JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

A SYSTEMATIC APPROACH ON EXTRACTION, SEGMENTATION AND CLASSIFICATION OF WEB AND SOCIAL MEDIA DATA

Rajashekar Nennuri[1], B. Mohammed Ismail[2]

[1]Research Scholar, Department of Computer Science Engineering, Koneru Lakshmaiah Education Foundation Deemed to be University, A.P, India. E-mail: rajasekharnennuri@gmail.com
[2]Professor, Department of Computer Science Engineering, Koneru Lakshmaiah Education Foundation Deemed to be University, A.P, India.

Received: 23.04.2020 Revised: 24.05.2020 Accepted: 21.06.2020

ABSTRACT: Health care industry is in high pressure to decrease the cost of expenditure and managing resource to improve patient care. By adding to his dramatic change in increase of chronic diseases and increase in population, consumer expectation changes in purchase and receives care and increase in usage of social media and technologies in mobile are modifying the healthcare obtaining and delivering ways. Now a day's all the industries are beginning with a data and analytics of digital information belongs to same industries to improve service and reduce cost. As there is a wide range in health care data in public domain platforms regarding the population, patient and people profile with their economical, medical and operative process. Social media is acting key role to decrease problems associated with healthcare problems. Social media platforms are having millions of health related problems. Developing tools is the major source to reduce these problems in terms of cost of efficiency. All we have to do is reducing the cost and making efficient health classifiers to analyze the data. We are proposing a classifier model which analyzes the patient condition related to ailment and also suggests for the related treatment. For this a supervised classifier model for health is proposed which conducts survey on patient healthcare. Ailments are also classified based on exact symptoms of illness using advanced clustering approaches in those models classified twitter model works more effectively by comparing risks than a normal traditional model.

KEYWORDS: Data extraction, Segmentation, Classification, Visual clustering approach, Ailments, Twitter classifier models.

 DOI: http://dx.doi.org/10.31838/jcr.07.14.202

I. INTRODUCTION

Data analytics in sciences is more important and a challenge now a days. Whereas IT industry is moving day by day in social media applications for public sectors. For public and private sectors it is bulbous field too. With an optimizing cost mining and analyzing the data of big volume which is generated by user content for proper quality of services is a big task. We have many advantages in health care. We can get bulk data, monitored data and also with less time we can access the data. In analyzing this kind of data generally novel approach are proposed for best service. We have many methods to analyze the extracted data moreover extracting the data and tagging the related words to health related terms in social network is the key issue. Many works related in different topics are observed for framing extraction, segmenting and analyzing the data. In which some are pointed to our context.

II. LITERATURE SURVEY

Data Extraction

Data extraction is the main problem in social media. While extracting huge volume of data from public platforms that to automatic data extraction and web complex information systems are involved is a big task. The problems include extraction, managing and usage of high volume data from web. Thus data extracted will be large I volume and low in quality and high in divergence. So it is a bi critical task to make authentic system. In search of

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

this context Authors David F.Barrero, David Camacho, and Maria D. R-Moreno [1] works was identified as similar works where automatic web extractions on Genetic Algorithms are done. As their research followed Genetic algorithms and Regular expressions for learning software entities automatically by approaching Evolutionary computational approach. They have extracted some kind of structures of web data which they are named as "Wrappers". We all know that the data found from web based are not structured. Web Data Extraction (WDE) is also unsolved problem. It is also related to extraction and management and usage of high web data. They used Wrappers which are special programs which extract data from documents and store in structured format[4]. They followed Regular expressions which is a well known and effective ways to form a structured data. These expressions will be in formal language, so that can be changed to Perl for conversion to engine. A graphical representation of composition agent architecture followed by authors can be seen in Figure 1. Their experimental evaluation was carried out in various phases in that case setting of various parameters for GA and regex evolution using MAS are important. At te end grammatical rules are used for regex to get composed regex.

They get best results with mutation probabilities between 0.01 and 0.02 thus with average of 0.015 was fixed throughout the experiment. The precision and recall values are more supportive to follow the experimental analysis. The data set is very much needed to make out precision and recall. They used ten documents from different origin contacting different URLS and phone numbers[5]. Thus they calculated precision and recall in different document.

