International Journal of Engineering Inventions
ISSN: 2278-7461, ISBN: 2319-6491, www.ijeijournal.com
Volume 1, Issue 10 (November2012) PP: 52-58

A New Integrated Machine Learning Approach for Web Page Categorization

B.V.Swathi[1], A.Govardhan[2]
[1]*Department of Computer Science MIPGS, Hyderabad*
[2]*Professor in CSE & Director of Evaluation, JNTU, Hyderabad*

Abstract:—Clustering is an unsupervised task whereas classification is supervised in nature. In the context of machine learning, classification of instances of a dataset is carried out by a classifier after the classifier is made to learn the model from a training dataset. The training data consists of instances which are labeled by a human expert. The labels are the classes into which the instances of the dataset are divided and are fixed by the human expert. The essence is that human intervention is required in the form of preparing the training data for the machine to carry out the task of classification. Clustering of large datasets is universally accepted as a difficult problem, since it tries to group instances together, without the helping hand of the human supervisor. Also, the time complexity of algorithms such as K-Medoids is unacceptably high, even for moderately large datasets. The work reported in this paper aims to integrate both clustering and classification and test approach in the domain of web page categorization. Rather than using training data created by a human expert for classification, clustering is used in preparing the training data for the classifier.

Keywords:—Classification, Clustering, Web Search, Snippets

I. INTRODUCTION

In the recent past, the World Wide Web has been witnessing an explosive growth. Information is kept on the web in various formats and the content is dynamic in nature. All the leading web search engines, namely, Google, Yahoo, Askjeeves, etc. are vying with each other to provide the web user with the appropriate content in response to his/her query. In most cases, the user is flooded with thousands of web pages in response to his query and it is common knowledge that not many users go past the first few web pages. In spite of the multitude of the pages returned, most of the time, the average user does not find what he/she is looking for in the first few pages he/she manages to examine. It is really debatable as to how useful or meaningful it is for any search engine to return thousands of web pages in response to a user query. In spite of the sophisticated page ranking algorithms employed by the search engines, the pages the user actually needs may actually get lost in the huge amount of information returned. Since most users of the web are not experts, grouping of the web pages into categories helps them to navigate quickly to the category they are actually interested and subsequently to the specific web page. This will reduce the search space for the user to a great extent. It is strongly believed and felt that the experience of a person using a web search engine is enhanced multifold if the results are nicely categorized as against the case where the results are displayed in a structure less, flat manner.

Classification and clustering are the two tasks which have been traditionally carried out by human beings who are experts in the domain of application. But in this electronic age, with the explosion in the amount of information available on the net, it is becoming increasingly difficult for human experts to classify or cluster all the documents available on the World Wide Web. Hence, it is increasingly evident that machine learning techniques be used instead of human experts, to carry of the tasks of document classification and clustering.

In the machine learning approach to text classification, the set of rules or, more generally, the decision criterion of the text classifier is learned automatically from training data. This approach is also called statistical text classification if the learning method is statistical. In statistical text classification, a number of good example documents (or training documents) from each class are required for training the classifier. The need for manual classification is not eliminated since the training documents come from a person who has labeled them where labelin*g* refers to the process of annotating each document with its class. Clustering algorithms group a set of documents into subsets or clusters. The goal is to create clusters that are coherent internally, but clearly different from each other. In other words, documents within a cluster should be as similar as possible and documents in one cluster should be as dissimilar as possible from documents in other clusters. Clustering is the most common form of unsupervised learning. No supervision means that there is no human expert who has assigned documents to classes. In clustering, it is the distribution and makeup of the data that will determine cluster membership In supervised classification the goal is to first learn and then to use a categorical distinction that a human supervisor imposes on the data. In unsupervised learning, of which clustering is the most important example, a similarity measure is used to group similar objects together.

