An Analysis of Public Perceptions for K12 Implementation in the Philippines using Web Content Mining Techniques and Wrapper Induction Algorithm

Angie M. Ceniza
University of San Carlos
School of Arts and Sciences
Department of Computer and
Information Sciences
angieceniza9@gmail.com

Christian V. Maderazo
University of San Carlos
School of Arts and Sciences
Department of Computer and
Information Sciences
cvmaderazo_madech@yahoo.com.ph

ABSTRACT

In recent years, Basic Education Curriculum is a 10 – year scheme, Grade 1 to 10. It has been modified to 12 years and we call it as the new curriculum or K-12 curriculum. Researchers would like to know the sentiments of the public about this implementation of the new educational system. Today, web is the best medium in expressing one's sentiments. Internet citizen makes use of blogs, social media or any other media in which they can express and exchange ideas. In this paper, we explore web content mining which is a web mining techniques in harvesting web pages. Wrapper induction algorithm and Opinion Lexicon were used to evaluate the web content as positive, negative and neutral. The research achieves 83.94% precision and 72.32% recall in measuring the performance of the system.

CCS Concepts

• Information systems~Web searching and information discovery • Information systems~Web search engines • Information systems~Web crawling • Information systems~Site wrapping

Keywords

Web Mining, Web Content Mining, Wrapper Induction Algorithm, Opinion Lexicon

1. INTRODUCTION

Perception analysis is often used to find out how people understand or feel about their situations or environments. The Philippines implemented a new curriculum which started last school year 2012-2013. This is the shift from the Basic Education Curriculum to the new K to 12 Curriculum. The said innovation in Philippine education has been made legal by the Republic Act 1033 or the Enhanced Basic Education 2013. There are many innovations introduced to the curriculum such as the extension of years spent in school. From the old 10-year scheme, Grade 1 to 10, it has been modified to 12 years. Sentiment analysis is a computational treatment of people's opinions, attitudes and emotions towards an entity [1]. It is one way of knowing the public sentiments regarding the implementation of the new K to 12 Curriculum. In gathering sentiments, web mining is the most common technique in collecting public opinions. Internet citizen post their opinions, ideas and suggestions on the web [2]. Web is most popular large data repository of containing broad of data and knowledge base, in which the information is hidden [8]. Researchers make use of web management methods and effective extraction of all related information to K to 12 curriculums from the web. To understand web mining we should know the data mining techniques in available. Figure 1 shows the taxonomy of web mining.

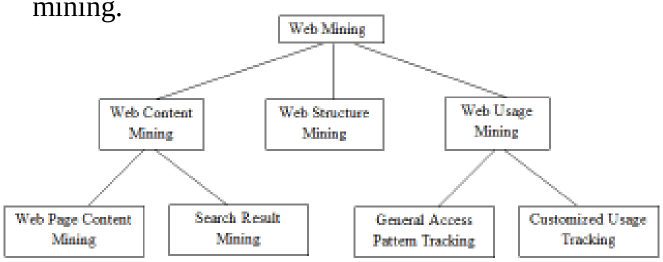

Figure 1.Taxonomy of Web Mining

Web content mining is the process of extracting knowledge from documents and content descriptions. Web structure mining is the process of obtaining knowledge from the organization of the web and the links between the web pages. Web Structure mining is the process of analyzing information about the visited web pages in order to discover unknown pattern. Figure 2 give the idea about web mining research which is divided into three categories.

These include (1) web content mining (2) web structure mining and (3) web usage mining.

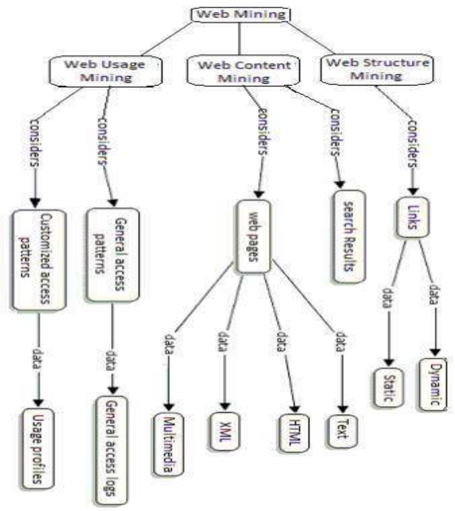

Figure 2.0 Categories of Web Mining

Researchers make use of web content mining. Through this, all related information regarding K-12 implementation was harvested and extracted by wrapper induction. Those data are processed and evaluated to analyze the public perceptions for K12 implementation in the Philippines.

