Large-Scale Web Page Classification

Sathi T. Marath
DNA 13
Ottawa, Canada
sathitm@gmail.com

Michael Shepherd
Computer Science
Dalhousie University
Halifax, Canada
shepherd@cs.dal.ca

Evangelos Milios
Computer Science
Dalhousie University
Halifax, Canada
eem@cs.dal.ca

Jack Duffy
Computer Science
Dalhousie University
Halifax, Canada
Jack.f.duffy@gmail.com

Abstract

This research investigates the design of a unified framework for the content-based classification of highly imbalanced hierarchical datasets, such as web directories. In an imbalanced dataset, the prior probability distribution of a category indicates the presence or absence of class imbalance. This may include the lack of positive training instances (rarity) or an overabundance of positive instances. We partitioned the subcategories of the Yahoo! web directory into five mutually exclusive groups based on the prior probability distribution. The best performing classification methods for a particular prior probability distribution were identified and used to design a content-based classification model for the complete (as of 2007) Yahoo! web directory of 639,671 categories and 4,140,629 web pages. The methodology was validated using a DMOZ subset of 17,217 categories and 130,594 web pages and we demonstrated statistically that the methodology of this research works equally well on large and small datasets.

1. Introduction

Over the past decade, web users have witnessed an exponential growth in the number of web pages accessible through popular search engines. Organizing the large volume of web information in a well-ordered and accurate way is critical for using it as an information resource. One way of accomplishing this in a meaningful way requires web page classification. Web page classification addresses the problem of assigning predefined categories to the web pages by means of supervised learning. This inductive learning process automatically builds a model over a set of previously classified web pages. The learned model is then used to classify new web pages.

Numerous classifiers proposed and used for machine learning can be applied for web page classification. These include Support Vector Machines (SVMs), k-Nearest Neighbor (k-NN), and Naïve Bayes (NB) classifiers. Empirical evaluations of these algorithms on selected small segments of popular web directories have shown that most of these methods are effective in web page classification. However, the effectiveness of these algorithms on very large web taxonomies like the Yahoo! directory and Open Directory Project (ODP) has not been thoroughly investigated. Web taxonomies like the Yahoo! directory and the Open Directory Project have hundreds of thousands of categories and millions of web pages. The sheer volume of categories and web pages makes large-scale web page classification an important consideration for web directories and search engines.

In contrast to the traditional benchmark datasets, web directories generally have complex statistical properties. This makes large-scale hierarchical web page classification significantly different from traditional text classification and from web page classification with limited categories and documents. Web directories usually exhibit a spindle distribution having more categories and documents in the middle of the hierarchy than at either the upper or the lower levels of the hierarchy.

Another distinguishing attribute of web directories is the skewed category distribution over the web pages. If we only consider the documents assigned directly to categories without counting in the documents assigned to their child categories, the number of documents per category follows the power law distribution [5, 6]. This indicates both imbalance within the dataset and the absence of a sufficiently large training set, i.e., rarity, within the dataset. In an imbalanced dataset, almost all examples belong to one class. When a machine learning algorithm is exposed to an imbalanced dataset, standard classifiers tend to focus on the large classes and ignore the small classes. In the case of rarity, the learning algorithm may find many different learning

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.229

IEEE computer society

rules within the decision boundary, all giving the same accuracy on the training data.

In web taxonomies such as Yahoo! the assignment of a web page into a category will not automatically grant this assignment to its parent categories or vice versa. The recursive assignment of the web pages of a category into its parent category helps to decrease the degree of rarity within web taxonomies. This process, however, results in the localized over-abundance of positive instances especially in the upper level categories of the hierarchy. When classifying categories with very large numbers of positive training instances, it is crucial to assess whether the classifier trained with a very large dataset is better than the one trained with a small subset of data. In theory, classifier performance should not be reduced when trained on a large dataset. However, classifiers using large datasets for training may not always be better, and may be slightly worse due to the much larger solution space.

The class imbalance, rarity and large-sample learning issues within a web directory dataset make applying classification algorithms on such directories very difficult. Earlier large scale web page classification research either overlooked the machine learning issues due to rarity, class imbalance and the over-abundance of training instances or addressed these issues using a common framework. This lead to either highly comprehensive or inadequate statements. At the time this research was conducted [8], the maximum number of web page categories ever classified is not more than 246,279 categories from Yahoo! [5, 6]. In their research the hierarchical SVMs lead to a Micro-F1 of 24%.