The methods used for text clustering include decision trees (S.Dumais et al., 1999; Hikargupta et al., 1997; U.Y.Nahm and R.J.Mooney, 2000; Y.Yang et al., 1999), statistical analysis (D.Freitag, 1999; T.Hofmann, 1999; H.Kargupta et al., 1997), neural nets (T. Honkela et al., 1997), inductive logic programming (W.W.Cohen, 1995; M.Junker et al., 1999), and rule-based systems(S.Scott and S.Matwin, 1999; S.Soderland, 1999). Most of the document clustering methods that are in use today are based on the Vector Space Model (K.Aas and L.Eikvil, 1999; G.Salton et al., 1975), which is a very widely used data model for text classification and clustering. The Vector Space Model represents documents as a feature vector of the terms that appear in all the document set. Each feature vector contains term weights of the terms appearing in that

document. Similarity between documents is measured using one of the several similarity measures that are based on such a feature vector. Classification can be done more efficiently by reducing the features using feature reduction algorithms.

The key input to a clustering algorithm is the similarity measure. In document clustering, the similarity/distance measure is usually vector space similarity or distance. Different similarity/distance measures give rise to different cluster formations. Thus, the similarity/distance measure is the main means by which the outcome of clustering can be influenced.

In this paper a new approach which integrates classification and clustering, called Integrated Machine Learning Approach(IMLA) is presented. In the process of integrating clustering and classification, the approach uses the Find-K algorithm (B.V.Swathi and A.Govardhan, 2009) and the modified QuickReductalgorithm (B.V.Swathi and A.Govardhan, 2009). The new method is applied to the domain of automatic web page categorization. Section 2 presents an overview of the proposed methodology with brief explanation of the steps involved. Section 3 presents the experimental setup and the results obtained. Section 4 presents the conclusion.

II. METHODOLOGY OF IMLA

The steps involved in the newly proposed technique are as follows:

1. Creating the dataset
2. Finding the Number of clusters(K) Using the Find-K algorithm
3. Labeling the Clustered Web Pages to Create the Training Dataset for the Classifier
4. Using the QuickReduct algorithm, reduce the dimensionality of the dataset
5. Classify the remaining dataset

The steps are elaborated in the following subsections

2.1 Creating the dataset

A dataset consisting of M web page snippets returned by any search engine in response to a given query should be initially created. This can be done either by manually copying all the web snippets into a text file, or by using a program. In order to clearly bring out the difference between the selected web snippets, the keywords forming the query submitted to the search engine should be removed from the snippets. This is done keeping in mind the fact that all the web pages contain these keywords in them even though they belong to different categories. These common words tend to mislead the categorization process and therefore, removed.

Once the dataset consisting of the web snippets is created, a part of it consisting of N web snippets, where $N < M$, is selected. The selection of this dataset is very critical in the accuracy of the overall result. In the real world scenario, the instances can be randomly picked and should form a sizeable portion of the initial dataset containing M pages.

2.2 Finding the clusters

The first step in creating an automatically labeled training dataset for a classifier is to cluster the data using a partitioning based clustering algorithm. The reason for using a partitioning based clustering algorithm is that it produces a set of flat, structure less, disjoint clusters, which can be treated as classes to which the instances of the dataset belong. But the biggest challenge of running a partitioning algorithm such as K-means or K-Medoids is to know the number of partitions K, which the algorithm needs as input. Find-K[3] algorithm uses the K-medoids clustering algorithm to automatically predict the correct number of clusters that are present in a given text dataset. In this paper, the Find-K algorithm has been used to predict the number of clusters. Further, the K-Medoid algorithm is used to carry out the clustering task on the dataset subsequent to the determination of the number of clusters.

2.3 Labeling the clustered web pages to create the training dataset for the classifier

In a way, the clustering task, carried out by an algorithm, can be viewed as similar to the creation of a training data by a human expert. There is, of course, a huge difference in the abilities of an algorithm and a human expert in assimilating the similarity or difference between a set of web page snippets. Web snippets are essentially made up of text and it is common knowledge that human brains, even today, are way ahead of machines in the area of language processing. But this presents us with the challenge. Clustering, which is an unsupervised machine learning activity, groups data instances based on their similarity. The measurement of similarity is very critical to the process of clustering since it acts as an index to which two instances may belong together.

The human expert, on the other hand, depends on the knowledge and experience he or she has gained over the years in attaching labels to the instance of a dataset.

Once the instances are clustered, the number of the clusters to which an individual instance belongs is attached as a label to that instance. In this way, all the instances are attached with their corresponding cluster number which acts as the class label. A training dataset is thus created.

