International Journal of Systems Signal Control and Engineering Application

MEDWELL PUBLICATIONS

Web Page Block Identification using Machine Learning Techniques

[1]Neetu Narwal, [1]Sanjay Kumar Sharma and [2]Amit Prakash Singh
[1]*Institute of Maharaja Surajmal, Banasthali Vidyapith, Rajasthan, India*
[2]*GGSIP University, New Delhi, India*

Key words: DOM, page block, radial basis network, support vector machine, neural network

Corresponding Author:
Neetu Narwal
Institute of Maharaja Surajmal, Banasthali Vidyapith, Rajasthan, India

Page No.: 67-73
Volume: 12, Issue 4, 2019
ISSN: 1997-5422
International Journal of Systems Signal Control and Engineering Application

Abstract: Internet has gained greatest acceptance as reservoirs of information. It has been observed that the web page along with main content comprises of noise (advertisement, external links). This noise content poses difficulty for various search engines to classify the web page accurately and provides distraction to the serious user interested in gathering data related to a topic. There are various segmentation techniques that partition the web page but very few have categorized the segmented block. In this study, we tried to categorize the page blocks extracted from segmentation. We have used web page segmentation algorithm for parsing the web page and extracted important features to build a dataset. Linear and nonlinear machine learning techniques to have been used to train dataset. In this experiment we also analyzed the importance of features for the learning process. We perceived that the embedded objects from external source have highest significance for block identification. In our experiment, the non-linear radial basis neural network resulted in best performance with an accuracy of 99.89%.

INTRODUCTION

Internet is the most common media nowadays to gather information regarding any topic. The user finds it convenient to pick information through web sites. But the web page along with the main content generally includes advertisements, copyright information, external hyperlinks and navigation panel for internal hyperlinks.

Web designer today provides customized advertisement on the web page which is based on the user's web usage history and hence, it distracts the user from his goal. In case of student, he has a tendency to view not only relevant but irrelevant content of the web page. Hence, to distinguish disparate information in a web page, we tried to build a model that categorize the web page content under three categories namely pure noise, mix of noise and main content and pure main content. The methodology used in the research comprises of the following steps, the first step is to parse the web page and segment them using visual and spatial information into a set of visual blocks. Then, extract important features from the block. Some of the features are directly gathered from the source code and others are computed using relative feature value in reference to the whole web page. We congregate features related to spatial information, content information, link information, formatting information and external source information. These features are further analyzed for their relevance and significance in block identification. Finally, the dataset is trained with various machine learning classification algorithm, i.e., support

vector machine, neural network and logistic regression and their results are compared in terms of accuracy, precision, recall and f-measure to learn the best block identification model. The major work done in this research includes:

- A block identification model that can categorize the block as pure noise, mix of noise and main content, pure main content
- Building a dataset of web sites belonging to five different categories, i.e., science, academics, fiction, sports, news from yahoo directory utility
- Using correlation coefficient measure, we gathered interesting insights into significance of features in building the model
- Some applications of block identification

Literature review: Many researchers have worked in the area of block identification and used different strategies for web page segmentation and content categorization, i.e., DOM styles, templates, entropy measure of terms, etc.

Song *et al*. (2004) have derived a block importance estimation model that automatically assign importance values to block in a web page and formulated it as a learning problem. Our work is closely related to the work by Song *et al*. (2004). We included some interesting features related to the block to build the corpus that is different from theirs. The experiment show that these new features have highest correlation measure and information gain values and hence, contribute in a major way to identify the block. Our block identification model gives better performance in all the learning methods as compared to them.

Bar-Yossef and Rajagopalan (2002) compared web pages of a given website to identify a template. Based on template they formulated rules to partition the web page. The count frequency of partition repeated across web pages of a website was used to differentiate content type. Our model is not web site specific and provide automatic identification of the block.

Lin and Hu (2002) designed a system named Info discover, that partitions a web page into several content blocks on the basis of table tags which is considered as the common tag to partition any web page. Entropy of each works is merged to form block entropy which is used to differentiate informative or redundant block.

Yi and Liu (2003a, b) derived a new style tree that represents both layout and content of a web page. They mapped the web pages of the site style tree to detect noisy information in the page. Their experiment shows that the noise elimination technique improves the classification and clustering of web page considerably. Their research needs overhead to construct style tree and comparison with web page of the given website to identify block type. We reduced the overhead of style tree construction and provided a generalized model that automatically identifies the block based on the feature.