Based on probability distributions, we partitioned the subcategories of the Yahoo! web directory into five mutually exclusive groups. The effectiveness of different data level, algorithmic and architectural solutions to the associated machine learning issues, such as class imbalance and rarity, was explored [8]. Later, the best performing classification technologies for a particular prior probability distribution were identified and integrated into the Yahoo! Web directory classification model [8]. The methodology was validated using a DMOZ subset of 17,217 categories and 130,594 web pages and we proved statistically that the methodology of this research works equally well on a large (Yahoo!) and a small dataset (DMOZ).

The remainder of this paper is organized as follows. Section 2 is the literature review. The dataset used for this research is discussed in Section 3. The experimental setup used for this research is discussed in Section 4. Section 5 is the results and discussions. The recommendations of this research are summarized in Section 6.

2. Literature Review

A large number of statistical learning methods have been applied to the text classification problem in recent years. Some of them are regression models, nearest neighbor classifiers, Bayesian probabilistic classifiers, decision trees, inductive rule learning algorithms, neural networks and on-line learning approaches. Since a large number of methods and results are available, a cross-method evaluation is important to comprehend the current status of the text categorization research. The comparison of different text and web page classification methods, however, is very difficult due to the absence of a cohesive methodology for the matter-of-fact evaluation. Cross-method comparisons with a limited number of methodologies have been reported in the literature. However, these types of small-scale comparisons can either lead to highly comprehensive statements that are based on inadequate observations, or provide limited insight into a global comparison among a wide range of approaches.

The lack of a standard data collection is the main bottleneck for cross-method comparison in text categorization research. For a given dataset, there are many possible ways to introduce inconsistent variations. For example, the popular Reuters news story corpus has multiple versions depending on differences in the training, test and evaluation set combinations. Whether the reported classifier performance on the different versions of Reuters is comparable is not clear [14]. Incomparability across different evaluation measures used in individual experiments is another concern in cross-experiment evaluation [14]. Many measures such as recall and precision, accuracy or error, Precision-Recall breakeven point or the F1-Measure have been proposed and used for classifier evaluation. Each of these measures is designed to evaluate some characteristic of the categorization. However, none of them conveys identical or comparable information. There exist some difficulties in comparing published results of text categorization methods when they are evaluated using different performance measures. In general, one should be very vigilant while comparing the published text categorization research.

Due to the aforementioned issues, a comprehensive evaluation of different classification methods has not been reported. However, Yang and Liu [15] evaluated fourteen classifiers using the Reuter's corpus. The k-Nearest Neighbor (k-NN) classifier has shown the best performance. Other top performing classifiers listed in their research were Linear Least Square Fit (LLSF) and Neural Net. Rule induction algorithms like SWAP-1, RIPPER and CHARADE, show apparently good performance. Relatively worse performance was reported for Rocchio and Naive Bayes classifiers.

In a different study [14], the robustness of SVM, linear regression (LLSF), logistic regression (LR), Neural Net, Rocchio, Prototypes, k-Nearest Neighbor (k-NN), and the Naive Bayes classifier were evaluated when applied to a dataset with skewed category distribution. For a skewed dataset, SVM, k-NN, and LLSF significantly outperformed Neural Net and Naive Bayes classifiers.

Different studies [10, 4, 5, 6, 14] have demonstrated that the SVM has high training performance and low generalization error. However, SVMs when applied to an imbalanced dataset, produce a less effective classification boundary skewed to the minority class [1].

In general, empirical evaluations of popular classification algorithms such as SVMs, k-NN, and Naïve Bayes classifiers on selected small segments of popular web directories have shown that most of these methods are effective in web page classification. However, available classification research on reasonably sized subsets of popular web directories conclude that in terms of effectiveness these classification algorithms cannot fulfill the classification needs of very large-scale taxonomies [2, 3, 5, 6, 13]. In their research, even with the best classifier setting, hierarchical SVMs lead to a Micro-F1 of 24% and a Macro-F1 of 12%. This study concludes that in terms of effectiveness neither flat nor hierarchical SVMs can fulfill the classification needs of very large-scale taxonomies. The skewed distribution of large web directories like Yahoo! with many extremely rare categories makes SVMs performance ineffective. Their research, however, overlooked the impact of class imbalance, and absolute rarity while classifying an imbalanced dataset.