2.4 Dimensionality Reduction Using Modified QuickReduct

Once the dataset set is clustered and subsequently labeled, the modified QuickReduct algorithm [4] is applied to the dataset to reduce its dimensionality. In one case, the original dataset containing 625 features (terms) was later represented by just 5 features after reduction.

2.5 Classify the test instances

Since only a part of the total dataset is used as the training set, the other part is now used to test the accuracy of classification. For the purpose of classification, an implementation of the well known C4.5 algorithm, known as J48, from the WEKA toolbox[18]has been used. This particular implementation of the classifier provides a facility wherein we can train the classifier with the training data and then obtain the predictions on the unlabelled test dataset. While using separate

training and test dataset files, it must be ensured that the internal representation of both the training set and the test set is exactly the same.

III. EXPERIMENTAL SETUP AND RESULTS

In order to obtain a perspective on the performance of IMLA, its results are compared with two traditional machine learning approaches.

- Pure Partitional Clustering (Using a Known K-value)
- Pure Classification with a human expert created training data

In order to carry out the experiments, 110 web snippets returned by the Google search engine in response to the query 'apple' have been manually collected. This dataset formed the basis for all the experiments carried out and reported in this work. The instances of the dataset have been identified to belong to 6 different categories.

3.1 Case I(80 Training, 30 test)

In this case, 80 out of the 110 web snippets have been selected to form the training dataset. To begin with, the Find-K algorithm is run on this dataset and the number of clusters has been found to be 6. The dataset is then clustered using the K-Medoids algorithm into 6 clusters. This dataset of 80 instances is now turned into a training data by attaching the corresponding cluster number/name to the individual instances. The remaining 30 instances are used as the test data and the results are noted. Further, the dataset containing 80 instances is labeled by the authors and again used as the training set, for comparison purposes. This is to compare the effectiveness of automatic generation of training data with the conventional method of human generated training data. These results are further compared with pure clustering, where the entire dataset consisting of 110 instances is clustered into 6 clusters.

The confusion matrices obtained after applying IMLA, classification and clustering are presented in Tables.1,2 and 3. From the confusion matrix, it is very easy to identify the number of true positives, false positives and false negatives. By examining a row, we can obtain the true positives and the false negatives. On the other hand, the column values provide us with the true positives and the false positives. Tables 4, 5 and 6 contain a summary of the results obtained using the three different methodologies, i.e., that of the newly proposed integrated method and the traditional clustering and classification. The comparison has been based on three parameters, namely, precision, recall and F-Measure. The precision, recall and F-Measure have been calculated in the following manner.

Precision = True Positives / (False Positives+True Positives)
Recall = True Positives/(False Negatives + True Positves)
F-Measure = 2* Precision*Recall/(Precision+Recall)

classified as >>	a	b	c	d	e	f
a = ipod	5	0	0	0	0	0
b = trailer	0	5	0	0	0	0
c = itunes	1	0	4	0	0	0
d = laptop	0	0	0	5	0	0
e = iphone	1	0	0	0	4	0
f = fruit	0	0	0	0	0	5

Table 1: Confusion matrix for IMLA-Case I

classified as >>	a	b	c	d	e	f
a = ipod	4	0	1	0	0	0
b = trailer	0	5	0	0	0	0
c = itunes	0	0	5	0	0	0
d = laptop	0	0	0	5	0	0
e = iphone	1	0	0	0	4	0
f = fruit	0	0	0	0	0	5

Table 2: Confusion Matrix for Classification-Case I

classified as >>	a	b	c	d	e	f
a = ipod	18	0	1	1	0	0
b = trailer	0	25	0	0	0	0

c = itunes	1	1	13	0	0	0
d = laptop	0	0	0	14	1	0
e = iphone	1	0	0	0	9	0
f = fruit	1	0	0	0	0	25

Table 3: Confusion Matrix for Clustering-Case I

	Precision		
Classes	IMLA	classification	clustering
Ipod	0.714	0.800	0.900
Trailer	1.000	1.000	0.961
Itunes	1.000	0.833	0.928
Laptop	1.000	1.000	0.933
Iphone	1.000	1.000	0.900
Fruit	1.000	1.000	1.000
Average	0.952	0.939	0.937