2. REVIEW OF RELATED LITERATURE

These are several lines of related work which are reviewed in this section.

2.1 Web Mining

The emerging field of web mining aims at finding and extracting relevant information that is hidden in Web-related data, in particular in (hyper-) text documents published on the Web. Web mining techniques could be used to build the information retrieval and extractor part, since building it manually is time-consuming and not scalable [5]. In the study of Wang [6], web mining is consist of pre-processing, pattern discovery and pattern analysis. Kosala and Blockeel [7] conducted survey that point some confusion regarding the usage of the term Web Mining and suggest three Web Mining categories. In the research of Li and Yu [9], web pages typically contains large amount of information that a web mining techniques enable to identify which is relevant and irrelevant information.

2.1.1 Web Content Mining

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content [10]. It describes the discovery of useful information from the web documents. In web content mining, the content maybe text, image, audio, video, metadata and hyperlinks. It is mainly based on research in information retrieval and text mining, such as information extraction, text classification and clustering, and information visualization. Web content mining could be differentiated from two points of view: the agent-based approach or the database approach. The first approach aims on improving the information finding and filtering and could be placed into the following three categories
[11]:

1. Intelligent Search Agents. These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.
2. Information Filtering/ Categorization. These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.
3. Personalized Web Agents. These agents learn user preferences and discover Web information based on these preferences, and preferences of other users with similar interest.

The second approach aims on modeling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it. Researchers find and filter the information of K to 12 curriculums using information filtering and categorization.

2.2 Wrapper Induction

Wrapper induction (WI) [12] aims to generate extraction rules, called wrappers, by mining highly structured collections of Web pages that are labeled with domains specific information. Many systems have been built that automatically gather and manipulate such information on a user's behalf. However, these resources are usually in different format depending on the user's preference, so extracting their content is difficult and time consuming. Most systems used customized wrapper [13] procedures to perform extraction task. Wrappers provide an effective mechanism to extract information for a given website, and can often be learned using a very small number of labeled examples [14].

2.3 Opinion Lexicon

Opinion words [15] are usually used in a number of sentiment analysis tasks. They are used as features in sentiment classification. The most commonly used lexicon sources are WordNet and Opinion Lexicon which exists in languages other than English. Building resources, used in Sentiment Analysis tasks, is still needed for many natural languages. In the research of Souza and et. al [16] opinion lexicons are linguistic resources annotated with semantic orientation of terms (positive and negative) and are important for opinion mining tasks. Hu and Lui [17] maintain and freely distribute a sentiment lexicon consisting of list of strings. Flekova and et.al [18] uses open lexicon in identifying frequent bigrams where a polar word switches polarity. Cruz and et.al [19] uses the opinion lexicon in computing the semantic orientation (positive and negative evaluative implications) of certain opinion expression.

3. METHODOLOGY

This section shows the conceptual framework of analyzing the public perceptions of K12 implementation in the Philippines using Web Content Mining Techniques and Wrapper Induction Algorithm. Figure 3 shows the conceptual framework of the research.

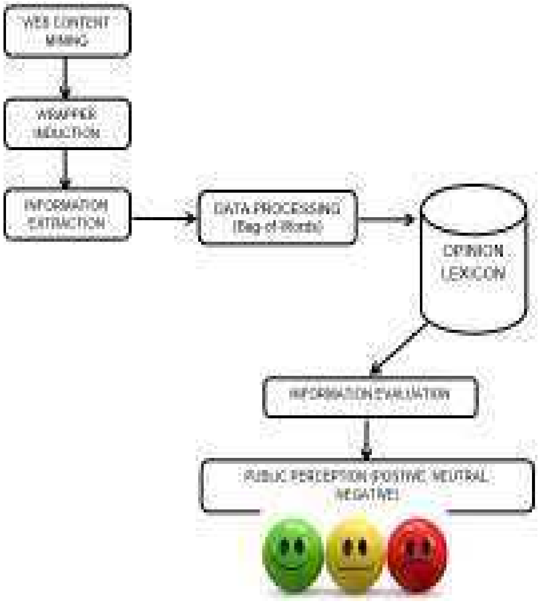

Figure 3.0 Conceptual Framework

3.1. Web Mining and Web Content Mining

This research makes use of web mining in gathering all the related information needed about K12 implementation in the Philippines. According to Lui [20], it aims to discover useful information or knowledge from the web hyperlink structure, page content, and usage data. We used Link Klipper Chrome extension tool to harvest the information we need. The researchers make use of Google [21] as the search engine using the search keywords “K12 Philippines”. There were 318 links being harvested. Web Content Mining is the process of extracting useful information from the contents of Web documents. Information Filtering/Categorization is the retrieval techniques to automatically retrieve, filter, and categorize the extracted information.