McCallum et al, [9] addressed the rarity issue through a Bayesian framework and shrinking the estimate parameters as one descends the hierarchy, improving the accuracy on classes with 50 documents or less by 10%.

The effectiveness of hierarchical SVM while classifying the top two levels of categories of the LookSmart dataset has also been studied. This research reported a macro-average F1 measure of 57.2% for the top level 13 categories and 47.6% for the 150 second level categories [2, 3]. There is a drop in performance in going from 13 to 150 categories. Conversely, this study uses the top two levels of the LookSmart categories only and the conclusions might not generalize to the case of classifying hundreds of thousands of categories.

Xue et al. [13] addressed the large-scale web page classification in a two phase process. In the first phase, a category-search algorithm is executed to acquire the category candidates for a given dataset. Based on the category candidates, the large scale hierarchy is pruned and classification is performed on the pruned subset of the original hierarchy. In this research, a statistical-language-model based classifier using n-gram features is used for classification. The performance of the proposed algorithm is evaluated on the Open Directory Project with over 130,000 categories**.** With this approach the Micro-F1 at the fifth level of the hierarchy is 51.8%, whereas for top-down based SVM classification algorithms the Mico-F1 at the fifth level is 29.2%.

The literature review cited in this section provides some insight to the average performance of different classification algorithms. Unfortunately, the few available web page classification studies on reasonably sized subsets of popular web directories conclude that in terms of effectiveness these classification algorithms cannot fulfill the classification needs of very large-scale taxonomies. Such web directories have hundreds of thousands of categories, deep hierarchies, class imbalance and rarity over the hierarchies. The class imbalance and rarity make applying classification algorithms to such datasets very difficult and the problem has not been thoroughly studied.

3. Dataset

This research [8] aims to design a unified classification model or framework for highly imbalanced hierarchical datasets. The complete Yahoo! web directory from 2007 of 639,671 categories and 4,140,629 web pages organized in a 17 level hierarchy was used in this research. Figure 1 illustrates the spindle distribution of the categories by hierarchical depth. In an imbalanced dataset, the prior probability distribution of a category indicates the presence or absence of class imbalance, rarity and large-sample learning issues due to the overabundance of positive instances. Based on the prior probability distribution, we subdivided the entire set of categories of the Yahoo! directory into 5 mutually exclusive groups as given in Table 1. The rationale for these 5 class sizes is explained later.

Classification algorithms, when applied to categories of 1000 or more labeled instances should address the machine learning issues due to class imbalance and large-sample learning. Conversely, classification algorithms, when applied to rare categories of 10 to 99 labeled web pages should address the machine learning issues due to class imbalance and rarity.

Table 1: The Prior Probability Distribution Range of Yahoo! Web Directory

Category Size	Number of categories	% of categories
More than 1000 labeled web pages	988	0.16
100 to 999 labeled web pages	10,025	1.58
10 to 99 web labeled pages	121,329	19.06
1 to 9 labeled web pages	504,240	78.82
Categories without any labeled web pages	3,089	0.38

Reasonably sized categories of 100 to 999 labeled web pages make up 1.58% of the categories. However, the abundance or shortage of negative instances in the sibling categories makes these categories imbalanced. There are 504,240 Yahoo! categories containing 1 to 9 web pages. This forms 78.82% of the total Yahoo! categories. Due to the extreme lack of training instances, no individual classifiers are designed for these categories and they are subsumed by their parent categories, reducing the actual number of categories by this number.

4. Methodology

A breadth-first approach was taken to the classification. We addressed the class imbalance problem by focused over-sampling and under-sampling of the negative instances. This prevents information loss due to the sub-sampling of positive instances.