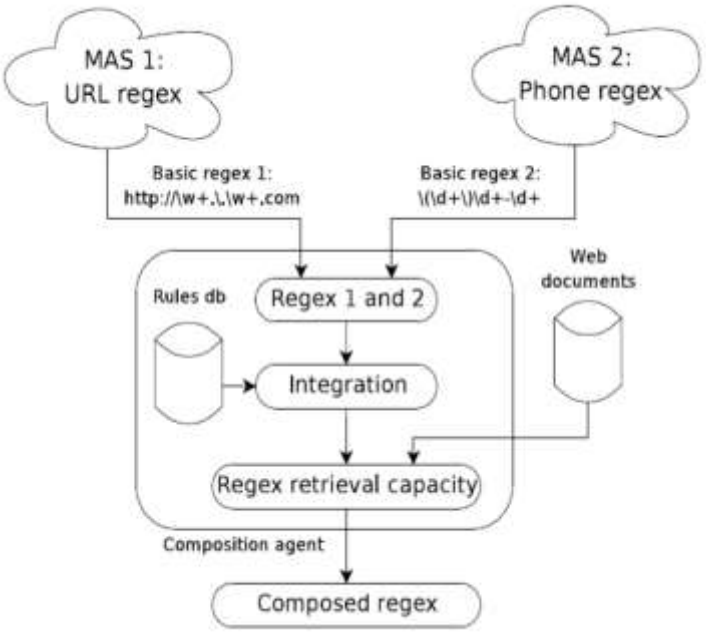


Figure 1: Composed Regex Formulation

They have verified precision and recall of both automated and manual in different parameters as string extracted are true of false for every string. Thus the results are quite impressive thus traditional and retrieved elements are ploted in Table 1.

Table 1: Extraction Capacity of Basic and Composed Regex. Table Shows Traditional and Retrived Positives Detected

	X				Y				(X)(Y)			
	Retr	Tpos	Prec	Recall	Retr	Tpos	Prec	Recall	Retr	Tpos	Prec	Recall
Document 1	5	5	1	1	0	0	-	-	5	5	1	1
Document 2	0	0	-	-	5	5	1	1	5	5		1
Document 3	5	5	1	0.5	5	5	1	0.5	10	10	1	1
Document 4	99	99	1	1	0	0	-	-	99	99	1	1
Document 5	10	10	1	1	0	0	-	-	10	10	1	1
Document 6	0	0	-	-	43	6	0.14	6	43	6	0.14	0.12

Document 7	20	20	1	0.21	773	12	0.16	0.12	97	32	0.33	0.33
Document 8	37	37	1	0.05	668	76	0.11	0.11	705	113	0.16	0.16
Document 9	24	24	1	0.13	88	1	0.01	0.01	112	25	0.22	0.14
Document 10	0	0	-	-	49	23	0.47	0.45	49	23	0.47	0.45
Average			1	0.56			0.41	0.33			0.63	0.62

Data Segmentation

Next to extraction, Segmentation of data will be done thus the data extracted will be huge that data will be sequentially transformed to data set. Those data sets are processed for classification. Works by Liangzhe Chen. Sorour E.Amiri, B.Aditya Prakash[2] on segmentation of data sequences are most likely to relate out topic. They worked on DASSA which is self guided effective algorithm which automatically detects segmentation on change in patterns. It involves a multilevel method which goes segmentation in level wise granularity.[11] They made data structures carefully leveling the information by bottleneck method with a MDL principle which represent each segment. This method will find out effective segmentation by novel average longest path optimization on segment graphs. At the end patterns are interpreted using outputs from DASSA. Hence these DASSA will goes with various real time datasets with variety of size. Hence we can find the time variance of segments when change in pattern of sequence was found. They worked on detection of Healthcare data sequence of Ebola disease. They gathered data related to Ebola infected persons of different ages and richer people and lower income young people[12][13]. They found the data by means of tweets posted by various sources. For the processing they have take a problem of finding a sequence and automatic segmentation and they divided total phase in to three properties first is the data has to process on algorithm even no forming of clusters and second find appropriate number and get cut points without user input automatically and finally complete within reasonable time for real dataset. They have taken a sequence of two segments as tabulated below in Table 2 and Figure 2.

Table 2: Sequence of Flu-cases Segmentation

	Age	Y	X	Income	Size	#Workers	#Vehicles
Segment 1	4.0	4.0	4.0	10.0	0.0	3.0	5.0
	4.0	3.0	4.0	10.0	0.0	3.0	2.0
	4.0	4.0	2.0	10.0	2.0	5.0	5.0
	4.0	3.0	4.0	10.0	2.0	3.0	2.0

	Age	Y	X	Income	Size	#Workers	#Vehicles
Segment 2	1.0	5.0	7.0	6.0	5.0	1.0	2.0
	4.0	5.0	7.0	3.0	0.0	1.0	1.0
	4.0	6.0	6.0	7.0	1.0	5.0	4.0
	2.0	5.0	5.0	7.0	0.0	3.0	2.0

Figure 2: Segmenting a Sequence of Words Appearing in Tweets

They have given how DASSA works as shown next

Algorithm 1 Pseudo- code for DASSA

Input: D
Output: The best segmentation S
1: [~X, p(~xjx)]=Cluster (D).//Finding data clusters using IB and MDL
2: Build a node for every possible time segment y.//ConstructingG
3: Add node s and t to represent the start and end time of D.
4: Create edges for adjacent y's.

