INTELLIGENT AGENT FOR PUBLICATION AND SUBSCRIPTION PATTERN ANALYSIS OF NEWS WEBSITES

W.D.R Wijedasa and Chathura De Silva

Department of Computer Science Engineering, Faculty of Engineering University of Moratuwa, Sri Lanka

ABSTRACT

The rapid growth of Internet has revolutionized online news reporting. Many users tend to use online news websites to obtain news information. When considering Sri Lanka, there are numerous news websites, which are subscribed on a daily basis. With the rise in this number of news websites, the Sri Lankan authorities of media face the issue of lacking a proper methodology or a tool which is capable of tracking and regulating publications made by different disseminators of news.

This paper proposes a News Agent toolbox which periodically extracts news articles and associated comments with the aid of a concept called Mapping Rules; to classify them into Personalized Categories defined in terms of keywords based Category Profiles. The proposed tool also analyzes comments made by the readers with the aid of simple statistical techniques to discover the most popular news articles and fluctuations in popularity of news stories.

KEYWORDS

News Articles; Mapping Rules; Personalized Classification; Category Profiles

1.INTRODUCTION

The rapid growth of Internet and the increased confidence on digitized information have revolutionized online news reporting. Such improvements provide efficient means and ways such as news websites and news blogs, for disseminating news information much quicker than ever before. The main objective of online news reporting is to provide up-to-date news as well as to increase news consumption and its usages.

When considering Sri Lanka, there are large number of news websites and news blogs in all three mediums, Sinhala, English and Tamil. These news websites and blogs publish news related to local issues, politics, events, celebrations, people, business, weather and entertainment. Such news websites are subscribed on a daily basis for up-to-date information. With the growth in this number of news websites and news blogs, the Sri Lankan authorities of media face the issue of lacking a proper methodology or a tool which is capable of carrying out the following. Identifying frequently published news topics, finding out topics related to a given set of keywords, identifying topics which increase popularity with respect to time and identifying news subscription patterns of users in terms of spotting out the regular users, which topics they subscribe most and whether they actively express their ideas via commenting.

Lacking of a proper methodology or a tool with previously mentioned capabilities has led the responsible parties of media to face several issues. Among these issues, difficulty of tracking and regulating publications made by different disseminators of news takes a major place. This major issue has led to several sub issues such as difficulty in identifying incidents which require more attention and difficulty in maintaining the fairness, accuracy and balance of coverage under each publication. Moreover the difficulty in regulating publications made by different parties in a way in which it meets the demands of the readers can be considered as another major issue.

A media system within a country directly influences its society in various ways and plays an important role in the economy of the country. It has been argued that a person's view is influenced more by the media system within the country than by his or her personal experiences. Additionally the public opinion and the awareness are also affected by the means by which news is being reported [1].

When considering the media system of Sri Lanka, online news reporting also plays a significant role that is equivalent to traditional paper based news media. As a result of the significant role played by the news websites within the Sri Lankan media system, it highlights that these online news websites and news blogs need to balance reporting of news while providing equal coverage on each and every view. Moreover fairness and accuracy of covered topics are also considered as equally important qualities. Such qualities can be maintained by solving previously mentioned issues faced by the Sri Lankan authorities of media.

This paper proposes an Intelligent Agent for publication and subscription patterns analysis of news websites and associated reader comments, using Personalized Classification [2] and simple statistical techniques.

2.PREVIOUS WORK

Media analysis has remained as a vast research domain, in social sciences for several decades. Traditional media analysis has been based on a process known as "coding", which involves manual analysis and annotation of small sets of news articles[1][3].

As the result of exponential growth of the World Wide Web and the Internet, online news reporting has been modernized with the most efficient ways of reporting news. Therefore the users and authorities of media are now experiencing overwhelming quantities of readily available news content which grows day by day. This has resulted in a great difficulty for the responsible parties of media to analyze the news content via coding.

A variety of research efforts have been carried out to provide an opportunity to monitor a vast number of news outlets, constantly and in an automated way [3] [4] [5] [6]. The required automation of analysis has been realized by the utilization of Artificial Intelligence (AI) techniques from the fields of Natural Language Processing (NLP), Text mining, Text Analytics and Machine Learning (ML) [3] [4].

Some of the existing automations of analysis are News Outlets Analysis and Monitoring (NOAM) [3], Lynda [4], News Analysis System (NAS) [5], NewsBlaster [6] and Google News. NOAM [3] is a data management system which gathers and monitors multi-lingual news content from 21 countries. Lynda [4] is a multi-purpose system which focuses mainly on the US media system. It has been developed to detect spatial and temporal distributions of Named Entities. NAS [5] is a prototype news analysis system which classifies and indexes news stories in real time while NewsBlaster [6] is a robust news tracking and summarization system of Colombia.

Extraction of news articles from various semi-structured and heterogeneous online sources is a major activity and a huge challenge for above mentioned automated systems. Various research efforts have been taken place to come up with algorithms and approaches for extracting news articles from online sources. Among such approaches, work in [7] has proposed a system called HTML2RSS which extracts content from HTML Web pages based on Document Object Model (DOM) structure to generate RSS files. Authors have introduced a concept called Mapping Rules [7] for extraction purposes. A Mapping Rule is defined as XML file containing XPath information, which specifies from which node of the DOM tree the content should be extracted. These Mapping Rules have been generated using manual help. A GUI has been provided for users to mark HTML elements such as text and links. Subsequently the XPath information of the selected elements has been extracted automatically by the system.