Gupta *et al*. (2003) have proposed a DOM-based content extraction method to allow information to be accessed over small-scale devices such as PDAs, mobile phones, etc.

Block identification model: Web page author organize their content in a rational manner such that the noise and main contents are arranged in separate blocks to discriminate the content relevance. Each block has certain significance in terms of its positioning and arrangement. Therefore, we proposed to construct a model that automatically identifies the block based on their features.

A block identification model uses machine learning technique to map the features (f_1, f_2, ..., f_n) of the block B_i with its block type T_j.

- B_i $[f_1,f_2 .. f_n] \rightarrow T_j$
- Block feature → Block identification

The first step of building block identification model is to perform web page segmentation and extract visual blocks. Web page is segmented using various techniques few of them are:

- DOM based segmentation
- Location based segmentation
- Vision based segmentation

In our research, we used DOM (Document Object Model) based approach. A web document is represented as a tree structure, arranged in a hierarchy. DOM provides us with a set of API functions that help to retrieve, manipulate and delete elements in the web page.

The web page displayed in Fig. 1 is the news website and by observing it we derive certain information regarding content arrangement and styling used by the web author.

- The web author has placed important content in the center
- Web site navigation panel is found below the header block and copyright information is present in the bottom
- The right panel is reserved for advertisements
- The color scheme used by the main content is minimal and has sparse use of formatting features
- In the given example the main content occupies approximately 50% of the web page covered area

From the observations, we conclude that the significance of a block is reflected by its position in context to the whole web page. The hyperlink available in

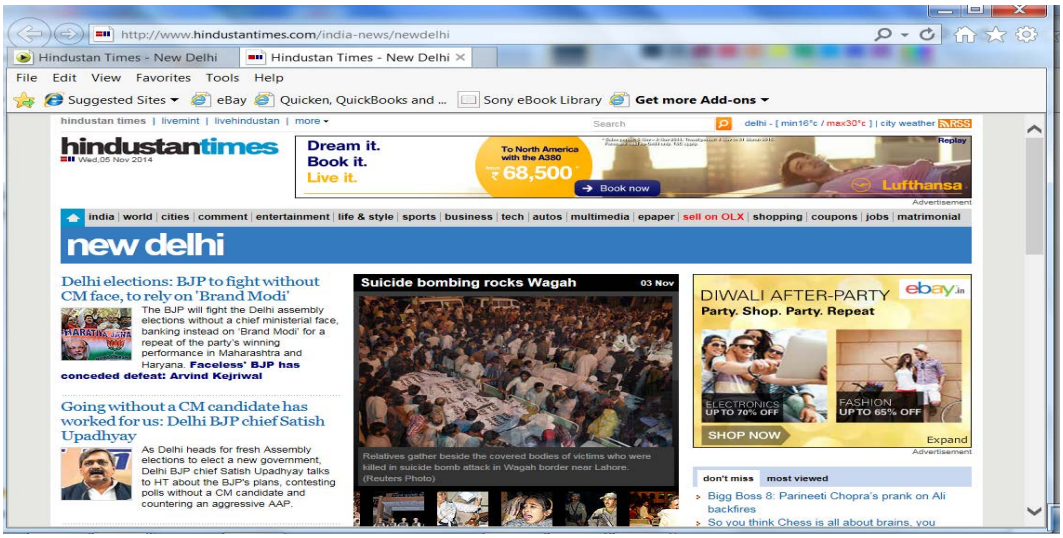

Fig. 1: Web page of Hindustan times web site showing presence of advertisements, external links, internal links with main content uploaded on 5th Nov 2104