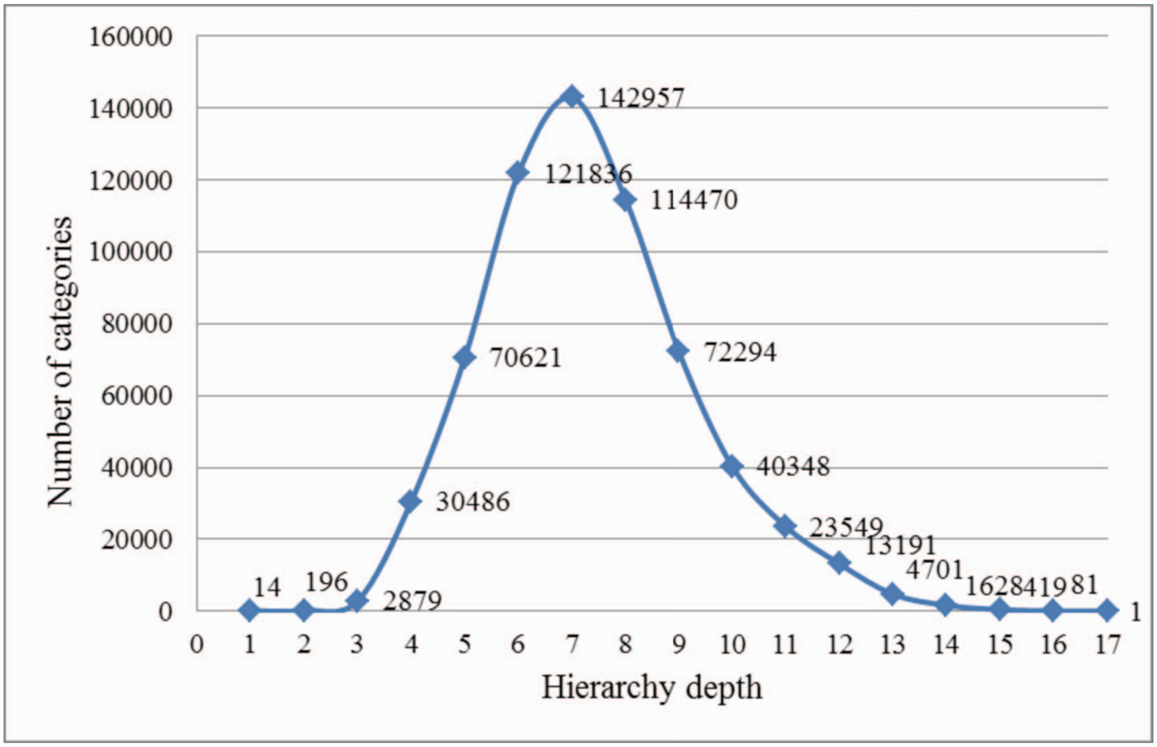

Figure 1. Spindled Category Distribution of Yahoo! Web Directory

The macro-averaged F1-measure is used to evaluate the performance of the rare classes. Whereas accuracy, the popular evaluation metric for classification, ignores the performance of rare categories, the F1-Measure does not. This is because both the True Positive (TP) Rate and the False Positive (FP) Rate used to calculate recall and precision are defined with respect to the positive class. The rarity problem associated with the classification of categories of very few labeled instances is addressed using adaptive over-sampling. The large-sample learning problem associated with the classification of the large categories of hundreds of thousands of positive instances is addressed using ensemble architectures.

Different methodologies were applied to the three groups of different size categories; more than 1000 labeled pages, 10 to 99 labeled web pages, and 100 to 999 labeled web pages. This last group of categories is described last as it was necessary to determine the partition boundaries between this size category and the larger and smaller sized categories.

4.1. Classification of Yahoo! Categories with 1000 or More Web Pages

In ensemble learning, sub-sampling is performed on the original dataset to create multiple samples. Member classifiers are trained from each sample. Thus, an ensemble classifier comprises a group of member classifiers. The category of a new web page is determined by voting by the member classifiers. If sufficient training instances have been taken, the ensemble formed by the multiple member classifiers can produce the same result as much larger samples [11]. The optimal sample size varies with the dataset. The sample size optimization followed in this research includes training multiple member classifiers taking increasingly larger samples, observing the trends, and stopping when no progress has been made. The sample size used in the first iteration is ensured to be the representative of the original dataset. A sample for a category, say C_i, is drawn from that category and all its child categories and combined with an equal number of negative instances. The size of the samples is increased gradually in the subsequent iterations to allow for the chances of improving performance with a larger sample size.