3.2 Wrapper Induction

Researchers developed a wrapper that will categorize the gathered information from unstructured to structured web pages. Each web page was extracted to create “bag-of-words”. Wrapper induction (WI) [12] is also known as “wrappers” that will generate extraction rules to easily mine the Web pages.

3.3 Bag-of-Words Model

This research makes use of the extracted information from wrapper induction algorithm. It generates “bag-of-words”, which is the basic unit of information. The document is represented as a bag of words model. It considers the whole web content for each harvested link. The bag of words model, the occurrence of each word will be classified as positive and negative through Opinion Lexicon [20] [22] in order to identify sentiments in each web pages.

3.4 Opinion Lexicon

There are 2,006 positive words and 4, 783 negative words. We used Opinion Lexicon [20] [22] which is a list of identified positive and negative (a – z) words. It includes mis-spelling words, morphological variants, slang and social-media mark-up. These words serve as trainings sets that enable us to identify positive and negative sentiments in a certain website. In this research, two words refers to words with negation word, such as “no”, “not”, “none”, “nothing”, “nowhere”, “neither”, “nobody”, “hardly”, “scarcely”, “barely”. The information was analyzed in order to identify the public perceptions of K12 implementation in the Philippines if it is positive, negative or neutral. Table 1.0 shows some of the harvested links, the number of positive words, number of negative words and its sentiment classification.

Table 1.0 Harvested links with the number of positive and negative words.

	Links	No. of Positive Words	No. of Negative Words	Result
1	https://www.linkedin.com/title/guidance-counselor-at-university-of-san-carlos	6	8	Negative
2	"http://www.spin.ph/basketball/pba/news/alex-compton-on-alaskas-play-in-middle-quarters-of-game-three-that-was-the-most-disappointing-stretch-of-basketball-that-i-recall-see-alaska-play "	6	0	Positive
3	http://www.hellominers.com/viewtopic.php?f=182&t=16611	21	0	Positive
4	https://iprice.ph/k/home-living/	12	0	Positive
5	http://www.golfdebiot.fr/index.php/political-dynasty-philippines-essay	30	4	Positive
6	http://blogs.worldbank.org/education/voices/category/tags/k-12	10	2	Positive
7	http://udyong.gov.ph/teachers-corner/4839-will-k-12-curriculum-improve-the-quality-of-philippine-education	19	3	Positive
8	http://www.bworldonline.com/content.php?section=Economy&title=adb-lends-300m-to-support-k-to-12-program&id=104113	7	0	Positive
9	http://apc.essc.org.ph/content/view/99/7/	16	0	Positive
10	http://www.repglass.net/research-paper-about-k12-education-in-the-philippines/	1	8	Negative

The graph in Figure 2.0 shows result of the perception analysis for K12 implementation in the Philippines. There were 3% of the links were evaluated to have a negative perception, 93% were positive perception and 4% were neutral perceptions.

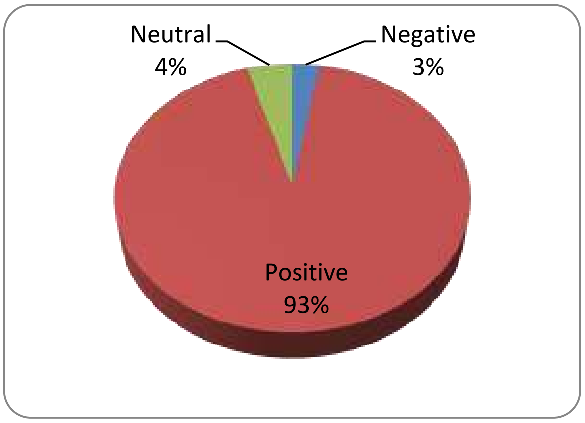

Figure 2.0 Public Perception for K12 Implementation in the Philippines

4. RESULTS

The measurement of performance evaluation is done by using precision and recall of the sentiment prediction.

$$Precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved}$$

(1)

$$Recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents}$$

(2)

The researchers harvested 381 links related to K12 Implementation in the Philippines. Each links were evaluated in the occurrence of positive and negative words. We have found out that out of 318 links there were 88 links considered as “irrelevant links” and 119 links that produces “invalid perceptions” because there were no positive, negative or neutral perceptions being identified. In other words, there were 111 links that enable us to analyze the public perception about K12 implementation in the Philippines. In 111 links, 103 links have positive perceptions, three links have negative perceptions and five links have neutral perceptions. The experiment results show that the approach used in the research achieves 83.94% precision and 72.32% recall.