Most of the automated systems such as NOAM [3], NAS [5], NewsBlaster [6] and Google News have utilized Clustering or Classification for analyzing the extracted news articles. Grouping news articles into stories describing a similar event is a major activity in such systems. Various research efforts have been taken place in order to identify the benefits of existing algorithms and to propose new approaches for grouping news articles. Among such efforts a novel classification approach named Personalized Classification has been presented in [2] [8] for classifying news documents. Personalized Classification enables users to define their own categories of interest on the fly, with the aim of automating the assignment of documents to such categories. The authors have used a set of keywords to define each Personalized Category and have utilized Support Vector Machines (SVM) to perform the classification. The training documents for Personalized Categories have been obtained from a pool of training documents with the use of a search engine and the set of keywords defining each category. Subsequently a classifier per Personalized Category has been created and trained for grouping the articles into stories.

3.METHODOLOGY

Section 3 will describe the approach carried out by this research to address the important issues identified in section 1, by coming up with a News Agent Toolbox. The proposed tool has been fundamentally developed to periodically analyze patterns and visualize the results for surveillance purposes.

The news content analysis process of the tool follows a three step methodology, namely the news content acquisition, pre-processing and analysis, as in existing systems such as NOAM. A probe was developed for periodical acquisition of news content, associated reader comments and news article sharing information made on social networks, from a set of pre-defined Sri Lankan news websites. The pre-processing step involves tokenizing, stop-word removal, stemming and representing the news content in a suitable form. The content analysis step involves major activities, such as classifying news articles into Personalized Categories [2] [8] and analyzing reader comments and sharing information associated with news articles. The high level architecture of the proposed tool is depicted in Figure 1.

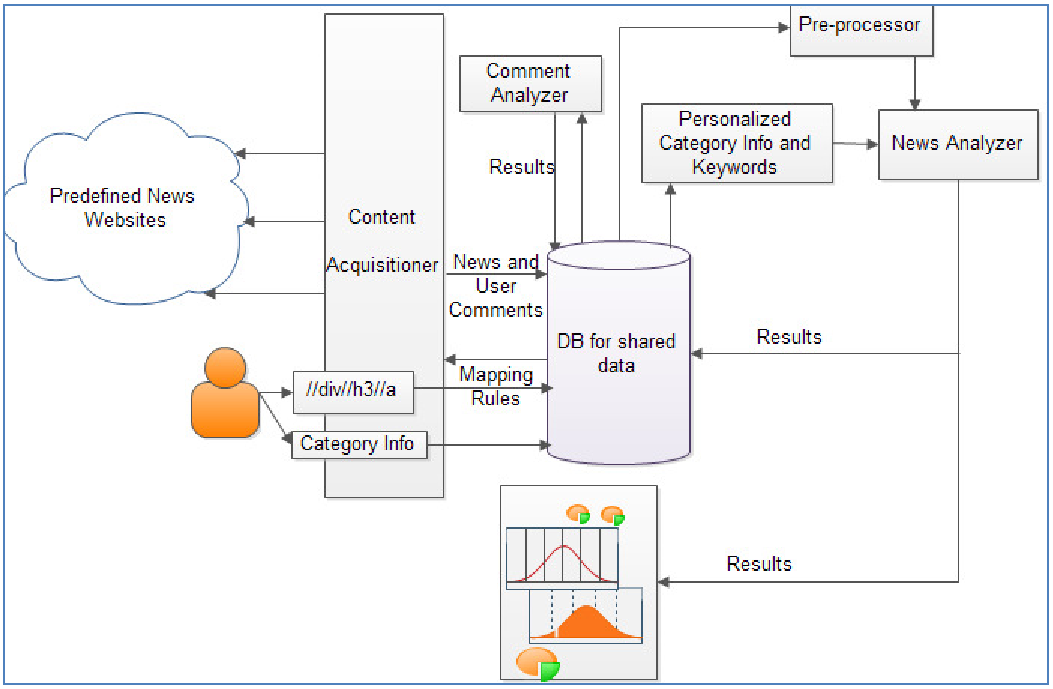

Figure 1. High level architecture of the proposed system

The tool has been designed using a modular architecture where each module is specialized for a particular task. According to the three steps methodology, the tool has been designed with four major modules namely the Content Acquisitioner, Pre-processor, News Analyzer and Comments Analyzer. Moreover a user friendly GUI and a visualization tool has been attached to the system to support user interaction and visualization of results.

3.1.Content Acquisitioner

The Content Acquisitioner module was developed to extract news articles, associated user comments and sharing information on social networks, from a set of predefined English news websites and news blogs.