Table 4: Comparison of Precision values

	Recall		
Classes	IMLA	classification	clustering
Ipod	1.000	0.800	0.900
Trailer	1.000	1.000	1.000
Itunes	0.800	1.000	0.867
Laptop	1.000	1.000	0.933
Iphone	0.800	0.800	0.900
Fruit	1.000	1.000	1.000
Average	0.933	0.933	0.933

Table 5: Comparison of Recall Values

	F-Measure		
Classes	IMLA	classification	clustering
Ipod	0.833	0.800	0.900
Trailer	1.000	1.000	0.980
Itunes	0.889	0.909	0.896
laptop	1.000	1.000	0.933
iphone	0.889	0.889	0.900
fruit	1.000	1.000	1.000
Average	0.935	0.933	0.935

Table 6: Comparison of F-Measure Values

It can be seen from the results that for the dataset under consideration, the average precision is better in the integrated method when compared to the other two methods. The F-measure value, however, is better than the traditional classification method but slightly lower than the pure clustering.

3.2 Case II (30 Training, 80 test)

The same procedure as explained in Case I is carried out here. The difference is that now the size of the training set is 30 and that of the test set is 80. The confusion matrix for this case using IMLA is presented in Table 7. The comparative results for the three methods are presented in Tables 8, 9 and 10. In this case, the newly proposed integrated method completely outperforms the traditional methods. The values of both precision and recall and thereby that of F-measure are much better than the pure classification and clustering technique.

classified as >>	a	b	c	d	e	f
a = ipod	15	0	0	0	0	0
b = trailer	0	20	0	0	0	0
c = itunes	0	1	9	0	0	0
d = laptop	0	0	0	10	0	0
e = iphone	1	0	0	0	4	0
f = fruit	0	0	0	0	0	20

Table 7 Confusion Matrix for IMLA- Case II

	Precision		
classes	IMLA	classification	clustering
ipod	0.938	1.000	0.900
trailer	0.952	0.944	0.961
itunes	1.000	1.000	0.928
laptop	1.000	1.000	0.933
iphone	1.000	0.455	0.900
fruit	1.000	1.000	1.000
Average	0.981	0.899	0.937

Table 8: Comparison of Precision Values

	Recall		
classes	IMLA	classification	clustering
ipod	1.000	0.867	0.900
trailer	1.000	0.850	1.000
itunes	0.900	0.900	0.867
laptop	1.000	0.900	0.933
iphone	0.800	1.000	0.900
fruit	1.000	1.000	1.000
Averages	0.950	0.919	0.933

Table 9: Comparison of Recall Values

	F-measure		
classes	IMLA	classification	clustering
ipod	0.968	0.929	0.900
trailer	0.976	0.895	0.980
itunes	0.947	0.947	0.896
laptop	1.000	0.947	0.933
iphone	0.889	0.625	0.900
fruit	1.000	1.000	1.000
Average	0.963	0.890	0.934

Table 10 Comparison of F-Measure Values

3.3 CASE III (42 Training, 68 test)

The confusion matrix obtained after classification by the C4.5 decision tree classifier is shown in Table 11. The comparison of results for this case for the three different methods is presented in Tables 12, 13 and 14. In this case, the newly proposed integrated method exactly equals the performance of classification method and outperforms the clustering method.

classified as >>	a	b	c	d	e	f

a = ipod	12	0	1	0	0	0
b = trailer	0	8	0	0	0	0
c = itunes	0	0	8	0	0	0
d = laptop	1	0	0	2	0	0
e = iphone	0	0	0	0	18	0
f = fruit	0	0	2	0	0	16

Table. 11 The confusion Matrix for Case III

	Precision		
classes	IMLA	classification	clustering
ipod	0.923	0.923	0.900
trailer	1.000	1.000	0.961
itunes	0.727	0.727	0.928
laptop	1.000	1.000	0.933
iphone	1.000	1.000	0.900
fruit	1.000	1.000	1.000
Average	0.941	0.941	0.937

Table 12: Comparison of Precision Values

	Recall		
classes	IMLA	classification	clustering
ipod	0.923	0.923	0.900
trailer	0.889	0.889	1.000
itunes	1.000	1.000	0.867
laptop	1.000	1.000	0.933
iphone	0.667	0.667	0.900
fruit	1.000	1.000	1.000
Average	0.913	0.913	0.933

