INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access
RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of First International Conference on Science, Technology & Management (ICSTM-2020)

Performance Analysis of Feature Selection Techniques for Text Classification

Hemlata Patel [1], Dr. Dhanraj Verma[2]
[1]Student,Dept. of Computer Science& Engineering, Dr. A.P.J. Abdul Kalam University, Indore,MP, India
[2]Professor,Dept. of Computer Science& Engineering, Dr. A.P.J. Abdul Kalam University, Indore,MP, India
hemlatapatel55@gmail.com[1]

Abstract

Internet is a suitable, highly available and low cost publishing medium. Therefore a significant data is hosted and published using websites. In this domain some amount of data is directly present for common people and some of data is not publically distributed. Such kinds of data are utilizable by service providers and administrators for business intelligence and other similar applications. In this presented work the web data analysis or mining is the key area of investigation and experimental study. The web data mining can be dividing in three major classes i.e. web content mining, web structure mining and web usages mining. In this work the web content mining and web usages mining is taken into consideration. First of all the web content mining is explored thus a system is developed for making comparative performance study of different content feature selection techniques. In this experiment the GINI index, Information Gain, DFS and Odd Ratio is compared using a real world collection of web pages. In order to classify the extracted features from the web contents the SVM (Support Vector Machine) is applied. The comparative study demonstrates the IG and GI is the suitable feature selection techniques that work well with the SVM classifier.

Keywords: Web Data Mining, GINI Index, Information Gain, K- Nearest Neighbour, Support Vector Machine.

1. Introduction

Web is a backbone of new generation technology, research, education, medical, engineering and a number of areas getting benefitted by web. From web documents and services using data mining techniques information is extracted automatically, this is the process of Web Mining. To discover useful data from the web using patterns is the main purpose of web mining. In this presented work the different formats of web data is explored and web mining techniques are investigated for identifying effective, efficient and accurate techniques of web data mining.

The web mining can be classified into three type's web content mining, web usages mining and structure mining [1]. A web mining process which extracts useful data from the web is called Content Mining. The contents are video, audio, text documents, structured records, and hyperlinks. In web content data delivered to the user in the form of a list, images, texts, tables, and videos. The number of webpages has increased to billions in the last few decades and it's still increasing. To search a query into billions of documents is a time-consuming task. By performing different mining techniques and by narrowing down the search data content mining extracts queried data, so it's easy to find data required by the user [3].Similarly, web usages mining explores the domain of hidden knowledge in web access log files. And finally the structure mining helps to optimize the accessibility of web pages and structure of the work [2]. While the volume of data from heterogeneous sources develops impressively, foresight and its strategies infrequently advantage from such accessible data. This work focuses on textual data and considers its utilization in foresight to address new research questions and incorporate different partners. This textual data can be gotten to and methodically analyzed through content mining which structures and aggregates data in a generally robotized way. By exploiting new data sources (for example Twitter, web mining), more entertainers and perspectives are incorporated, and more accentuation is laid on the investigation of social changes.

In this presented work the web content mining and web usages mining is the main area of investigation. Thus using a real world application the applicability of the web mining techniques is demonstrated in this work. The proposed work is a promising approach for motivating the researchers to employ different data mining techniques for solving real world issues.

2. Literature Survey

The exploration of potential of text mining for foresight by considering different data sources, text mining approaches, and foresight methods are used by authors [4]. In this paper authors extracts patterns and reduces data dimensions of BSS usage by exploring time series representation and clustering of BSS usage data [5]. This paper provides over three decades long (1983–2016) systematic literature review on clustering algorithm and its applicability and usability in the context of EDM [6]. The author's goal of review is to make available a comprehensive and semi-structured overview of WCM methods, problems and solutions proffered. They have 57 publications including journals, conferences, and workshops in the period of 1999-2018 as a review on this subject [7]. This paper provide author's try to give a brief idea regarding web mining concerned with its techniques, tools and applications [8]. Two different feature selection methods are investigated in this paper on the spam reviews detection. Bag-of-Words and words counts. Different machine learning algorithms were applied such as Support Vector Machine, Decision Tree, Naïve Bayes and Random Forest [9]. In this research an effort to address such uncertainty which is based on a data set derived from profiling data set available publicly. The conventional text feature extraction approach is applied to identify the most significant words in the data set [10]. This paper provide an improved global feature selection scheme (IGFSS) where the last step in a common feature selection scheme is modified in order to obtain a more representative feature set is proposed[11]. This paper shows an introduction of a fuzzy term weighing approach that makes the most of the HTML structure for document clustering [12].