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

5: Calculate the edge weights as the Euclidean distance betweenthe corresponding conditional cluster distribution p(~xjy).
6: S = DAG-ALP (G, h, s, t).//Finding the ALP as S

They also got efficient results that with different domains as epidemo,ogy, motion sequence, social media, Ebola, PUC-rio like that they have taken many datasets with min and max values and got effective results[16][17]. Here are some resultant clusters of infection status that they have got after evaluation.

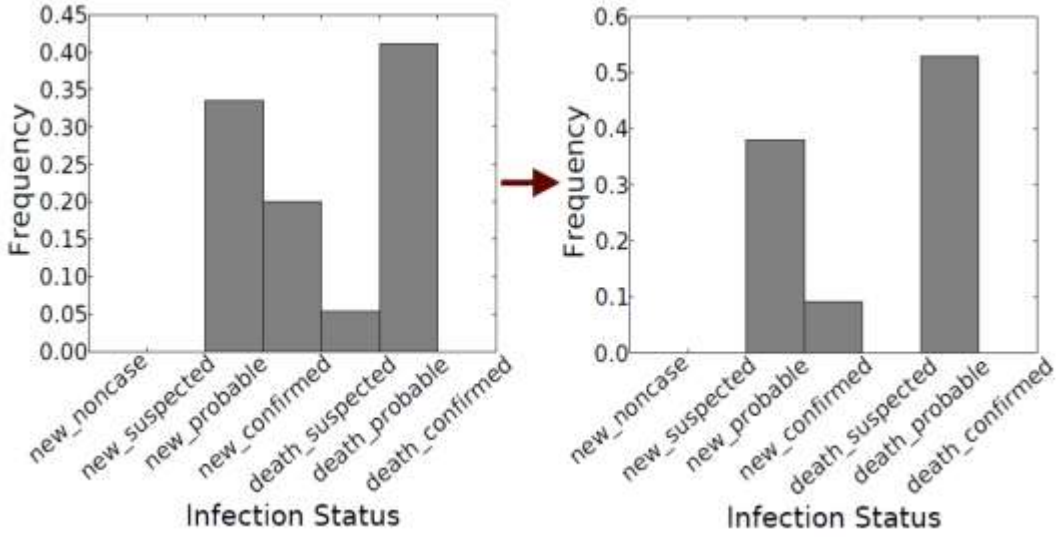


Figure 3: Infection Status

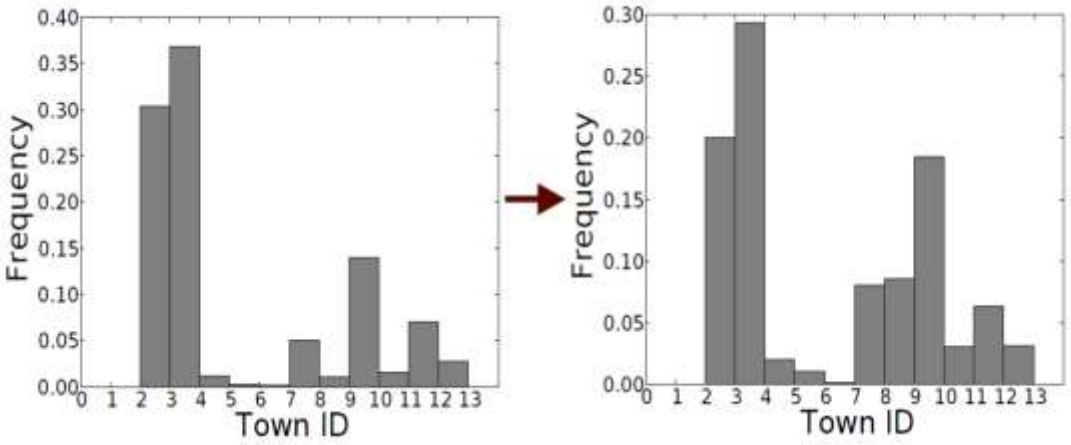


Figure 4: Infection Towns

They conclude that Fires simply does not output any clusters for many segments and it cannot detect the same good segmentations as DASSA. They believe problems would happen to other traditional clustering algorithms as well.