In this procedure, it is important to analyze the achieved benefit before moving to the next iteration of a higher sample size. For each member classifier, the average of the TP and the FP after 3-fold cross validation is calculated. A quality factor, defined as a function of the average TP and FP of the member classifiers and the variance of the TP and the FP is used to measure the quality of ensembles formed on each iteration. A high TP and low FP is desirable for producing a classifier with high recall and precision. The higher the variance, the greater will be the dissimilarity between the member classifiers. Hence the Quality Factor (QF) of an ensemble formed of N member classifiers is defined as,

$$\text{QF} = \text{Average of TP} / (\text{Average of FP} \times \text{Average variance of TP} \times \text{Average variance of FP}).$$

The slope in degrees between two consecutive QFs (y-axis) and the normalized average sample size (x-axis) is measured. A slope of zero degrees means there is no improvement in the classifier performance between the two consecutive sample sizes and no further iteration is needed. However, this is an optimal situation that rarely happens. If the slope in degrees between two consecutive QFs and the average sample size is less than or equal to the predefined threshold of one degree, the member classifiers formed from the higher sample size is taken as the optimal ensemble classifier for C_i. These member classifiers will be used for voting to determine the category of a new web page.

4.2. Classification of the Yahoo! Rare Categories with 10 to 99 Web Pages

The lack of training instances and the negative or positive dominant class imbalance are the main machine learning issues associated with the classification of Yahoo! rare categories of 10 to 99 labeled instances. Over-sampling is a popular data level approach to address rarity in which instances may occur in the same sample multiple times. In its basic form, it duplicates the rare category dataset. However, crude over-sampling may result in classifier over fitting.

Weiss proposed an adaptive re-sampling architecture to address classifier over fitting and related issues associated with random sampling that showed much superior performance [11, 12]. Inspired by this research, we designed a modified version of Weiss's adaptive re-sampling approach to address the rarity associated with Yahoo! categories of 10 to 99 web pages.

4.2.1. Adaptive Over-Sampling. In basic over-sampling, the instances for over-sampling are randomly selected with a probability 1/N, where N is the full sample size. The basic theory of classifier design using adaptive sampling is to iteratively induce new classifiers by increasing the weight of erroneously classified cases in the training set in the next iteration. However, in adaptive-over sampling, for each case in the full training set, a record of the current solution performance is kept. Cases with a large error for the current solution are over-sampled with increased frequency. In the subsequent iterations, the instances for over sampling are selected with probability P_i which is determined by the relative probability of the error of the case i. This process avoids over fitting.

For each category C_i of 10 to 99 web pages, in the 1st iteration, available positive instances of C_i and all child categories are drawn and combined with an equal number of negative instances from the sibling categories of C_i and a single classifier after 3-fold cross validation is designed. Three-fold cross validation was used throughout rather than a higher n-fold to reduce computational complexity. The average F1-Measure of classification, average recall of C_i and average recall of all child categories are measured.

In the 2nd iteration, the training instances of C_i and all child categories of C_i with lower average recall than a predefined threshold recall of 75% are identified and over-sampled by 10%. The test dataset is not over-sampled. The macro-averaged F1-Measure of C_i is calculated. The new average recalls of the C_i and child categories are calculated and used for deciding the over-sampling criteria in the next iteration. If the increase in the F1-Measure of two consecutive iterations is less than 1% no further over-sampling is performed for the category C_i. Otherwise the algorithm proceeds to the next iteration and repeats the procedure of iteration-2.

4.3. Classification of Imbalanced Yahoo! Categories with 100 to 999 Web Pages

Before designing a classification architecture for this group, we analyzed the appropriateness of this arbitrarily fixed range.

In our ensemble architecture, sub-sampling is performed on the original dataset to create multiple member classifiers. The sample size optimization followed in this research includes training multiple member classifiers taking increasingly larger samples, observing the trends, and stopping when no progress has been made. If the ensemble learning turns unprofitable for the given Yahoo! subcategory, our ensemble model converges into a single classifier and undergoes three fold cross validation. Of the total 988 Yahoo! categories of 1000 or more positive instances, 24 categories converged like this into single classifiers. This includes 89.56% of the categories of less than 1700 positive instances. Based on this, we defined the lower limit category size for the ensemble architecture to be 1000 instances.

In the adaptive over-sampling based learning architecture designed to classify rare categories with 10 to 99 web pages, all categories of more than 75 web pages are oversampled by less than 20%. Whereas, all categories of less than 45 web pages are oversampled by more than 50%. The highest over-sampling is performed with the Yahoo! categories of less than 20 labeled web pages. Based on this, we defined the upper limit category size for adaptive over-sampling to be 100 instances as categories larger than this do not exhibit rarity.