3. Proposed Methodology

The proposed investigation of web mining is now focused to explore the domain of content mining and relevant feature selection techniques. This includes the design of data model which is used for accomplishing the desired objective. In this context a web mining model is demonstrated as given figure 1. The different component of the model is explained here.

1) Web Page Dataset: we had downloaded a significant amount of web pages from different subjects and designed a syntactic dataset. The data set is organized in a way by which the subdirectory consist of the class labels and the directory contents or web pages are treated as data instances to be classify in target subjects or domains.

2) Data Preprocessing: The entire web data preprocessing involve three main steps: a) Removal of HTML tags, b) Removal of special characters, and c)Removal of stop words.

3) Feature Selection: That technique helps to reduce the data dimension and regulate the requirements of the computational resources such as time and memory. In this work we involve four popular feature selection techniques used for web content mining.

a)GINI Index:let S is the set of samples and having k number of classes$(c_1, c_2, \dots, c_k)$. According to the classes we define k sub categorize of data such that$\{1, 2, \dots, k\}$. Then GINI index of S can be defined using [16].

$$\mathrm{Gini}(S) = 1 - \sum_{i=1}^{k} p_i^2 \qquad (1)$$

Where p_i is the probability which is calculated using i^{th} sample of S and complete set of S. however the minimum value of GINI is 0, which shows maximum utility of data. Similarly if the distribution of class and data is uniform then the GINI demonstrate the maximum value to 1 which shows minimum utility of data. In order to use the technique for text classification it is used as a measuring function of data impurity with respect to class labels associated with data. So according to previous consideration the lower value of GINI indicates the higher applicability of the attribute for classification.

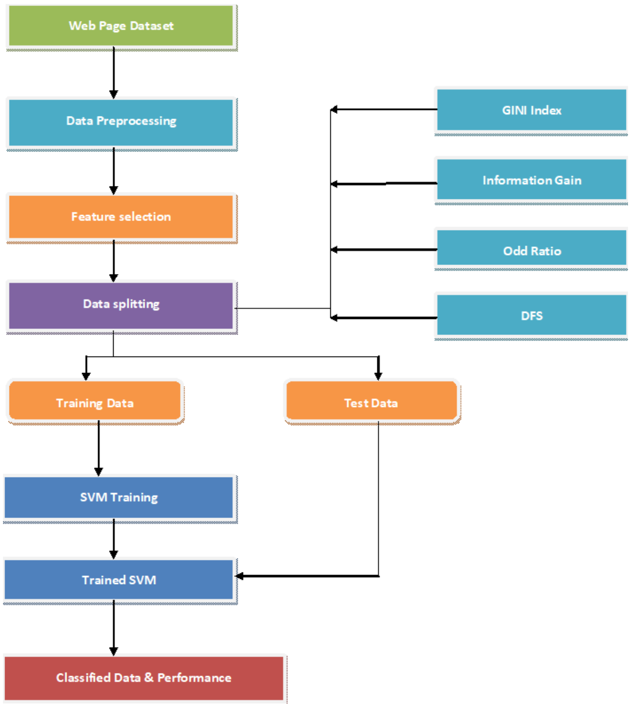


Fig.1. . Feature selection model

B) Information Gain: Information Gain (IG) measures how the features are. In text analysis, IG is used to measure the relevance of attribute A in class C. The higher the value of IG between classes C and attribute A, demonstrate the higher the relevance between classes C and attribute A [17].

