 days	Mon 1/18/16	Fri 1/22/16
14	Deliverables of Phase 2 (Elaboration)	0 days	Sun 1/24/16	Sun 1/24/16
15	Internal phase 3 implementation	25 days	Mon 1/25/16	Fri 3/4/16
16	Internal phase 3 finish	0 days	Sat 3/5/16	Sat 3/5/16
17	Internal phase 4 implementation	20 days	Mon 3/14/16	Fri 4/8/16
18	Internal phase 4 finish	0 days	Sat 4/9/16	Sat 4/9/16
19	Documentation and software maintenance	5 days	Mon 4/11/16	Fri 4/15/16
20	Deliverables of Phase 3 (Construction)	0 days	Sun 4/17/16	Sun 4/17/16