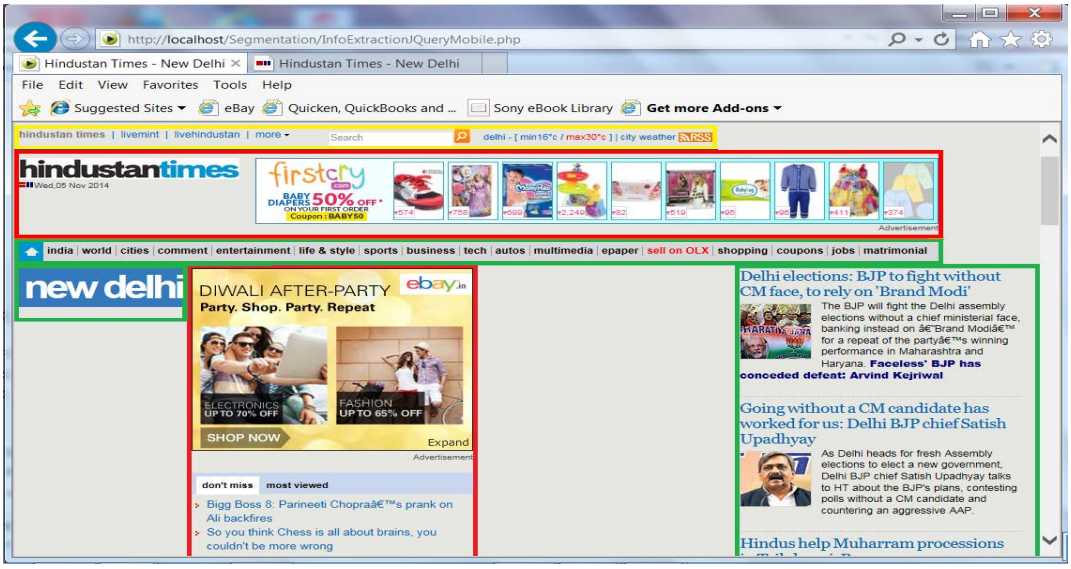

Fig. 2: Display visual block output after applying web page segmentation algorithm on the web page shown in Fig. 1

the block is also considered as an important feature in deciding the significance of the block. There are two types of hyperlink, i.e., link to the web page of the same domain or to other domain. Navigation panel or the main content comprises of internal links whereas the advertisement comprises of external links. Figure 2 displays visual blocks extracted from web page displayed in Fig. 1. The visual blocks are demarcated with colored boxes where red color symbolizes noise block, yellow color for mix of noise and main content and green color for main content.

Our web page segmentation algorithm use top down approach to parse the web page hierarchy structure and the nodes obtained are recursively parsed until it reaches a threshold size. The nodes having covered area below the minimum threshold size are merged with their sibling nodes to obtain the visual block.

Dataset preparation: We choose several websites belonging to five different categories, i.e., news, science, academics, health and sports to build a robust dataset. The visual block extracted from each web page is scanned for

the features. These features are segregated in different dimensions and analyzed for their significance in building model classifier.

- Spatial features
- Formatting features
- Content features
- Hyperlink features
- Embedded features

Spatial features: Spatial features are related to the positioning information of the page block with reference to the web page. The spatial features used in our model are:

Block_Top, Block_Left, Block_Width, Block_Height

These features are extracted using DOM API functions on the visual block. Since, the size of web page may span two or more screen size, hence, block_height feature needs to be normalized using its relative value that is measured in percentage with reference to the web page. Two blocks of similar height in two different web sites have distinct meaning where in one the web page spans single window and in other web page spans multiple windows. We also compute two relative features from the above features:

Rel_Block_Top, Rel_Block_Height

Generally the width of the web page does not span two or more web page. Therefore, we can avoid normalizing the width and left of the page block.

Formatting features: Formatting features represents the formatting styles applied on the visual block. We have used four formatting features in our work.

Block_Font_Size Block_Font_Weight
Block_Backgroundcolor Block_Color_Pattern

Block_font_weight is a value that specifies the relative density of text specified along with the font. Since, we have multiple font-size and font-weights in a given block, therefore, we used maximum font-size and weight of the text contained in the block.

Single block may incorporate multiple colors in the text or it may have single color. Block_color_pattern feature represents the count of different colors used inside the block.