$$I(C, A) = H(C) - H(C|A) \qquad (2)$$

Where, $H(C) = -\sum_{cEC} p(C) \log p(C)$, the entropy of the class, and $H(C|A)$ =1 is the conditional entropy of class given attribute, $H(C|A) = -\sum_{cEC} p(C|A) \log p(C|A)$. Since Cornell movie review dataset has balanced class, the probability of class C for both positive and negative is equal to 0.5. As a result, the entropy of classes H(C) is equal to 1. Then the information gain can be formulated as:

$$I(C, A) = 1 - H(C|A) \qquad (3)$$

The minimum value of $I(C, A)$ occurs if only if $H(C|A)$= 1 which means attribute A and classes C are not related at all. On the contrary, we tend to choose attribute A that mostly appears in one class C either positive or negative. On the other words, the best features are the set of attributes that only appear in one class. It means the maximum $I(C, A)$ is reached when P (A) is equal to $P(A|C_1)$ resulting in $P(C_1| A)$ and $H(C_1| A)$ being equal to 0.5. When$P(A) = P(A|C_1)$, then the value of $P(A) = P(A|C_2)$results in $P(C_2|A) = 0$ and$P(C_2|A) = 0$. The value of $I(C, A)$ is varied

c) DFS: The probabilistic feature ranking metric DFS. Its requirements emphasize that, terms present in a number of classes should be ranked higher than other terms; terms rarely occur in a single class and doesn't present in other classes are irrelevant and should be ranked lower; terms which frequently occur in a single class but doesn't occur in other classes are highly distinguishing, should be scored higher. DFS metric assigns score values between 0.5 and 1.0

$$DFS(t) = \sum_{j=1}^{M} \frac{P(C_j|t)}{P(\bar{t}|C_j) + P(t|\bar{C_j}) + 1} \quad (4)$$

Where, M is the number of classes, $P(C_j)$ is probability of j^{th} class and $P(\bar{t}|C_j)$ is probability of absence of term t when class C_j is given while $P(t|\bar{C_j})$ is feature likelihood when classes other than C_j are given.

d) Odd Ratio: It is a likelihood ratio. It's numerator is the multiplication of $t_p$and t_n and denominator is the multiplication of $f_p$and f_n. It presents the likelihood of feature occurrence to a class. It prioritizes those features having high occurrence rate to a particular class but ignores features which frequently occur in other classes. It also doesn't take into account irrelevant and redundant features. It's mathematical formulation is given as.

$$OR = \frac{t_p \text{ X } t_n}{f_p \text{ X } f_n} \qquad (5)$$

Odds ratio performs well on small number of features.

4) Data Splitting: After feature selection of the approach the system returns a feature vector which is used further for experimentation or learning with the supervised learning algorithm.

5) Training Set: The data splitting create two sub sets of entire web content data features first 70% of randomly selected data instances are used here for the classifier training.

6) Testing Set: Additionally the 30% of randomly selected data is used for testing of the trained model.

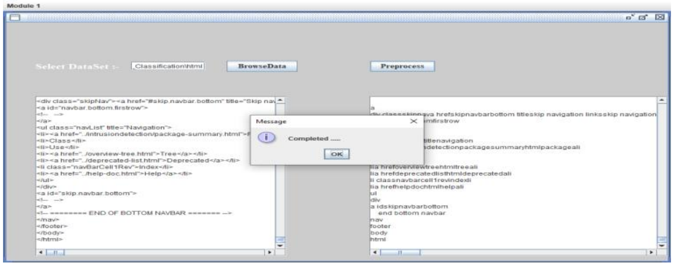
7) SVM Training: the SVM is a supervised learning model which is mostly used for classification of binary data, which is used the concept of hyper plain for differentiating between two classes.

8)Trained SVM: the SVM algorithm used for make training on the extracted features from the different feature selection techniques. After taking training from the input features the algorithm can identify the similar patterns.

9) Classified Data and Performance: based on test data classification using the trained SVM the system measures the performance in terms of accuracy and error rate. At the same time the system also computes the efficiency of the system in terms of time consumed and memory usages.

