International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 06 | June -2018 www.irjet.net p-ISSN: 2395-0072

Multi-Faceted Approach to Automated Classification of Business Web Pages

Venkatesh R[1]

[1]RRD, Chennai, India

Abstract - *Data Services Companies routinely aggregate commercial business content automatically from entity websites on a large scale. A key product is the base information on new and upcoming businesses which have a potential value for business information databases and CRM products. A legal name, contact information, business activity and name of principal/leadership team are valuable basic information that can be aggregated. However there is cost involved in crawling all the pages to identify the right page where the four key information points are located. A business website has an average of 5 pages and pinging and crawling each page involves IP cost at granular level and virtual machines capacity at a high level. The efficiency of the current process and algorithms in use is not satisfactory and results in increased cost of operation. However, because the current algorithms are "Black Box", companies find it hard to tweak the algorithm to better suit its needs and achieve higher accuracy figures. The research is aimed at building an effective model to classify pages and reduce the crawl volume and control running costs. The objective is to predict the most likely pages to pass on for extraction and the most unlikely pages for bypassing.*

Key Words: Information Retrieval; Information Science; Data Mining; Supervised Learning; Webpage Classification

1. INTRODUCTION

Data Services companies routinely aggregate commercial business content automatically from entity websites on a large scale. A key product is the base information on new and upcoming businesses which have a potential value for business information databases and CRM products. Address and names of Executives are valuable basic information that can be aggregated along with a multitude of other requirement based data. Companies have proprietary method to discover websites of new websites and aggregate base information. However, there is cost involved in crawling all the pages to identify the right page where the four key information points are located. A business website has an average of 5 pages and pinging and crawling each page involves IP cost at granular level and virtual machines capacity at a high level.

Amongst the multitude of steps orchestrated by these software, a prominent activity is to classify if a page has relevant content or not. The platform channels all relevant pages to a downstream extraction process. Companies face a challenge with significant False Negatives in classification impacting the revenues as every business data point missed due to incorrect classification is potential revenue loss. Companies are looking for alternative statistical model(s) built and deployed on top of its text mining and data collection engines that could deliver better and more accurate results. The objective of the research is to improve the current web page classification engine used at data services companies by building an effective predictive model to classify the pages and reduce the crawl volume and control running costs. The model will predict if a web page contains executive information and Business Address or not, thereby improving revenue potential, reduce Operational costs via improved coverage.

1.1 Limitations

The devising of the strategies to build the model majorly using information in URL text was limited by its current capacity for data collection. Hence the data may not be exhaustive of all possible key words that might appear in URL texts of webpages containing the target information

Since the project uses supervised learning algorithms, the target variables required manual flagging which:

- Limited the size of the training set available
- The logic and methodology of data collection is outside the scope of the project and hence has not been discussed in detail.

1.2 Challenges

There is no prior classification of the pages. The researcher had to examine each link and manually classify the training set thereby increasing data preparation efforts or limiting the size of the training set.

The page links provided were generated six months ago. Some of these links are likely to be obsolete or inactive.

Data for page characteristics has to be generated using automated means to manage time. Automation might not yield 100% data due to safeguards against scraping or inherent flaws of the page, limiting the size of the training set.

1.3 Machine Learning

The following list of Machine Learning algorithms were used for the final model: due to the fact that dataset contains

a high class imbalance, with respect to both the target variables.

- Lasso
- Logit
- Random Forest
- Naïve Bayes
- C5.0 – Rule Based
-

1.4 Model Evaluation & Validation

In machine learning, the classifier is evaluated by a confusion matrix. Considering the case of imbalanced data assessment, accuracy places more weight to the common classes than on the rare classes, which in turn makes it difficult for the classifier to perform well on the rare classes. It becomes a misleading indicator because of which additional metrics are developed by combining sensitivity and specificity. The Area under the ROC Curve (AUC) of the model is to be used as the evaluation metric, as the data is highly imbalanced and using accuracy as the sole measure of performance evaluation will paint a false picture.

1.5 Data Approach

The methodology follows a 4-stage process – gathering data from web pages, identifying characteristics and preparing a data frame to represent the characteristics, applying various classification techniques to classify the pages and evaluating the models.