Table 13: Comparison of Recall Values

	F-Measure		
classes	IMLA	classification	clustering
ipod	0.923	0.923	0.900
trailer	0.941	0.941	0.980
itunes	0.842	0.842	0.896
laptop	1.000	1.000	0.933
iphone	0.800	0.800	0.900
fruit	1.000	1.000	1.000
Average	0.917	0.917	0.934

Table 14: Comparison of F-Measure Values

IV. CONCLUSIONS

A new approach which integrates classification and clustering, is presented in this paper. The approach has been applied to the problem of web page categorization. The accuracy of the results obtained by the newly proposed integrated method has been compared with the results obtained by the classical methods applied separately. From the results obtained it can be concluded that the new approach gives very encouraging results. It is also observed that when the size of the training data is smaller, the results are better. This could be due to the reason that clustering results are better when working with a small number of well defined instances. This also reduces the noise in the clusters thereby increasing the subsequent classification accuracy. When compared to the traditional classification it has the advantage of not requiring a human expert to supply examples. In the new approach, only a small number of instances need be clustered, which generally decreases the

noise. The major contribution of the proposed approach is that it completely automates the process of web page categorization by eliminating the need for a human expert at all stages.

REFERENCES

1. G. Salton and M.J. McGill. Introduction to Modern Information Retrieval. *McGraw- Hill Computer Science Series, New York: McGraw-Hill*, 1983.
2. G. Salton. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. *Reading, Mass.: Addison Wesley*, 1989.
3. B.V.Swathi and A.Govardhan. Find-k: A New Algorithm for Finding the k in Partitioning Clustering Algorithms. *International Journal of Computing Science and communication Technologies,*2(1) : 286-272, Aug 2009.
4. B.V.Swathi and A.Govardhan. A Modified Rough Set Reduct For Web Page Classification. *International Journal of Computer Applications in Engineering Technology and Sciences*, October 2009.
5. S. Dumais J. Platt, D. Heckerman, and M. Sahami. Inductive Learning Algorithms and Representations for Text Categorization. *Seventh Int'l Conf. Information and Knowledge Management,* pp.148-15, Nov. 1998.
6. H. Kargupta , I. Hamzaoglu, and B. Stafford. Distributed Data Mining Using an Agent Based Architecture. *Knowledge Discovery and Data Mining,* pp.211-214, 1997.
7. U.Y. Nahm and R.J. Mooney. Mutually Beneficial Integration of Data Mining and Information Extraction. *17th Nat'l Conf. Artificial Intelligence (AAAI-00),* pp.627-632, 2000.
8. Y. Yang J. Carbonell, R. Brown, T. Pierce, B. Archibald, and X. Liu. Learning Approaches for Detecting and Tracking News Events. *IEEE Intelligent Systems,* 14 (4): 32-43, 1999.
9. D. Freitag and A. McCallum. Information Extraction with HMMs and Shrinkage. *AAAI-99 Workshop Machine Learning for Information Extraction,* pp. 31-36, 1999.
10. T. Hofmann. The Cluster-Abstraction Model: Unsupervised Learning of Topic Hierarchies from Text Data. *16th Int'l Joint Conf. Artificial Intelligence (IJCAI-99),* pp.682-687, 1999.
11. T. Honkela S. Kaski, K. Lagus, and T. Kohonen. WEBSOM—Self-Organizing Maps of Document Collections. *WSOM '97, Workshop Self-Organizing Maps,* pp.310-315, June 1997.
12. W.W. Cohen. Learning to Classify English Text with ILP Methods. *Fifth Int'l Workshop Inductive Logic Programming*, pp. 3-24, 1995.
13. M. Junker M. Sintek, and M. Rinck. Learning for Text Categorization and Information Extraction with ILP. *First Workshop Learning Language in Logic, J. Cussens*, pp. 84-93, 1999.
14. S. Scott and S. Matwin. Feature Engineering for Text Classification. *16th Int'l Conf. Machine Learning (ICML-99*), pp. 379- 388, 1999.
15. S. Soderland. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34(1-3): 233-272, 1999.
16. K. Aas and L. Eikvil. Text Categorisation: A Survey. *Technical Report 941, Norwegian Computing Center*, June 1999.
17. G. Salton A. Wong, and C. Yang. A Vector Space Model for Automatic Indexing. *Comm. ACM,* 18, (11): 613-620, Nov. 1975.