4. Implementation

Using the developed user interface we have tried to deliver the functional aspects of the proposed framework. The design and explanation are given as:



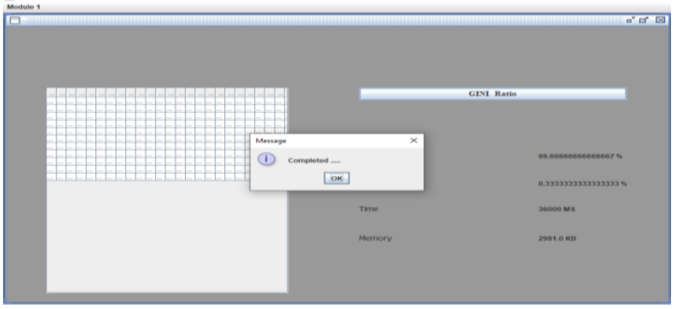
```
<div class="skipNav"><a href="#skip.navbar.bottom" title="Skip na
<a id="navbar.bottom.firstrow">
<!--   -->
</a>
<ul class="navList" title="Navigation">
<li><a href="../intrusiondetection/package-summary.html">F
<li>Class</li>
<li>Use</li>
<li><a href="../overview-tree.html">Tree</a></li>
<li><a href="../deprecated-list.html">Deprecated</a></li>
<li class="navBarCell1Rev">Index</li>
<li><a href="../help-doc.html">Help</a></li>
</ul>
</div>
<a id="skip.navbar.bottom">
<!--   -->
</a>
<!-- ======== END OF BOTTOM NAVBAR ======= -->
</nav>
</footer>
</body>
</html>
```

```
a
div classskipnava hrefskipnavbarbottom titleskip navigation linksskip navigation
mfirstrow
titlenavigation
detectionpackagesummaryhtmlpackageali
lia hrefoverviewtreehtmltreeali
lia hrefdeprecatedlisthtmldeprecatedali
li classnavbarcell1revindexli
lia hrefhelpdochtmlhelpali
ul
div
a idskipnavbarbottom
  end bottom navbar
nav
footer
body
html
```

Message

Completed .....

Fig. 2. Data selection



Message

Completed .....

99.66666666666667 %

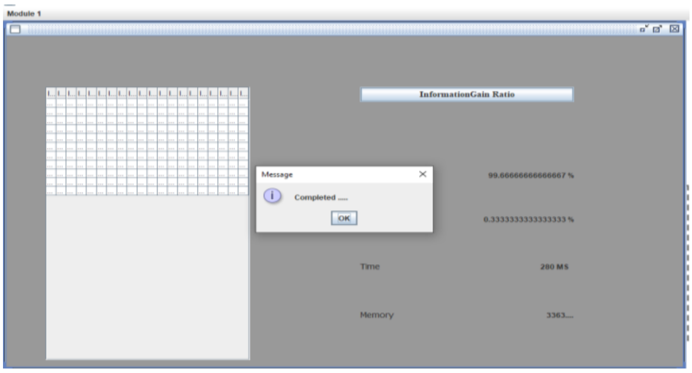
0.3333333333333333 %

Time 36000 MS

Memory 2981.0 KB

Fig. 3. GINI ratio generated

Figure 2 shows the selection of HTML data set which is available in local storage. Further in next figure 3 the feature selection technique is implemented with the concept of GINI Index.



Message

Completed .....

99.66666666666667 %

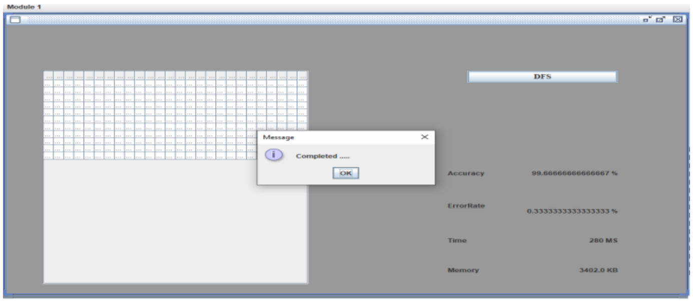
0.3333333333333333 %

Time 280 MS

Memory 3363....