Classification

As we go for classification next to segmentation, works by Madan Krishna Murthy[3] on classification of large scale data sets fixes our next requirement. He worked on a large data set Contacting word sequences and classifying then by true or false statements depending upon our required query processing[6][14]. Here are some resultant clusters of infection status that they have got after evaluation. The data will be collected for processing from various sources and structures to a data set for classification that identifiers attempt to detect every sensitive information piece[8][9]. Their goal is to guarantee that even identifier remain in published data the advisory

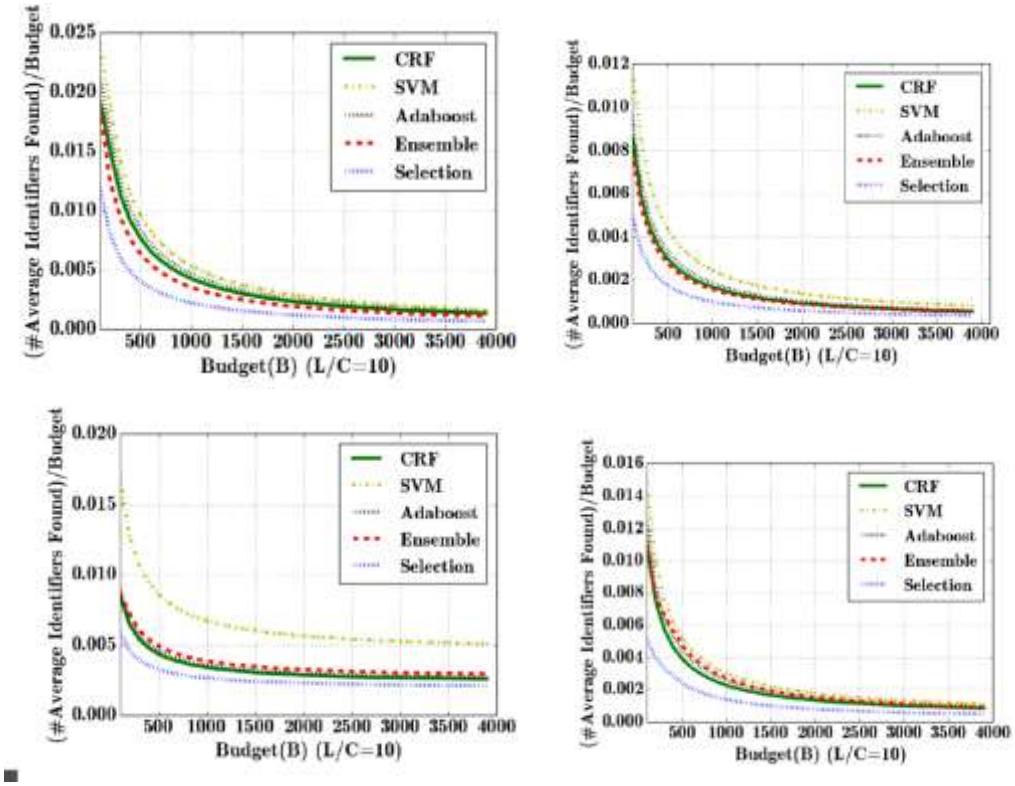
ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

cannot easily find them. He has taken positive and negative statement sin to consideration framed a Greedy sanitation algorithm[15].

Algorithm 1 GreedySanitize(X), X : training data.

$H \leftarrow \{\}, k \leftarrow 0, h_0 \leftarrow \emptyset, D_0 \leftarrow X,$
repeat
  $H \leftarrow H \cup h_k$
  $k = k + 1$
  $h_k \leftarrow$LearnClassifier(D_{k-1})
  $D_k \leftarrow$RemovePredictedPositives(D_{k-1}, h_k)
until $T(H \cup h_k) - T(H) \geq 0$
return H

The ratio of the average number of sensitive identifiers found by the attacker and the adversarial budget, while the publisher applies classifiers CRF, SVM, Adaboost, Ensemble, and Selection which allows the publisher to choose a learner with highest accuracy from fCRF, SVM, Adaboost, Ensembleg for GS (L/C=10). (a) i2b2, (b) VUMC, (c) Enron, (d) Newsgroup, and (e) Reuters datasets[7].