Content and hyperlink features: Content features are related to the information in terms of text, images, tables contained inside the block. Hyperlink features are related to the count and type of link present inside the block. We have used five content and hyperlink features in our work:

Block_Text_Length Block_No_of_Img
Block_No_of_Tables Block_No_of_Intl_Link
Block_No_of_Extn_Link

Block_text_length is the length of the text contained inside the block. Block_no_of_img and Block_no_of_ tables are the count of images and tables placed inside the block. Block_no_of_intl_link measures the count of links to the same web page or within the same domain. Block_no_of_extn_link measures the count of external links present in the page block, external links refer to the links to different domain, it is analyzed by checking the href attribute of anchor tag <a href=..", the presence of http:// along with non different domain name signifies the external link. An advertisement block usually contain link to external website. The navigational panel may also contain links but these links are generally to the same domain. We have also computed five relative features values:

Rel_Block_Text_Length Rel_Block_No_of_Img
Rel_Block_No_of_Tables Rel_Block_No_of_Intl_Link
Rel_Block_No_of_Extn_Link

Embedded features: Web page has certain external objects which are embedded in the web content while displaying it on the web browser. These embedded objects may belong to the same domain or it may belong to other domain. Hence, it is necessary to identify the source of these objects. Mostly objects with external source are advertisements or external links to other web domain. Embedded features used in our work includes:

Block_no_of_iframe, Block_no_of_javascript,
Block_no_of_extn_src.

We also considered three relative features:

Rel_Block_no_of_iframe, Rel_Block_no_of_javascript,
Rel_Block_no_of_extn_src.

Learning algorithm: To derive the significance of each block, we have used the approach of learning by example, where the test datasets are manually pre labeled with class and trained to build a model. Each block is represented as (x, y) where x is set of features of the block and y is the class. In this study, we have used linear and non-linear Support Vector Machine (SVM), Artificial Neural Network (ANN) techniques to train the model.

Support vector machine: The Support Vector Machine (SVM) classifier is a tool for classification developed by Vapnik (1992), it aims at finding the best hyper-plane that separates the feature space of a class from the other.

The data for training is a set of feature (represented as vector) x_i along with their class y_i. For some feature d, the $x_d \in R_d$ and $y_i = \pm 1$. The equation of hyper-plane is:

$$<w, x>+b = 0 \quad (1)$$

where, $w \in R_d$, <w,x> is the inner product of w and x, b is a real number. To find the best hyper plane, we find w and b that minimize ||w|| such that for all data points (x_i, y_i):

$$y_i(<w_i x_i>+b) \geq 1 \quad (2)$$

The supports vectors are the x_i on the boundary, those for that $y_i(<w_i x_i>+b) = 1$.

SVM is generally useful for two-class problem where we find the hyper plane that separates the data points with maximum margin. In our research, we have three classes hence we used the concept of single class against all and build a multiclass SVM classifier model.

In our experiment we have also used non-linear approach of SVM. It is generally useful in problem where data are not linearly separable and hence, each is mapped using some kernel function to higher dimensional feature space formed by the nonlinear mapping of n-dimensional feature set into k-dimensional feature space (k>n) through use of function φ(x).

The kernel function K (x, y) is defined on a linear space S and a function φ that map x to S such that:

$$K(x,y) = <\varphi(x), \varphi(y)> \quad (3)$$

The dot product takes place in the space S. The class of kernel functions which are popularly used are polynomial, radial basis function and certain sigmoid function. In our experiment we have tested all three-kernel function but the radial basis function shows good result as compared with other.

Radial basis function (Gaussian): Kernel function K maps the attribute x using the following function to provide larger feature set. For some positive number σ:

$$K(x, y) = \exp(-<(x-y),(x-y)>/(2\sigma 2)) \quad (4)$$

Neural network: Artificial Neural Networks (ANN) is best known for solving problems that can't be solved using conventional algorithms. When new input is provided to the ANN Model, it produces an output similar to the closest matching training input pattern.

In neural network model architecture, each node at input layers receives input values, do processing on it and send it to the next layer. The key feature of neural networks is that it learns the input/output relationship through training. The response of the neural network is reviewed and the configuration is refined until the analysis of the training data reaches a satisfactory level.

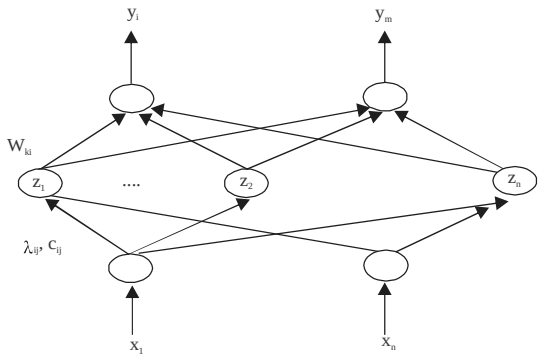

Fig. 3: RBF neural network structure

In our experiment we used feed forward neural network to build a model. The input layer has 24 neurons and output layer has 3 neurons. We have used sigmoid activation function to train the model.