Fig. 4. Information gain calculated

The figure 4 shows the implementation of Information Gain based feature selection. The figure 5shows the calculation of DFS based feature selection technique. Similarly the next figure 6 shows the odds ratio based feature selection approach.



Message

Completed .....

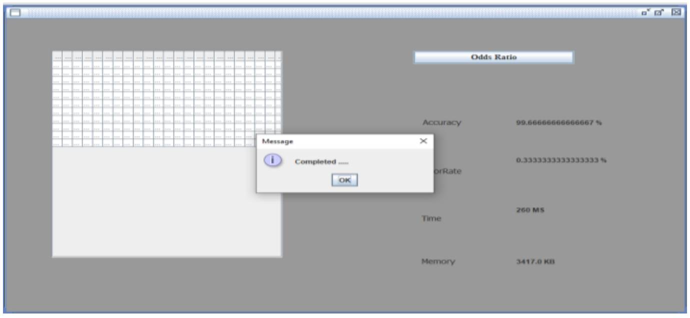
Accuracy 99.66666666666667 %

ErrorRate 0.3333333333333333 %

Time 280 MS

Memory 3402.0 KB

Fig. 5. Calculating DFS



Message

Completed .....

Accuracy 99.66666666666667 %

orRate 0.3333333333333333 %

Time 260 MS

Memory 3417.0 KB

Fig.6. Odds Ratio Calculated

5. Result Analysis

The aim of this experimental scenario is to obtain an efficient feature selection technique for implementing the web content mining based applications. In this context a comparative analysis is conducted between different feature extraction techniques. There are four parameters are used to compare the performnce.

1) Accuracy- That can be measured using the ratio of total correctly classified and the total patterns to be classified. That can also be represented using the following equation:

$$accuracy = \frac{totalcorrectlyclassified}{totalpattern\ \ stoclassify} X100 \quad (9)$$

Table.1.Accuracy (%)

No. of files	GINI Index (%)	Information Gain (%)	DFS (%)	Odds Ratio(%)
50	96.33	96.37	95.47	96.34
102	95.52	96.91	96.88	96.72
185	97.06	97.46	97.52	97.19
274	97.59	97.21	98.71	97.69
348	98.22	98.84	99.03	98.15
410	99.81	99.27	99.79	98.72
502	99.38	99.65	99.81	99.17

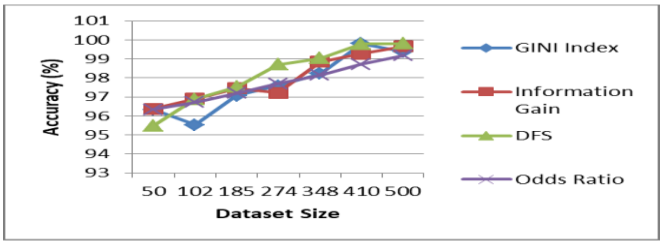


Chart.1.Accuracy (%)

The accuracy of the implemented feature extraction techniques is given in chart 1and table 1.

2) Error rate- This is a ratio of misclassified test samples and the total samples for classification. That can be calculated using the following equation:

$$error\ rate = \frac{total\ \ misclassified\ \ \ samples}{total\ \ samples\ \ to\ classify} X100 \quad (10)$$

Table 2. Error rate (%)

No. of files	GINI Index (%)	Information Gain (%)	DFS (%)	Odds Ratio(%)
50	3.67	3.63	4.53	3.66
102	4.48	3.09	3.12	3.28
185	2.94	2.54	2.48	2.81
274	2.41	2.79	1.29	2.31
348	1.78	1.16	0.97	1.85
410	0.19	0.73	0.21	1.28
502	0.62	0.35	0.19	0.83

The error rate of the different feature extraction techniques is shows in chart 2 and table 2.