Negative or positive dominant class imbalance is the only machine learning issue associated with the classification of Yahoo! categories of size 100 to 999 instances. The class imbalance issue is addressed by either over or under-sampling the negative dataset. This prevents the information loss of positive instances due to sub-sampling. Both these sampling techniques decrease the degree of class imbalance by altering negative dataset distribution. Text content from the body of the web pages is used for classification. Taking complete positive instances and an equal number of negative instances from the sibling categories, 3-fold cross validation is conducted on each category of this group.

4.4. Yahoo! web directory classification model

Integrating the above-mentioned three classification architectures, we designed a hierarchical machine learning model for the content based classification of the Yahoo! web directory. The distribution of ensemble, sub-sampled and adaptively over-sampled classification units in the designed Yahoo! classification model is summarized in Table 2.

The model contains 132,342 classification units distributed in 14 layers. Each unit of this hierarchical classification model maps a Yahoo! web directory category of more than 10 labeled instances. Of the total 132,342 classification units, 988 units consist of ensembles of multiple member classifiers. The remaining classification units are modeled after sub-sampling or adaptive over-sampling. The first layer of the model contains 14 ensemble units only (more than 1000 instances). The second layer of the model contains 196 classification units, of which, 133 units are ensemble classification units and the remaining units are sub-sampled classification units designed for 63 categories of 100 to 999 instances. The third layer includes 995 adaptively over-sampled units designed for categories of 10 to 99 instances.

5. Results

This research was conducted on the Atlantic Computational Excellence Network (ACEnet). ACEnet is a High Performance Computing (HPC) environment providing distributed HPC resources, visualization and collaboration tools. A maximum entropy classification algorithm was used to induce the classifiers. The maximum entropy classifier, when used for web page classification, does not require feature selection or feature extraction prior to the classification. The MEGA Model Optimization Package is used to implement the maximum Entropy Classifier (MEGA Model Optimization Package, 2007). In this research, a number of different classifiers and architectures were evaluated before selecting the final configuration and these evaluations may be found in [8].

5.1. First Calculation of Macro F1

The average classifier performances across the hierarchy depth for different groups of Yahoo! categories are summarized in Table 3. The overall macro-average F1 Measure is 81.02.

Whether the methodology of this research will work equally well when applied to the content based hierarchical web page classification of larger or smaller dataset is also examined. These experiments were conducted on a hierarchical subset of the DMOZ web directory. At the time of our crawling in October, 2009, there were 602,410 categories and 4,519,050 web pages in the topmost 14 levels of the DMOZ web directory. Like the Yahoo! web directory, most of the DMOZ categories are extremely rare with fewer than 10 labeled web pages.

The effectiveness of the integrated model classification model developed for the Yahoo! Web directory classification was validated using a DMOZ subset of 17,217 categories and 130,594 web pages. This dataset was downloaded from the Large Scale Hierarchical Text classification (LSHTC) Pascal Challenge [7]. The LSHTC Challenge is a hierarchical text classification competition using large datasets based on the DMOZ web directory. The detailed category distribution of this subset is summarized in Table 4.

Table 2: The Structural Information of the Yahoo! Classification model

Level	Adaptively over-sampled class. units	Sub-sampled classification units	Ensemble classification units
1	0	0	14
2	0	63	133
3	995	728	233
4	8,420	2,222	347
5	11,530	2,895	138
6	28,705	2,161	87
7	33,997	1,082	33
8	16,569	529	3
9	10,981	229	
10	5,776	84	
11	3,043	22	
12	1,002	10	
13	268		
14	43		
Total	**121,329**	**10,025**	**988**

There are 62 DMOZ categories containing 1000 to 100,000 web pages. Ensemble learning combined with the Maximum Entropy Classifiers, the best performing classification technology when applied to the Yahoo! categories of 1000 or more positive instances, is applied to classify the DMOZ categories of this group.