Several challenges were encountered while developing the extraction functionality of the Content Acquisitioner module. Non-uniform structures of news websites and variations in organization of news articles take a major place among these challenges. Each news website has its own structure and its own way of organizing news articles. Such arrangements lead to difficulties in coming up with an extraction methodology which suits all news websites with different structures. Existence of news websites which do not follow specifications and websites with malformed html content is also challenging. Moreover, many news websites place additional content such as videos, advertisements and navigational elements in between news content, making the extraction process of full text news content a complicated task.

In order to minimize the impact of above mentioned challenges and to come up with an extraction methodology which suits many sites with different arrangements, the work proposed in [7], Mapping Rules were used for each news website in order to extract navigational links that point to full stories of news articles.

3.1.1.News Content Extraction

A four step methodology was used to extract news articles from a set of user provided news websites.

Step 1:XPath based Mapping Rule generation for each website

As the first step of the extraction process, Mapping Rules [7] per news website will be generated with the manual assistance. The user desired news website will be loaded via the tool. The user will be provided with a functionality to select sample links pointing to news articles, within web pages where news articles are required to be extracted. These selected sample links will be used to distinguish links pointing to news articles from links pointing to additional content such as videos, advertisements and etc. Subsequently the XPath of the user selected sample links will be determined by the tool. Next the identified XPath will be processed in a way that can be used to aid the extraction of all news links appearing in similar paths within a given web page.

Sample XPath expression identified for a selected news link.
/html /body /div [3] /div [10] /div [2] /div [4] /div [2] /h5 /a

Sample XPath expression which has been processed to aid the extraction of news links appearing in similar paths.
/html /body /div /div /div /div /div /h5 /a

Step 2: Creation of DOM trees for web pages containing news article links

A DOM tree structure is generated for each webpage from which the articles are required to be extracted, using HtmlAgilityPack [9] parser.

Step 3: Extraction of links pointing to news articles based on Mapping Rules

Links pointing to news articles which fall under given XPath based Mapping Rules are extracted by traversing through DOM trees of the web pages.

Step 4: Extraction of full text news content from news articles

Nboilerpipe [10] library, developed by Christian Kohlschütter, was used for extracting textual news content from news articles pointed by the links extracted in the previous step.

3.1.2. Reader Comments Extraction

A three step methodology was used to extract the author of the comment, commented date and the comments associated with the extracted news articles.

Step 1: XPath based Mapping Rule generation

As the first step of the comments extraction process, three separate Mapping Rules will be generated for each website, in order to aid the extraction of commented authors together with the commented dates and the comments.

Step 2: Creation of DOM trees for news articles containing user comments

A DOM tree structure is generated for each news article webpage.

Step 3: Extraction of comments from news articles based on Mapping Rules

The commented authors and their corresponding comments and the commented dates within each article web page are extracted by traversing through DOM trees.

Many news websites facilitate users to share news articles via social networks such as Facebook, Twitter, LinkedIn, Google plus and etc. Social sharing allows users to explicitly publish the URL of a news article or to like the URL of a news article. The statistics of such social sharing information on Facebook, Twitter, LinkedIn and Google plus were extracted utilizing below described two step methodology [11].

For each extracted news article the corresponding social network API call was made to the given endpoint incorporating the article's URL. In next step the retrieved response in terms of JSON objects were queried to extract the counts of likes and shares made for the extracted news article.

3.2. Pre-processor

The major functions of the pre-processor module are to perform an initial cleansing of each article, tokenizing, removal of stop-words and stemming the content of news articles. The Porter stemming algorithm by Martin Porter [12], developed based on an explicit list of suffixes together with the criterion under which each suffix to be removed from a word was utilized for stemming news articles.

3.3. News Analyzer

The main functionality of the News Analyzer component is to classify news articles into a set of predefined news stories defined under general categories such as politics, business, health and etc. The approach adopted for the classification process is Personalized Classification presented in work [2] and [8].

3.3.1. Personalized Classification of News Articles

Personalized Classification is a process which enables users to create their own classification categories on the fly. This classification methodology is capable of classifying documents under diverse user interests, compared to general classification process with a fixed set of pre-defined categories.

The News Analyzer component which utilizes Personalized Classification approach follows a three step methodology for creating personalized categories. As the first step it enables an administrative user of the tool to define general categories such as politics, business, health, sports and etc. Subsequently it enables the user to define more specific news stories (Personalized Categories), under each general category. As the final step the user will be allowed to define a set of keywords to describe each personalized news category. A personalized category defined in terms of keywords will be named as keyword based Category Profile [2] [8]. Keyword based Category Profiles simplifies the user effort, compared to selecting adequate number of training documents for each Personalized Category.

3.3.2. Keyword based Category Profiles

In this research each Personalized Category is defined in terms of a list of user specified keywords instead of using a set of training documents. The Personalized Classification process based on keyword Category Profiles, would work optimally when the provided keywords have a high discriminatory power in distinguishing news articles under a desired category from the articles under the other categories.

In order to facilitate a higher discriminating power for keywords, a weighting scheme tabulated in Table 1 was introduced. The proposed system enables users to specify a higher weight on keywords which may occur infrequently and a lower weight on certain keywords which may occur frequently in many news articles.