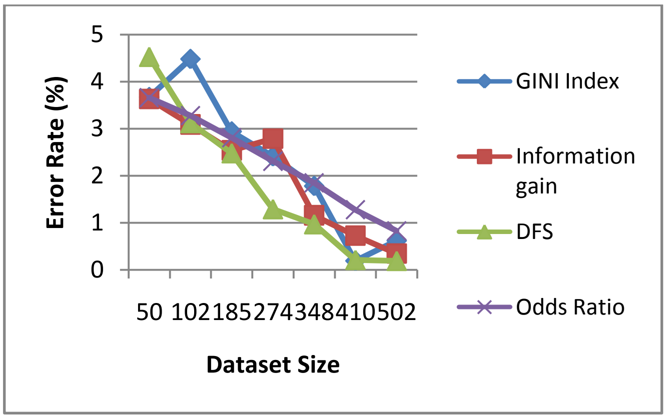


Chart 2. Error rate (%)

3) Time Consumption- The amount of time consumed for classification is calculated using the following formula:

$$time\ consumed = end\ time - start\ time \quad (11)$$

Table 3. Time consumed in MS

No. of files	GINI Index	Information Gain	DFS	Odds Ratio
50	93	96	108	101
102	178	181	195	184
185	266	279	298	280
274	429	452	485	461
348	521	548	579	557
410	658	670	701	682
502	818	851	977	912

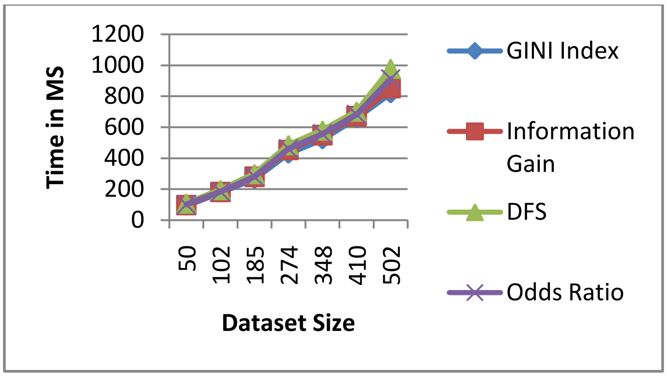


Chart 3. Time consumption (MS)

The performance of the implemented feature selection algorithms in terms of time consumption is given using figure 3 and table 3.

4) Memory Usage -The amount of total memory utilized for execution of an algorithm is measured here as the memory consumption or usages.

$$memory\ usage = total\ memory - free\ memory \quad (12)$$

Table.4 Memory Usages in KB

No. of files	GINI Index	Information Gain	DFS	Odds Ratio
50	3016	3828	3490	3495
102	3445	3713	3378	3359
185	3358	3925	3564	3716
274	3158	3840	3672	3573
348	3259	3996	3749	3658
410	3480	4019	3557	3458
502	3369	3957	3689	3691

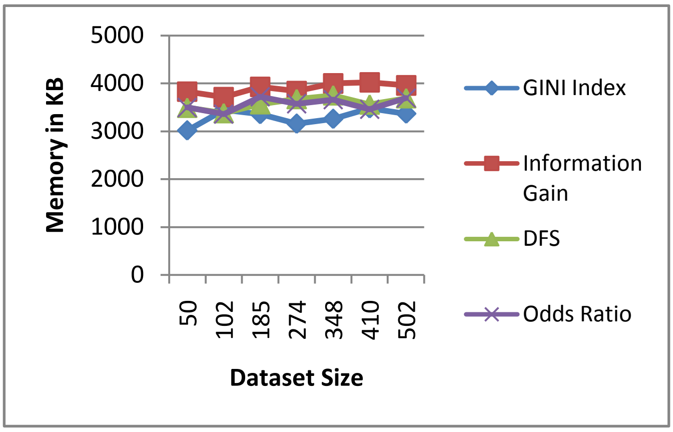


Chart 4. Memory Usage (KB)

The memory usage of the implemented algorithms for feature selection is explained in chart 4 and table 4.

Comparative Analysis of Feature Extraction Techniques

In this experimental evaluation the different feature extraction techniques namely GINI Index, Information Gain, DFS and Odds ratio is used with the SVM classifier. In this experiment different parameters that are accuracy, error rate, memory usage, time consumption is measured for performance evaluation. Below in table 5 gives the summary of the implemented techniques in terms of mean performances.