In our research, we have also used non-linear radial basis neural network. It is feed forward neural network with single hidden layer that uses radial basis activation function for hidden neurons are called Radial Basis Function (RBF) network.

The RBF neural network has an input layer, a radial basis hidden layer and an output layer as shown in Fig. 3. The parameter c_{ij}, l_{ij} are center and standard deviations of radial basis activation functions. We have used gaussian activation function in our experiment:

$$\sigma(\gamma) = \exp(-\gamma^2) \quad (5)$$

MATERIALS AND METHODS

Experiment: In this study the measured accuracy of the block identification model and the feature importance and their influence on the learning model is stated.

The dataset comprises of different category of websites picked from yahoo directory utility for science, academics, fiction, sports, news. We have extracted web block from 120 web pages from 80 different web sites, giving us total 1500 blocks used in our study. These blocks are then manually labeled as pure noise block, mix of noise and main content and pure content block.

We conducted experiment using six different classifiers linear svm, non-linear svm and neural network, rbf neural network, logistic regression, multi class classifier. We evaluated the results using different measures such as accuracy, specificity, precision, recall, f-measure.

Table 1: Features with their correlation coefficient and information gain measure

Feature name	Correlation coefficient	Info. gain	Feature name	Correlation coefficient	Info. gain
Block_No_of_Iframe	0.4524	0.1962	Rel_Block_Width	0.1082	0.025
Rel_Block_Extn_src	0.3918	0.1962	Block_No_of_Intl_link	0.1044	0.1029
Block_No_of_java	.02862	0.1830	Rel_Block_link_length	0.1043	0.1193
Block_Extn_src	0.2049	0.1250	Block_No_of_Tbl	0.1043	0
Rel_Block_No_of_Out_link	0.1805	0.0791	Rel_Block_height	0.0972	0.024
Block_No_of_Img	0.1686	0.0400	Block_Link_Length	0.0941	0.1791
Block_Font_Size	0.1597	0.1150	Rel_Block__No_of_Tbl	0.0895	0.0396
Rel_Block_No_of_Img	0.1575	0.0446	Rel_Block_Text_length	0.0837	0.1101
Block_Top	0.1548	0.1170	Block_Width	0.0582	0.0763
Block_No_of_out_link	0.1303	0.1072	Block_Text_len	0.0575	0.2634
Block_Height	0.1127	0.7290	Block_Left	0.0215	0.398
Block_Font_Weight	0.1106	0.0691	Rel_Block_No_of_Intl_link	0.0190	0.1175

Table 2: Comparison of general features and relative features

Feature set	Accuracy	Specificity	Precision	Recall	F-measure
Feature set	0.9383	0.8912	0.7987	0.8199	0.8039
Relative feature set	0.8041	0.6934	0.4447	0.5486	0.4892
Aggregate feature set	0.9766	0.9605	0.9170	0.9379	0.9273

Table 3: Comparison of block identification of various classifier models

Feature set	Accuracy	Specificity	Precision	Recall	F-measure
Logistic regression	0.8920	0.8790	0.8180	0.8340	0.8220
Multi class classifier	0.8560	0.8540	0.8010	0.8210	0.8040
Linear SVM	0.8158	0.6928	0.4395	0.5250	0.4784
Non linear RBF SVM	0.9795	0.9659	0.9254	0.9494	0.9368
Linear feed forward neural network	0.9766	0.9605	0.9194	0.9368	0.9368
Non linear radial basis network	0.9989	0.9806	0.9546	0.9722	0.9754

Features significance in block identification: There are total 24 features that include 15 features extracted explicitly from web page blocks, 9 features that are calculated in context to the complete page in percentage. Here we analyzed the relative importance of all features using information gain and correlation coefficient for classification. The result is shown in Table 1.

The discriminative correlation coefficient value in Table 1 show that count of iframe objects, embedded objects from external source and count of javascript code in the block have the highest correlation value which signifies that embedded objects have higher significance in block identification. The top and height feature of the block also have significant measure that contribute for block identification.