Table1. Weighting scheme used for keywords

Weight	0.2	0.4	0.6	0.8	1.0
Indication	Very Low	Low	Medium	High	Very High

A sample Personalized Classification category will be illustrated below.

***General Category**: HEALTH*
***Personalized Category**: DCD and Whey Protein products detected*
***Occurrence of Keywords**: Multiple Occurrences*
Keywords and Weights:

Dicyandiamide	*1.0*
DCD	*1.0*
Whey Protein	*1.0*
Botulism	*1.0*
Clostridium	*1.0*
Milk powder	*0.4*

3.3.3. Classification Rank Calculation

This research utilizes a classification rank calculation methodology for the classification process of news articles. The classification rank is an indication of how well, a given news article is connected to the classified Personalized Category. According to the classification rank calculation methodology, initially the news articles will be classified into two categories namely the user desired Personalized Category or OTHER category. All news articles with a classification rank above zero will get assigned to the user desired Personalized Category and the articles with a zero classification rank will get assigned to the category of OTHER.

The proposed classification process is depicted in Figure 2.

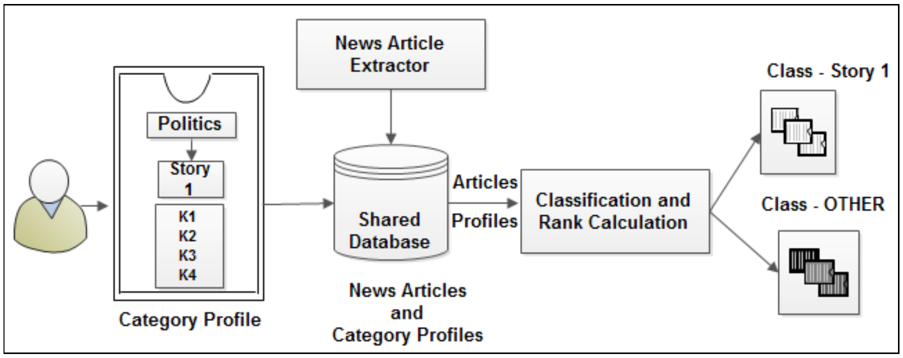

Figure 2. Personalized Classification process

The classification rank of each news article is calculated based on the *tf-idf* values of the keywords describing the desired Personalized Category. The *tf-idf* weight of a keyword $\mathbf{k_i}$ in a news article $\mathbf{a_j}$ is computed from the term frequency **Freq** $(\mathbf{k_i}, \mathbf{a_j})$ and the inverse document frequency as provided in Eq. 1 [2].

$$\textbf{\textit{tfidf}}\,(\mathbf{k_i, a_j}) \quad = \mathbf{Freq}\,(\mathbf{k_i, a_j}) \;*\; \mathbf{log_2\ N / DF(k_i)} \qquad \mathbf{(1)}$$

According to the Eq.1, **N** is the number of news articles extracted on the user specified date and $\mathbf{DF(k_i)}$ is the number of news articles in the article collection N, having keyword $\mathbf{k_i}$ occurring at least once.

The classification rank of an article is defined in terms of the summation of the *tf-idf* weights of all keywords describing the desired Personalized Category denoted by **cp** [2].

$$\mathbf{rank(a_j)} = \sum_{\mathbf{k\epsilon\ cp}} \textbf{\textit{tfidf}}\,(\mathbf{k_i, a_j}) \qquad \mathbf{(2)}$$

Three additional equations for calculating ranks were derived utilizing Eq.1 and Eq.2 with the aim of increasing the accuracy of the classification process and reducing the inaccurate articles being classified into desired categories with high rank values.

The Eq.3 will calculate weighted *tf-idf* value of a keyword $\mathbf{k_i}$ in a news article $\mathbf{a_j}$ by incorporating weights defined for the keywords describing the desired Personalized Category. This aims to reduce the rank values of articles having one to two keywords with high frequency, which might not exactly describe the desired news story.

$$\textbf{\textit{wgt\_tfidf}}\,(\mathbf{k_i, a_j}) = [\boldsymbol{W_i} * \mathbf{Freq}\,(\mathbf{k_i, a_j})] \;*\; \mathbf{log_2\ N / DF(k_i)} \qquad \mathbf{(3)}$$

According to Eq.3 $\boldsymbol{W_i}$ will denote the user defined weight given for keyword $\mathbf{k_i}$ describing the Personalized Category. Based on Eq.3 weighted summation and weighted average based rank calculations were derived and will be denoted by Eq.4 and Eq.5 respectively.

$$\mathbf{wgt\_rank(a_j)} = \sum_{\mathbf{k\epsilon\ cp}} \textbf{\textit{wgt\_tfidf}}\,(\mathbf{k_i, a_j}) \qquad \mathbf{(4)}$$

$$\mathbf{wgt\_avg\_rank(a_j)} = \sum_{\mathbf{k\epsilon\ cp}} \textbf{\textit{wgt\_tfidf}}\,(\mathbf{k_i, a_j}) / \mathbf{total}(\boldsymbol{W_i}) \qquad \mathbf{(5)}$$

According to Eq.5, **total**($\boldsymbol{W_i}$) will denote the total of the user defined Weights, of the keywords describing the desired Personalized Category. The classification process of the proposed system was evaluated by calculating the classification rank of news articles utilizing Eq.5.