Table.5.Comparisons between feature extraction techniques

Parameter	GINI + SVM	IG + SVM	DFS + SVM	OR + SVM
Accuracy	97.57	97.51	98.17	97.71
Error Rate	2.43	2.05	1.83	2.29
Memory Usage	3297.85	3896.57	3585.57	3564.28
Time	423.28	439.57	477.57	453.85

According to the given table 5, we can see that DFS leads in accuracy and error rate, by producing less error rate and higher accuracy in classification. On the other hand Information Gain and GINI Index based technique acceptable for time requirements of the classification system. in this context the GINI Index and Information Gain found acceptable for accuracy and less time requirements. Thus in further both the feature selection techniques are investigated with different supervised learning classifiers.

6. Conclusions and Future Work

In this presented work the web data mining is the main area of investigation. Therefore the web usages mining and web content mining is studied and the relevant methods are demonstrated. In web content mining the web pages are involved for experimentation because the most of web contents are published using HTML pages. These web pages includes different formatting tags and text contents therefore it complex to process and classify using any basic machine learning method. Therefore first some feature extraction techniques are explored namely GINI Index, Information gain, DFS and odds ratio. All these methods are basically measuring the ranks of the text features for selecting most appropriate according to their defined class labels. The experimental study offers different techniques and methods that useful for different kinds of data mining approaches used in web data mining.

Future work of this experiment is further extended to find suitable and efficient classifier for web content classification. Therefore the selected two feature selection techniques namely GINI Index and Information Gain will utilize with the three popular supervised learning classifiers namely SVM, SVR and k-NN.

References

Journals

[1] Y. Li, S. Arora, et al(2018) "Using web mining to explore Triple Helix influences on growth in small and mid-size firms", Technovation 76-77, 3–14

[2] S. K. Pal, et al(2002), "Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions", IEEE Transactions on neural networks, Vol. 13, No. 5,

[3] M. J. H. Mughal,(2018) "Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview", International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6.

[4] V. Kayser, and K. Blind, (2016) "Extending the knowledge base of foresight: The contribution of text mining", Technological Forecasting & Social Change xxx

[5] D. Li, Y. Zhao and Y. Li,(2019) "Time-Series Representation and Clustering Approaches for Sharing Bike Usage Mining", Volume 7.

[6] A. Dutt, et al(2017) "A Systematic Review on Educational Data Mining", Volume 5, 2017, 2169-3536,

[7] M. O. Samuel, et al(2019) "A Systematic Review of Current Trends in Web Content Mining", IOP Conf. Series: Journal of Physics: Conf. Series 1299 (2019) 012040, IOP Publishing, doi:10.1088/1742-6596/1299/1/012040

[8] A. Kumar, and R. K. Singh (2016), "Web Mining Overview, Techniques, Tools and Applications: A Survey", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 03 Issue: 12 |

[9] W. Etaiwi, and A. Awajan,(2017), "The Effects of Features Selection Methods on Spam Review Detection Performance", International Conference on New Trends in Computing Sciences, 978-1-5386-0527-1/17 $31.00

[10]Z. Zuo, et al(2018) "Grooming Detection using Fuzzy-Rough Feature Selection and Text Classification", 978-1-5090-6020-7/18/$31.00

[11]A. K. Uys(2016) al, "An improved global feature selection scheme for text classification", Expert Systems With Applications 43 82–92

[12]A. P. G. Plaza, et al(2019) "Using Fuzzy Logic to Leverage HTML Markup for Web Page Representation", IEEE Transactions On Fuzzy Systems,

[13]H. Park and H. C. kwon, (2011) "improved GINI index algorithm to correct feature selection Bias in Text Classification", IEICE Trans. INF.& SYST, VOL E94D, No 4.

[14]A. I. Pratiwi, and Adiwijaya,(2018) "On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis", Hindawi Applied Computational Intelligence and Soft Computing Volume, Article ID 1407817, 5 pages

[15]M. N. Asim, et al "Comparison of Feature Selection Methods in Text Classification on Highly Skewed Datasets", 978-1-5386-2969-7/17/$31.00 © IEEE.