We conducted experiment for analyzing the relative importance of features and their role in building the finest classifier model. We used the correlation coefficient measure to order the feature set and then made five dataset having 5, 10, 15, 20, 24 features, respectively. On these dataset feed forward neural network classifier is tested.

The accuracy of the model improves significantly with the increase in the feature set size. The results also depicts that fifteen features dataset provides higher accuracy and precision as compared to twenty features. The results show that accuracy and f-measure is almost same for fifteen and twenty-four features dataset. In our research, we have used twenty-four features for training the model (Fig. 4).

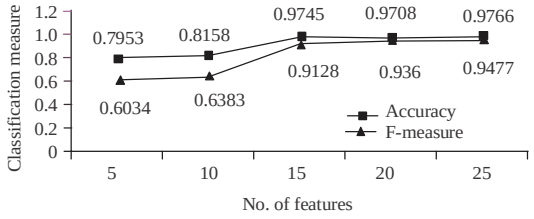

Fig. 4: Classification performance measure varies with number of features

To analyze the influence of features and relative features on the block identification model, we experimented it using feed forward neural network to train the classifier by using two category of features set and the result is depicted in Table 2. The result shows that accuracy of the model improves significantly when relative feature set is combined with the block feature set.

Comparison of learning method: We implemented six classifiers for training our model, logistic regression, multi class classifier, linear SVM, non-linear SVM, linear feed forward neural network and radial basis neural network. The results obtained by these models are shown in Table 3.

The result depicts that non-linear classifier outperforms linear classifier in terms of accuracy, recall, precision. The radial basis network shows the best results with accuracy of 99.89% and recall of 97.22%.

RESULTS AND DISCUSSION

Application of block identification: The block identification plays a significant role in various web applications. The output of the model can be utilized for web content personalization, content segregation, search engine crawlers, viewing the web page on small screen device, etc.

Most of the search engine crawlers typically collect keywords from the complete web site to build up their database. The web page usually contains some noise content along with important relevant content. Noise content can mislead the correct classification of the web content. Web page classification results may be significantly improved if only main content is picked for analysis.

Block identification can be utilized for topic specific search where user is interested in finding the useful content related to any topic from different web site. The main content from different web sites can be clubbed and displayed to the user.

Another useful application of block identification is displaying selective content of web site on small screen devices. Due to limited screen space, main content and internal links information is sufficient to be displayed to the user.

CONCLUSION

To distinguish disparate information in the web page, we tried to build a model that use machine learning techniques like support vector machine, neural network and logistic regression to correctly categorize the content of a web page. The results obtained show that nonlinear radial basis neural network outperforms other techniques with accuracy of 99.89%. Among all the features the embedded objects feature shows highest significance for block identification. Spatial features, i.e., top and height also have high correlation coefficient value.

REFERENCES

Bar-Yossef, Z. and S. Rajagopalan, 2002. Template detection via data mining and its applications. Proceedings of the 11th International Conference on World Wide, May 7-11, 2002, Honolulu, Hawaii, USA., pp: 580-591.

Gupta, S., G. Kaiser, D. Neistadt and P. Grimm, 2003. DOM based content extraction of HTML documents. Proceedings of the 12th international conference on World Wide Web, May 20-24, 2003, NewYork, USA., pp: 207-214.

Lin, S.H. and J.M. Ho, 2002. Discovering informative content blocks from Web documents Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, ACM, Edmonton, Alberta, pp: 588-593.

Song, R., H. Liu, J.R. Wen and W.Y. Ma, 2004. Learning important models for web page blocks based on layout and content analysis. SIGKDD Explor. Newsl. 6: 14-23.

Vapnik, V., 1992. Principles of Risk Minimization for Learning Theory. In: Advances in Neural Information Processing Systems, Moody, J.E., S.J. Hanson and R.P. Lippmann (Eds.). Vol. 4. Morgan Kaufmann Publishers, Inc., San Fransisco, USA.

Yi, L. and B. Liu, 2003a. Web page cleaning for web mining through feature weighting. Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI-03), August 9-15, 2003, Acapulco, Mexico, pp: 43-48.

Yi, L., B. Liu and X. Li, 2003b. Eliminating noisy information in web pages for data mining. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, Washington, DC. New York, pp: 296-305.