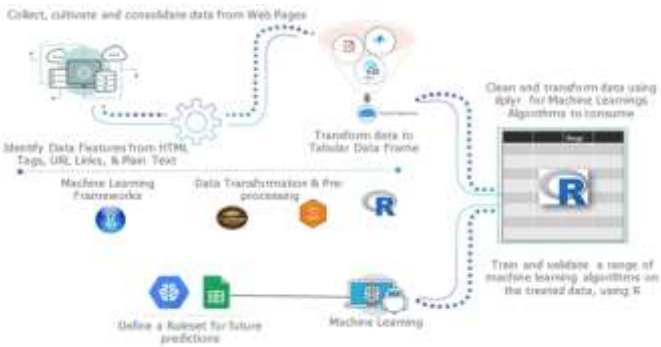

Fig -1: The Figure depicts the 4-Stage Process mentioned above

1.6 Literature Review

In the paper, "Naive Bayes Web Page Classification with HTML Mark-Up Enrichment,"[1] the study concentrates only the plain text of the HTML documents and are based on frequencies, and two weight functions which combine several criteria including information from HTML mark-up. In Nidhi Saxena's[2] paper "An Improved Technique for Web Page Classification in Respect of Domain Specific Search", the researchers use WSD, POS and ignore Meta Data in webpages in addition to highly intelligent Machine Learning models in order to reach at high accuracy of classification. It focuses only on HTML content for web page classification. In the article "Automated Classification of Web Sites using Naïve Bayes Algorithm," A. Patil[3] uniquely suggests that the home page of an HTML based website can be a source of conclusive information for classification of web pages into very broad categories. Therefore, the modelling strategies have been designed to concentrate on extracting specific features to help with the same.

What distinguishes the research from the work described in the research works above is the fact that the researcher have approached the problem with the goal of reducing the cost incurred as a result of web crawling and extensive text mining. Hence, our model greatly reduces the time and complexity involved with rigorous text mining and instead greatly relied upon the information present in URL link texts in order to devise a strategy to predict the presence of our target information in them. However, the researcher have also included one variable with features derived directly through the usage of text mining into the HTML pages of webpages for certain complementary information based on domain experience and examined its performance and significance. In addition to this, what also separates our work from the existing body of the work is our end goal to classify webpages on the basis of a highly business-specific goal rather than making broad classifications.

2. Modelling

There is a high class bias with only 914 webpages out of 11430 containing address information i.e. about 7.99% of the webpages; and 286 webpages out of the 11430 webpages containing executive information i.e. about 2.5% of the web pages.

As there were two different variables to predict, the problem was approached as a Multi Label Imbalance Problem. As the chances of success in fitting a single model onto a multilabel dataset were very slim. The researcher decided to build separate models to predict both the variables and predict the chances of a page containing address information or executive information, separately.

2.1 Resampling

As the dataset contains a high class imbalance, with respect to both the target variables, various resampling techniques, such as Ovun over Sampling and SMOTE were considered. However, resampling was not done as resampled data would not be representative of the population and would give a probability distribution that is not related to the actual scenario. Hence, no resampling was done. The data was split into randomly sampled training and test datasets, for the sake of thorough cross validation in the ratio 85:15, containing 9715 and 1715 records respectively.

2.2 Class Predictions

The AUC of each model was computed and the intersection of sensitivity and specificity curve plotted as the cut-off distribution was considered the optimal distribution for the below mentioned models.

2.2.1 Address Flag

Table 1. Performance of Different Models on Train Data for Address Flag

Algorithm	AUC	Accuracy	Specificity	Sensitivity	Negative Prediction Rate	Positive Prediction Rate
Lasso	88.53	82.11	82.17	81.40	28.30	98.08
Logit	88.72	83.53	80.10	83.83	29.98	97.99
Random Forest	93.63	87.91	83.85	88.26	38.18	98.44
Naïve Bayes	84.57	80.88	73.51	81.51	25.58	97.27
C5.0 – Rule Based	90.97	87.13	79.84	87.76	36.06	92.04

By comparing the Top Five models that performed well on the data, we can see that Random Forest with a Gini Split and a C5.0 with a Rule Based Tree is performing at similar levels with minimal drop in AUC and higher Negative Prediction Rate. As stated earlier, AUC has been taken as the primary evaluation metric, as the data is highly imbalanced.

With the business requirements and objectives in consideration, the rule based C5.0 is recommended for identifying pages with Address Information, as the performance of the model has been consistent in the training and testing phase. The Rulesets generated by the algorithm can also be extracted and the underlying logic can be implemented in any existing infrastructure to reduce the cost of implementation and enhance integration of the model.