3.4. Comment Analyzer

The main functionality of the Comments Analyzer component is to analyze the most popular news articles published within a given period, graphically plot sudden fluctuations in popularity of news stories and to provide a detailed profile for each commenter.

3.4.1. Popularity Score for News Articles

The Comments Analyzer component enables users to obtain a list of most popular news articles which have been published within a user desired period. Popularity of each article was calculated utilizing the statistics of the below factors.

- Number of unique users who have made comments.
- Number of days the article has received comments.
- Number of times that the news article has been shared, liked and commented via Facebook.
- Total number of times that the article has been shared via Twitter.
- Total number of times that the article has been shared via Google plus.
- Total number of times that the article has been shared via LinkedIn.

A popularity score was calculated for each news article utilizing a four steps methodology. As the first step statistics of above listed factors were calculated utilizing the comments and sharing information extracted via the Content Acquisitioner module. Subsequently the calculated statistics were analyzed with the aid of IMB SPSS Statistics [13] toolkit in order to come up with weights for each factor. As the third step, the weights obtained from the previous step were utilized to calculate a popularity score for each article using Eq.6.

$$\textbf{popularity\_score = (Weight[i] *Factor[i]) /age} \qquad \textbf{(6)}$$

According to the Eq.6 **Weight[i]** will denote the weight of the **i** th factor, where **Factor[i]** will denote the total number of cases available for the **i** th factor for a given news article. The number of days since the article was published will be denoted via variable **age**.

As the final step, the articles will be ranked according to the popularity scores and the top most **k** articles will be presented to the users of the News Agent toolbox.

3.4.2. Profile for Commenters

The Comments Analyzer module maintains user profiles for commenters of news websites. The main purpose of maintaining user profiles is to present various statistics related to comments made by users and their commenting behaviour.

3.4.3. Popularity Analysis of News Stories (Personalized Categories)

The main purpose of analyzing the popularity of news stories is to provide a graphical visualization of sudden increases in popularity of news articles belonging to a specific news story (Personalized Category). The popularity analysis of news stories will follow a four step methodology as described below.

As the first step the tool will classify the news articles based on the user provided period and the news story (Personalized Category). As the next step the tool will calculate the total number of comments, Tweets, Facebook shares, LinkedIn shares and the Google plus shares that the classified articles belonging to the given news story has received. Subsequently the tool will calculate sudden increases in popularity by considering the slope (the rate of change) between the two data points which represent the total number of comments and sharing statistics. As the final step the tool will plot how the popularity of the news story has been change over the user given period.

4. RESULTS

The proposed system was evaluated utilizing the accuracy of news article extraction process of Content Acquisitioner and the classification results of the News Analyzer modules. The extraction accuracy was simply evaluated using the percentages of correctly and incorrectly extracted articles. The classification results were evaluated by comparing the results obtained via the system with the results obtained via a manual classification process.

The Weighted Average rank calculation given by Eq.5, was utilized for the classification evaluation process.

4.1. Article Extraction Accuracy

The accuracy of the News Article Extraction process was evaluated utilizing data, which have been extracted in months of October and November. Around six thousand full text news articles were manually examined in order to identify news articles and links which have not been extracted as expected. The news articles which included only the commenting sections made by the users, the text content placed on navigator sections, the text content of footer sections and news links pointing to advertisements, leave comments and contact us web pages were considered as articles and links which have been extracted incorrectly.

The Figure 3 will illustrate the number of news articles which have been extracted accurately and inaccurately considering the total number of news articles extracted per day between the chosen time period.

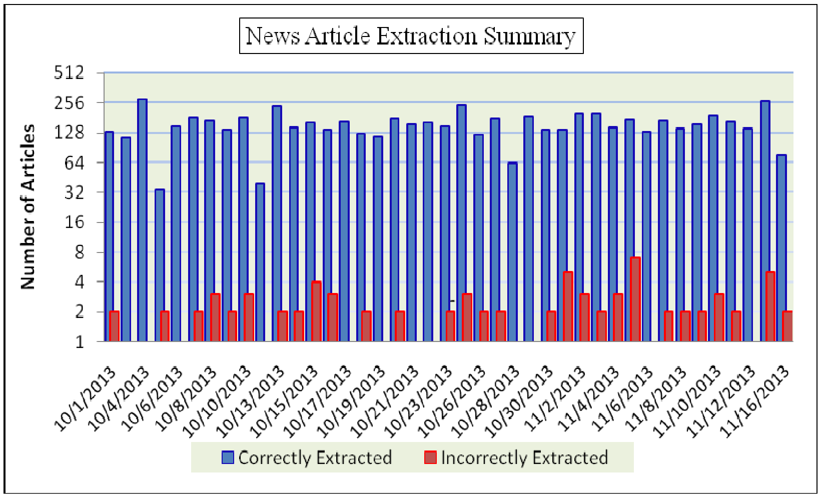

Figure 3. Accuracy of article extraction process

4.2. Article Classification Accuracy

The classification process of the News Analyzer component was evaluated creating several timely news stories (Personalized Categories) under three main categories such as Politics, Health and Business. The classification accuracy of each news story was measured considering the period of time where multiple news websites have reported the related news event.