There are 632 DMOZ categories containing 100 to 999 web pages. These are reasonably sized categories for efficient machine learning; however, dominance or scarcity of negative instances of the sibling categories results in the class imbalance. In addition to the class imbalance, 5,649 categories of this DMOZ subset are rare categories of 10 to 99 positive training instances. Adaptive over-sampling combined with Maximum Entropy Classifiers, the best performing categorization technique for Yahoo! categories of 10 to 999 web pages, is used to classify these categories. There are 10,873 Yahoo! categories containing 1 to 9 web pages. Due to the lack of training instances, no individual classifiers have been designed for this group.

Table 3: Average F1 Measure Achieved for Yahoo! Web Directory Classification

Level	Categories with 1000 or more labeled instances	Categories with 100 to 999 labeled instances	Categories with 10 to 99 labeled instances
1	95.74		
2	95.01	91.33	
3	90.84	84.90	88.68
4	88.87	84.65	87.09
5	87.21	82.29	85.71
6	85.22	82.16	83.39
7	81.97	78.09	81.14
8	78.06	72.61	80.33
9		69.63	78.36
10		59.50	75.82
11		56.74	74.31
12			72.48
13			71.24
14			69.74
Average	**87.86**	**76.19**	**79.02**

Table 4: Detailed category distribution of DMOZ subset

Level	Total Categories	Categories with 1000 or more labeled instances	Categories with 100 to 999 labeled instances	Categories with 10 to 99 labeled instances	Categories with 1 to 9 labeled instances
1	9	9	0	0	0
2	311	33	112	118	48
3	2522	15	242	1144	1121
4	6993	4	175	2276	4538
5	7382	1	103	2111	5167

The average classifier performance of the DMOZ subset is summarized in Table 5. The macro-averaged F1-Measure of the DMOZ subset achieved in this research is 84.85%. The highest average F1-Measure reported for this dataset in the LSHTC Pascal Challenge is 35.49% [7]. In their research, whether any hierarchy pruning or expansion was performed prior to the classification is not clear. A comparison of the average classifier performance on the Yahoo! Web directory and the DMOZ subset is summarized in Table 6.

A two-variable Chi-Square test of independence was conducted to check for significant difference in the average F1-Measures between the Yahoo! web directory and the DMOZ subset if any, across the 3 groups of Yahoo! and DMOZ subset categories (categories with more than 1000 labeled instances, categories with 100 to 999 labeled instances and categories with 10 to 99 labeled instances).

Table 5: Average Classifier Performance of DMOZ Subset

Category Group	# of Categories	Macro F1	Rare Category Recall
1000 or more	62	86.27	79.68
100-999	632	84.30	77.50
10-99	5649	83.99	77.86
1-9	10,873	Macro Averaged Recall: 70.58	

Table 6: Comparison of Macro F1-Measure for Yahoo! And DMOZ

Category Group	Yahoo!	DMOZ
1000 or more	87.86	86.27
100-999	76.19	83.99
10-99	79.02	84.30

The Chi square test for significant difference was 0.56 (p < 0.975). There is no significant difference in the method's ability to predict in either dataset. Thus we conclude that the method is equally appropriate for very large and smaller datasets, such as the Yahoo! and the DMOZ datasets.

5.2. Final Node F1 and Tree Induced Error

An alternative evaluation method is one in which the macro F1-measure is calculated only for the final class that the web page resides in. In addition, a tree induced error can also be calculated.

In the LHSTC challenge the tree induced error is defined as follows: "For any pair of categories (u,v) the tree induced error earned by predicting the label u when the correct label is v is the tree distance between u and v. The tree distance between two nodes is defined to be the number of edges along the unique path from u to v". In order to use the macro-average cumulative tree-induced error here we add all the tree distances and divide them by the total number of web pages.

The Final Node macro averaged F1-Measures for the Yahoo! web directory and the DMOZ subset are 78.34% and 81.98%, respectively. The macro-average tree induced error for the Yahoo! and DMOZ subsets is 3.44 and 2.24 respectively. The lowest tree induced error reported in the DMOZ LHSTC challenge was 3.08 and the achieved macro-F1 was 34 %. The leaf node based macro-F1 and tree induced error for Yahoo! and DMOZ are summarized as Tables 7 and 8, respectively.

The lower values for the tree induced error for the DMOZ dataset as compared to the Yahoo! Dataset may be explained by their relative sizes. The Yahoo! Dataset is considerably larger and thus a misclassified instance may be further from its appropriate category in the hierarchy than a misclassified instance in the DMOZ dataset.