5. CONCLUSIONS

In this research, the use of web mining enables the researchers to gather information related to K12 implementation in the Philippines. We make use of the web content mining and wrapper induction in order harvest and transform unstructured to structured web contents. In the sentiment classification, the information gathered that produces “bag of words” and were evaluated with Opinion Lexicon was used in the sentiment classification. The result of the research shows us a good level result about the public perceptions of K12 implementation in the Philippines. We aim to extend the system to harvest social networking post and to retrieve sentiments on in portable document format and power point documents.

6. ACKNOWLEDGEMENT

We would like to extend our outmost gratitude to the Department of Computer and Information Science, School of Arts and Sciences and University of San Carlos for giving us the opportunity to conduct this research.

7. REFERENCES

[1] MEDHAT, W., HASSAN, A., AND KORASHY, H. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal 5, (2014), 1093-1113.

[2] FUMKRANZ, J. (n.d.). Web Mining.

[3] RASCHKA, S. Naive Bayes and Text Classification I-Introduction And Theory. ARXIV PREPRINT ARXIV 1410, 5329 (2014).

[4] MEDHAT, W., HASSAN, A., AND KORASHY, H. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal 5, (2014), 1093-1113.

[5] KOSALA, R., BLOCKEEL, H., & NEVEN, F. (n.d.). Web Mining 5.6.

[6] WANG, Y., (2000), Web Mining and Knowledge Discovery of Usage Patterns. Cs 748T Project. (2000), 1-25.

[7] KOSALA, R., AND BLOCKEEL, H. Web Mining Research: A Survey. ACM Sigkdd Explorations Letter 2, 1 (200), 1-15.

[8] SIDDIQUI, A., & ALJAHDALI, S. (2013, May). Web Mining Techniques in E-Commerce Applications. International Journal of Computer Applications, 8(8), 39-43.

[9] YI, L., & LUI, B. (n.d.). Web Page Cleaning for Web Mining through Feature Weighting.

[10] UPADHYAY, G., & DHINGRA, K. (2013, November). Web Content Mining: Its Techniques and Uses. International Journal of Advanced Research in Computer Science and Software Engineering, 3(11), 610-613. doi:2277 128X

[11] COOLEY, R.; MOBASHER, B.; SRIVASTAVA, J.; Web mining: information and pattern discovery on the World Wide Web. Tools with Artificial Intelligence,1997. Proceedings.

[12] KUSHMERICK N., Wrapper induction for Information Extraction, PhD Thesis, Department Of computer Science, Univ. Of Washington (1997).

[13] KUSHMERICK, N. (1999, March 10). Wrapper Induction: Efficiency and Expressiveness. Artificial Intelligence 118 (2000) 15–68. Retrieved August 7, 2016.

[14] MANSOURI, A., AFFENDEY, L., & MAMAT, A. (2008, February). Named Entity Recognition Approaches. IJCSNS International Journal of Computer Science and Network Security, 8(4), 339-344. Retrieved August 8, 2016.

[15] LIU, B. Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data.
2nd ed. 2011, XX, 622 p.

[16] SOUZA, M., VIEIRA, R., BUSETTI, D., CHISHMAN, R., AND ALVES, I. M. Construction of a portuguese opinion lexicon from multiple resources. STIL, (2011).

[17] HU, M., AND LIU, B. Mining and summarizing customer reviews. In *KDD '04* Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (ACM, New York, NY, USA, August 2004).

[18] FLEKOVA, L., RUPPERT, E., AND PREOTIU-PIETRO, D. Analysing domain suitability of a sentiment lexicon by identifying distributionally bipolar words. In 6th Workshop on Computational Approaches to Subjectivity, Sentiment And Social Media Analysis WASSA (2015).

[19] CRUZ, F. L., TROYANO, J. A., ORTEGA, F. J., AND ENRÍQUEZ, F. Automatic expansion of feature-level opinion lexicons. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (June 2011).

[20] LIU, B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer-Verlag Berlin Heidelberg, 2011.

[21] SRIVASTAVA, T., DESIKAN, P., AND KUMAR, V. Web mining–concepts, applications and research directions. Foundations and advances in data mining, (2005), 275-307.

[22] LIU. B., HU, M., AND CHENG, J. Opinion observer: analyzing and comparing opinions on the Web. In *WWW '05: Proceedings of the 14*th international conference on World Wide Web (ACM, New York, NY, USA, May 2005).