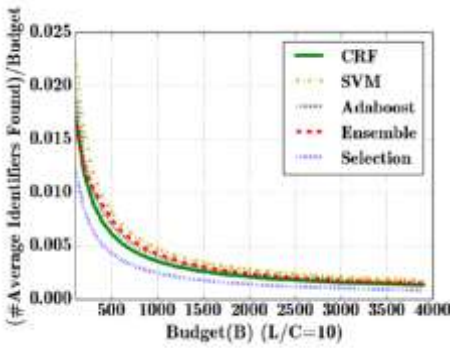


Figure 5: The Ratio of the Average Number of Sensitive Identifiers Found by the Attacker and the Adversarial Budget

III.CONCLUSION

Innovative approaches for data extraction, segmentation and classification was carried out. These experiments show limitations such as linear nature of the GA codification is not the best option to represent a hierarchical structure such as a regex as a result difficulties to define a fine-grained fitness function is able to evaluate not only all the regex, but also its parts. For these reasons the next step to follow is to use other evolutionary classifier algorithms, such as genetic programming and grammatical evolution that overcome this limitations thus we need to develop a classification model.

IV.REFERENCES

[1] David F. Barrero, David Camacho, and Maria D. R-Moreno. Automatic Web Data Extraction Based on Genetic Algorithms and Regular Expressions, *Data Mining and Multiagent Integration, Springer, LLC* 2009

[2] Liangzhe Chen, Sorour E. Amiri, B. Aditya Prakash. Automatic Segmentation of Data Sequences, *Association for the Advancement of Artificial Intelligence*, 2018.

[3] Madan Krishnamurthy, Khalid Mahmood, Pawel Marcinek. A Hybrid Statistical and Semantic Model for Identification of Mental Health and Behavioral Disorders using Social Network Analysis, *Advances in Social Networks Analysis and Mining,* 2016

[4] Mosa Salah, Basem Al Okush, Dr. Mustafa Al Rifaee, A Comparison of Web Data Extraction Techniques, *Jordan international conference on electrical engineering and information technology, IEEE,* 2019

[5] David Camacho, Maria D. R-Moreno, David F. Barrero, and Rajendra Akerkar. Semantic wrappers for semi-structured data extraction. *Computing Letters (COLE),* 4(1), 2008.

[6] Longbing Cao, Chao Luo, and Chengqi Zhang. Agent-mining interaction: An emerging area. *In AIS-ADM*, pages 60–73, 2007

[7] Chen, X. C.; Steinhaeuser, K.; Boriah, S.; Chatterjee, S.; and Kumar, V. 2013. Contextual time series change detection. *In SDM.*

[8] Sreemanth Pisupati, Mohammad Ismail.B, "Image Registration Method for Satellite Image Sensing using Feature based Techniques" *International Journal of Advanced Trends in Computer Science and Engineering* 9(1),490-593,Feb 2020

[9] Ghousia Anjum, T.Bhaskara Reddy, Mohammed Ismail.B, Alam, M., Tahernezhadi, M. "Variable Block Size Hybrid Fractal Technique for Image Compression" *Proceedings IEEE 6th International Conference on Advanced Computing & Communication Systems* March 2020

[10] K.Naga Lakshmi, Y. Kishore Reddy, M. Kireeti, T.Swathi Mohammad Ismail. B" Design and Implementation of Student Chat Bot using AIML and LSA" *International Journal of Innovative Technology and Exploring Engineering* 8 (6), 1742-1746, April 2019.

[11] Mohammad Ismail.B, V.Harsha Vardhan, V.Aditya Mounika, K.Surya Padmini "An Effective Heart Disease Prediction Method Using Artificial Neural Network "*International Journal of Innovative Technology and Exploring Engineering'* 8 (8), 1529-1532, June 2019.

[12] K.Srinivas, Mohammed Ismail.B "Testcase Prioritization With Special Emphasis On Automation Testing Using Hybrid Framework" *Journal of Theoretical and Applied Information Technology* 96(13) 4180-4190 July 2018

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

[13] Rahul Shahne, Mohammed Ismail, CSR Prabhu "Survey on Deep Learning Techniques for Prognosis and Diagnosis of Cancer from Microarray Gene Expression Data" *Journal of computational and theoretical Nanoscience* 16 (12), 5078-5088, Dec 2019.

[14] K.Rajendra Prasad, I. Surya Prabha, N. Rajasekhar, M.Rajasekhar, Social Data Analytics by Visualized Clustering Approach for Health Care, *Progress in Advanced Computing and Intelligent Engineering, Springer* 2017

[15] K.Rajendra Prasad, N. Rajasekhar, T Vishnu Vardhan Reddy, Improving of clustering results for speech data by visual approach, *SCOPES* 2016

[16] Masini Geetha Yadav, Rajasekhar Nennuri, Data Mining based Modern and Advanced Design and Development Applications, *International Journal of Pure and Applied Mathematics*, 2018

[17] Rajasekhar Nennuri, M Geetha Yadav, N Shalini, Time Series Classification in Data Mining & Clustering: A Review, 2017.