The following keyword based Category Profile was utilized to obtain test results for news story "*Commonwealth (CHOGM) 2013*".

General Category: *POLITICS*
Personalized Category: *Commonwealth (CHOGM) 2013*
Occurrence of Keywords: *Multiple Occurrences*
Keywords and Weights:

Commonwealth	*1.0*
CHOGM	*1.0*
Commonwealth Heads Government Meeting	*1.0*
Commonwealth Secretariat	*1.0*
Commonwealth Leaders	*1.0*
Summit	*0.4*

Figure 4 will illustrate the test results obtained for the classification process of news articles related to the news story "*Commonwealth (CHOGM) 2013*". The graph illustrates the number of news articles classified by the News Agent tool box compared to the number of news articles classified manually. The gap between the two will indicate the number of incorrect articles which have been classified by the proposed tool.

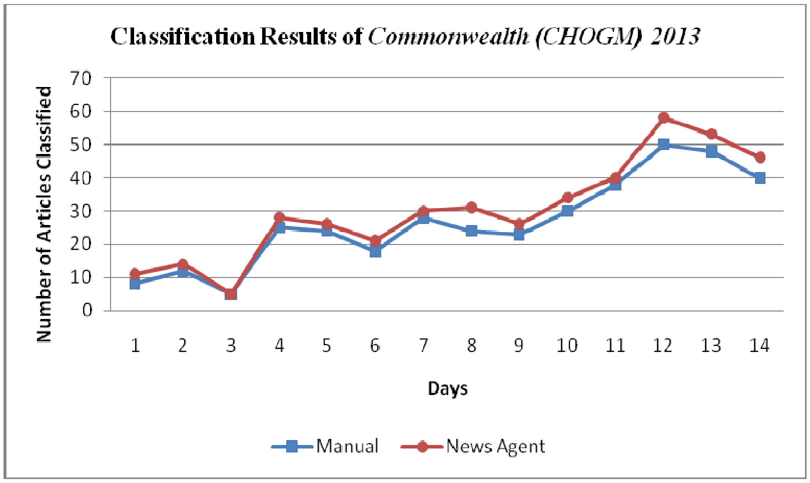

Figure 4. Test results obtained for Commonwealth (CHOGM) 2013

The following keyword based Category Profile was utilized to obtain test results for news story "*Whey Protein detected in dairy products*".

***General Category**: HEALTH*
***Personalized Category**: Whey Protein detected in dairy products*
***Occurrence of Keywords**: Multiple Occurrences*
Keywords and Weights:

Whey Protein	*1.0*
Botulism	*1.0*
Clostridium	*1.0*
Milk powder	*0.4*
Dairy products	*0.4*
Bacteria	*0.4*

Figure 5 will illustrate the test results obtained for the classification process of news articles related to the news story "*Whey Protein detected in dairy products*".

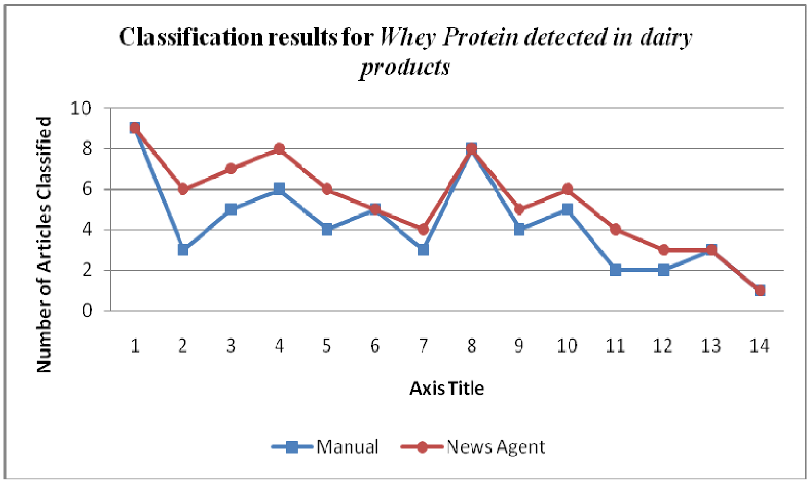

Figure 5. Test results obtained for Whey Protein detected in dairy products

The following keyword based Category Profile was utilized to obtain test results for news story "*Central Bank related news*".

***General Category**: BUSINESS*
***Personalized Category**: Central Bank related news*
***Occurrence of Keywords**: Multiple Occurrences*
Keywords and Weights:

Central Bank	*1.0*
CB	*1.0*
Monitor	*0.2*
Report	*0.4*

Figure 6 will illustrate the test results obtained for the classification process of news articles related to the news story "*Central Bank related news*".