Table 2. Performance of Different Models on Test Data for Address Flag

Algorithm	AUC	Accuracy	Specificity	Sensitivity	Negative Prediction Rate	Positive Prediction Rate
Lasso	88.5	80.92	81.43	80.87	27.60	97.98
Logit	89.91	82.8	81.43	82.92	29.92	98.03
Random Forest	90.05	85.55	74.29	86.56	33.12	97.41
Naïve Bayes	85.64	79.33	77.14	79.53	25.23	97.49
C5.0 – Rule Based	89.03	86.02	77.86	86.76	34.49	97.76

2.2.2 Executive Flag

Table 3. Performance of Different Models on Train Data for Executive Flag.

Algorithm	AUC	Accuracy	Specificity	Sensitivity	Negative Prediction Rate	Positive Prediction Rate
Lasso	95.61	93.42	67.07	94.11	23.03	99.01
Logit	95.94	87.83	89.96	87.77	16.02	99.70
Random Forest	99.11	96.93	89.56	97.12	44.94	99.72
C5.0 – Rule Based	99.94	99.33	98.39	99.36	80.07	99.96

When the Top 4 models built to identify pages with executive information are compared, it can be seen that all the algorithms have a really high Accuracy and Positive prediction rate, primarily due to the high class imbalance; however, C5.0 and Random Forest have again outperformed the other models in controlling classification. Due to the business reasons previously mentioned, the researcher would recommend the implementation of the rule based C5.0, over the Random Forest, as the predictions are easier to interpret and the logic is easier to integrate into business applications.

Table 3. Performance of Different Models on Train Data for Executive Flag.

Algorithm	AUC	Accuracy	Specificity	Sensitivity	Negative Prediction Rate	Positive Prediction Rate
Lasso	96.03	93.72	75.68	94.12	22.22	99.43
Logit	96.66	88.02	86.49	88.06	13.85	99.66
Random Forest	98.07	96.83	81.08	97.18	38.96	99.57
C5.0 – Rule Based	98.67	98.41	81.08	98.80	60.00	99.58

3. CONCLUSIONS

A prediction algorithm based on URL text alone is sufficient, at the initial level, to reduce hit count and the rigor of text analytics as above 90% accuracy figures were obtained in indicative tests.

Key words are a significant determinant for the presence of Executive Information in a webpage. The following keywords were found to be highly significant.

"history", "company", "people", "staff", "Board", "site", "about", "team".

Key words are a significant determinant for the presence of Business Address Information in a webpage. The following keywords were found to be highly significant.

"collections", "gallery", "News", "location", "blog", "Services", "Index", "Product", "contact", "Home_page".

After the initial stage, an independent model can be built on top of features derived from HTML bodies of websites to further increase efficiency, but it is not highly recommended if cost reduction is the primary criterion.

For predicting presence of Business Address and Executive Information in a webpage, Lasso, Logit (Logistic regression), Random Forest, C5.0 – Rule Based were found to be highly efficient. Any one of these or an ensemble model of two or more of these models can be deployed to replace the existing models.

Hidden variables may be looked for in order to feed into the recommended models to increase the accuracy figures in real time scenarios.

Way forward for data service companies is to automate the models built and build a Robotic Process Automation which would significantly reduce their Operational losses.

The classification models discovered sets of keywords displaying statistically significant bearing on the likelihood of finding Executive information and Business Address in the webpages examined. Deployment of such models at the stage of URL scraping, which precedes actual web text scraping, were found to be highly potent in increasing the strike efficiency and hence increasing the revenue of data service companies.

The performance figures derived from the entire exercise was satisfactory and the researcher was able to achieve accuracy figures greater than the target of 80% with multiple models.

REFERENCES

[1] V. Fernandez; R. Unanue; S. Herranz; A. Rubio. Naive Bayes Web Page Classification with HTML Mark-Up Enrichment. ICCGI 2006.

[2] V. Chandra; N. Saxena. An Improved Technique for Web Page Classification in Respect of Domain Specific Search, IJCA, 2014, Volume 102-4, 10.5120/17801-8615.

[3] A. Patil, B.V. Pawar , Automated Classification of Web Sites using Naive Bayesian Algorithm, IMECS 2012, Volume 1.

BIOGRAPHIES

Venkatesh Radhakrishnan is an experienced Research Analyst and Data Scientist with a demonstrated history of working in the Research & Analytics industry. With a Masters in Business Analytics and a Bachelors in Information Science, he is passionate about Neural Networks, Q-Learning Agents, Automation, Explainer Models, and Automated Machine Learning.