Table 7: Leaf node based macro-F1

Category	Yahoo!	DMOZ
1000 or more labeled instances	83.13	84.09
100-999 labeled instances	74.88	80.51
10 to 99 labeled instances	77.02	81.33

Table 8: Tree Induced Error

Category	Yahoo!	DMOZ
1000 or more labeled instances	3.40	2.07
100-999 labeled instances	3.69	2.22
10 to 99 labeled instances	3.24	2.43

5.3. Caveat

Although our results on the DMOZ data set appear to be much superior that those obtained in the LSHTC challenge, the results in the challenge were obtained using the whole number of categories whereas our approach subsumes categories with fewer than 10 instances into the parent category. As classification models for scarce categories are less robust, the LSHTC results would be expected to be not as good as our results.

6. Conclusions

Traditional classification algorithms assume the target classes of the dataset share similar prior probability distributions. However, in real-world datasets like web taxonomies this similar prior probability assumption does not hold. For example, popular web directories have hundreds of thousands of categories, deep hierarchies, class imbalance and rarity (very small classes) within the dataset. These properties make applying classification algorithms to such datasets very difficult. The available classification results on reasonably sized subsets of popular web directories conclude that in terms of effectiveness, the popular classification algorithms cannot fulfill the classification needs of very large-scale taxonomies.

Earlier large scale web page classification research either overlooked the machine learning issues due to rarity, class imbalance and over-abundance of training instances or addressed these issues using a common framework. Therefore they are insufficient to address the challenges of large-scale web page classification problem, or provide only limited insight to the real challenges of large-scale web page classification.

In this research we partitioned the subcategories of the Yahoo! web directory into five mutually exclusive groups based on the prior probability distribution and machine learning issues associated with such an imbalanced distribution. A classification model for the complete Yahoo! web directory of 639,671 categories and 4,140,629 web pages was designed and implement. Afterward, the methodology was cross-validated using a DMOZ subset of 17,217 categories and 130,594 web pages and we demonstrated statistically that the methodology of this research works equally well on both large and small datasets.

7. References

[1] Akbani, R., Kwek ,S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. *15th European Conference on Machine Learning (ECML)*, (pp. 39-50).

[2] Chen, H. (2000). Bringing order to the web:automatically categorizing search results. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* , 145-152.

[3] Dumais, S.,& Chen, H. (2000). Hierarchical classification of Web content. *SIGIR*, (pp. 256-263).

[4] Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). RCV1: a new benchmark collection for text categorization research. *Journal of Machine Learning Research* , 361-397.

[5] Liu,T., Yang, Y., Wan,H., Zeng,H., Chen,Z.,& Ma,W. (2004). Support Vector Machines Classification with A Very Large-scale Taxonomy. *SIGKDD Explorations* , 1-36.

[6] Liu, T., Yang, Y., Wan, H., Zhou, Q., Gao, B., Zeng, H., Chen, Z. & Ma, W. (2005). An Experimental Study on Large-Scale Web Categorization. WWW 2005., Chiba, Japan, pp. 1106-1107.

[7] LSHTC Pascal Challenge, [http://lshtc.iit.demokritos.gr/node/23] Last accessed on February,19,2012

[8] Marath, S. (2010). Large-Scale Web Page Classification. [http://dalspace.library.dal.ca/bitstream/handle/10222/13130/Marath_Sathi_PhD_CSCI_December_2010.pdf?sequence=1]

[9] McCallum, A., Rosenfeld, R., Mitchell, T. and Ng, A.Y. (1998). Improving Text Classification by Shrinkage in a Hierarchy of Classes. ICML, Madison, Wisconsin.

[10] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys , 1-47.

[11] Weiss, S. M., & Indurkhya, N. (1998). *Predictive data mining-A practical approach.* Morgan Kaufmann Publishers.

[12] Weiss, S. M., Apte, C., Damerau, F.J., Johnson, D.E., Oles, F.J., Goetz, T., Hampp, T. (1999). Maximizing text-mining performance. *Intelligent Systems and Their Applications* , 63 – 69.

[13] Xue,G., Xing, D., Yang,Q., & Yu,Y. (2008). Deep Classification in Large-scale Text Hierarchies. *SIGIR.*

[14]Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval* , 69-90.

[15] Yang, Y.,& Liu, X. (1999). A re-examination of text categorization methods. *SIGIR* , 42-49.