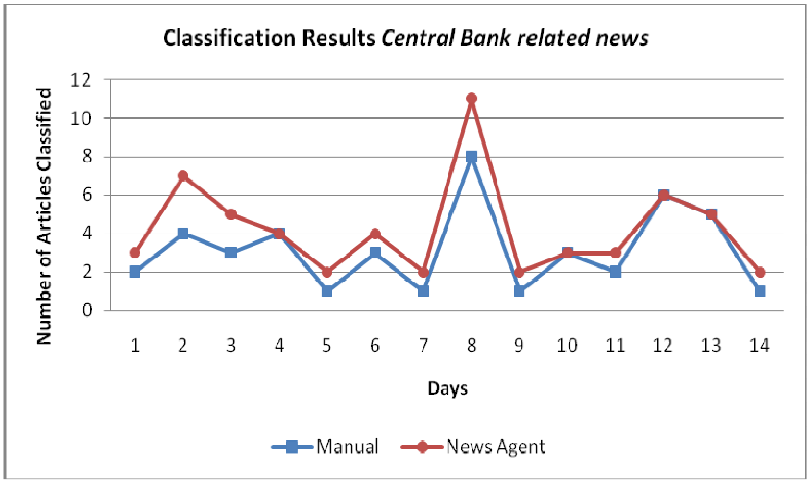

Figure 6. Test results obtained for Central Bank related news

According to the test results obtained for classifying news articles into above mentioned news stories, it is clear that the accuracy of the proposed system is much closer to the results obtained via the manual classification process. According to the conducted evaluations, the News Analyzer module functions with an approximate accuracy of 81% to 82%.

4.2.1 Discriminative Power of Keywords vs. Accuracy

The accuracy of the classification process is dependent upon the degree of the discriminating power of keywords defined under Category Profiles of the news stories. The higher the keywords are capable of distinguishing news articles of one news story from another, more accuracy can be obtained by the proposed classification process.

The test results obtained by the following sample test scenarios show that Category Profiles with keywords having high discriminating power provide more accuracy and Category Profiles with keywords having low discriminating power provide less accuracy. Figure 7 and Figure 8 will illustrate how the classification accuracy has been altered when the discriminating power of the keywords are changed.

The following keyword based Category Profile was utilized to obtain test results for news story "*Dengue outbreaks in Sri Lanka*" with high discriminating power keywords.

General Category: *HEALTH*
Personalized Category: *Dengue outbreaks in Sri Lanka*
Occurrence of Keywords: *Multiple Occurrences*
Keywords and Weights:

Dengue	*1.0*
Dengue mosquito	*1.0*
Dengue virus	*1.0*
Dengue Hemorrhagic Fever	*1.0*
Fever	*0.4*
Mosquito breeding	*0.4*

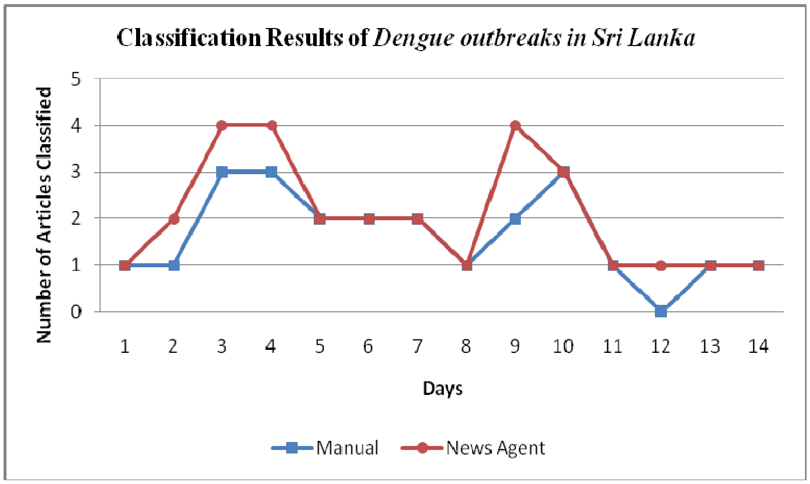

Figure 7. Results for keywords with high discriminating power

The following keyword based Category Profile was utilized to obtain test results for news story "*Dengue outbreaks in Sri Lanka*" with low discriminating power keywords.

General Category*: HEALTH*

Personalized Category*: Dengue outbreaks in Sri Lanka*

Occurrence of Keywords*: Multiple Occurrences*

Keywords and Weights:

Dengue	*1.0*
Mosquito	*1.0*
Virus	*1.0*
Fever	*0.4*
Breeding	*0.4*

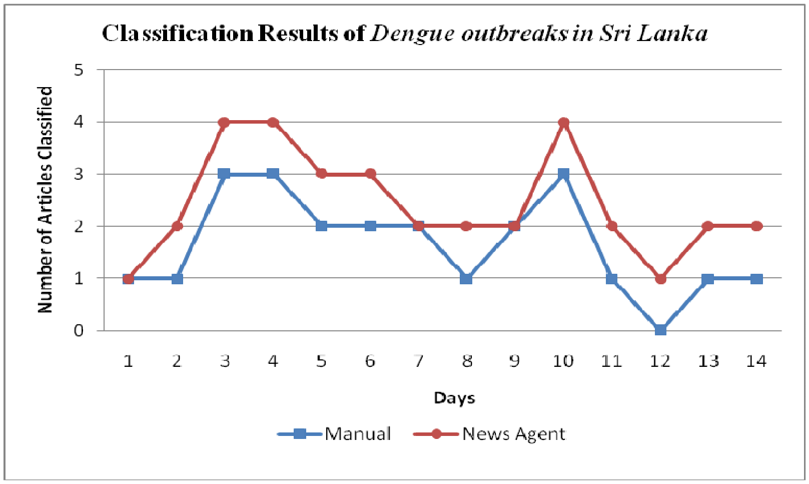

Figure 8. Results for keywords with low discriminating power

5. CONCLUSIONS

In summary, this research attempts to develop an Intelligent Agent, for analyzing readily available news content and the subscription patterns of Sri Lankan news websites and news blogs. The aim of this research is to take proper initiatives to mine heavily available online news content and the subscription data, in order to provide an efficient way to surveillance different publication and subscription patterns. The main objectives of this research are to assist the Sri Lankan authorities of media to track and regulate different publications made by different disseminators of news, to assist identifying incidents which requires more attention and news events increasing popularity over time, to assure proper balance, accuracy and equal coverage of news reporting.

The proposed system has been evaluated utilizing the accuracy of news article extraction process and the classification process. The extraction accuracy has been simply evaluated using the percentages of correctly and incorrectly extracted articles. The classification results have been evaluated using several news stories together with the keyword profiles for general categories such as Politics, Health and Business. The test results show that the extraction process functions with a 98% of accuracy and the classification process functions with an approximate accuracy of 81%.

ACKNOWLEDGEMENTS

I would like to convey my sincere gratitude to my supervisor, Dr. Chathura De Silva, Senior Lecturer, Department of Computer Science and Engineering, University of Moratuwa for providing the valuable research idea, guidance, constructive comments and extensive support throughout this research.

I also express my sincere gratitude to our MSc Project Coordinator Dr A. Shehan Perera, Senior Lecturer, Department of Computer Science and Engineering, University of Moratuwa for providing extensive support and encouragement by conducting weekly sessions, forums and progress presentations throughout the year.

REFERENCES

[1] I. Flaounas, ‘Pattern analysis of news media content’, University of Bristol, 2011.

[2] A. Sun, E.-P. Lim, and W.-K. Ng, ‘Personalized classification for keyword-based category profiles’, in Research and Advanced Technology for Digital Libraries, Springer, 2002, pp. 61–74.

[3] I. Flaounas, O. Ali, M. Turchi, T. Snowsill, F. Nicart, T. De Bie, and Cristianini, ‘NOAM: news outlets analysis and monitoring system’, in Proc. of the 2011 ACM SIGMOD international conference on Management of data, 2011, pp. 1275–1278.

[4] L. Lloyd, D. Kechagias, and S. Skiena, ‘Lydia: A system for large-scale news analysis’, in String Processing and Information Retrieval, 2005, pp. 161–166.

[5] R. J. Kuhns, ‘A news analysis system’, in Proceedings of the 12th conference on Computational linguistics-Volume 1, 1988, pp. 351–355.

[6] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, S. Sigelman, and M. Summarization, Tracking and Summarizing News on a Daily Basis with Columbia’s Newsblaster, Proceedings of Human Language Technology Conference. San Diego, USA, 2002.

[7] Geng, Q. Gao, and J. Pan, ‘Extracting content for news web pages based on DOM’, IJCSNS International Journal of Computer Science and Network Security, vol. 7, no. 2, pp. 124–129,2007.

[8] C.-H. C. A. S. Ee and P. Lim, ‘Automated online news classification with personalization’, in 4th international conference on Asian digital libraries, 2001.

[9] "Parsing HTML Documents with the Html Agility Pack" [Online]. Available: http://www.4guysfromrolla.com/articles/011211-1.aspx. [Accessed: 29-Apr-2013].

[10] Kohlschütter, P. Fankhauser, and W. Nejdl, 'Boilerplate detection using shallow text features', in Proceedings of the third ACM international conference on Web search and data mining, New York, NY, USA, 2010, pp. 441–450.

[11] 'Get Social Share Counts - A Complete Guide', CUBE3X. [Online]. Available: http://cube3x.com/2013/01/get-social-share-counts-a-complete-guide/. [Accessed: 10-Oct-2013].

[12] 'Porter Stemming Algorithm'. [Online]. Available: http:// tartarus.org /martin /PorterStemmer/index.html. [Accessed: 03-Jun-2013].

[13] 'IBM SPSS Statistics', 29-Oct-2013. [Online]. Available: http://www-01.ibm.com/software/analytics/spss/products/statistics/. [Accessed: 14-Oct-2013].

Authors

W.D.R Wijedasa received her M.Sc. Degree in Computer Science Engineering fro m University of Moratuwa Sri Lanka in 2014. She has been working as a Software Engineer at IFS RnD Pvt Limited from 2010. Her research interests are Data mining, Artificial Intelligence and Agent Technologies.

Dr. Chathura De Silva received his MEng. Degree and Ph.D. Degree from National University of Singapore. He has been working as a Senior Lecturer at Department of Computer Science E ngineering University of Moratuwa Sri Lanka. He is the current Head of Department of Computer Science Engineering University of Moratuwa.