ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 3, No. 2, 2008, *pp*. 083-089

Combined Feature Selection and classification – A novel approach for the categorization of web pages

K. Selvakuberan, M. Indradevi, Dr. R. Rajaram

Innovation Labs (Web 2.0), Tata Consultancy Services, M/s. Tata Consultancy Services Ltd.,
Chennai - 600096, Tamil Nadu, India

(*Received March 10, 2008, accepted April 22, 2008*)

Abstract. The World Wide Web is growing rapidly and the Internet users are still increasing day by day. Increasing with the number of users, the need for automatic classification techniques with good classification accuracy increases as search engines depend on previously classified web pages stored as classified directories to retrieve the relevant results. Machine learning techniques for automatic classification gains more interest as the classifier improves its performance with experience. In this paper we propose a method called Combined Feature Selection and Classification for effective categorization of web pages. Our experimental results show that our proposed approach improves the classification accuracy with the optimum number of attributes. We experimented with four machine learning classifiers (CV Parameter Selection, Logit Boost, Random Committee and VFI).Our results effectively improve the accuracy.

1. Introduction

At present, the number of Web-pages on World Wide Web is increasing significantly. There is an exponential increase in the amount of data available on the web recently. According to the number of pages available on the web is around 1 billion with almost another 1.5 million are being added daily. This enormous amount of data in addition to the interactive and content-rich nature of the web has made it very popular. Describing and organizing this vast amount of content is essential for realizing the web’s full potential as an information resource.(John M.Pierre 2001, Tom M.Mitchell 1999) However, these pages vary to a great extent in both the information content and quality. Moreover, the organization of these pages does not allow for easy search. So an efficient and accurate method for classifying this huge amount of data is very essential if the web is to be exploited to its full potential. This has been felt for a long time and many approaches have been tried to solve this problem. Web page classification techniques use concepts from many fields like Information filtering and retrieval, Artificial Intelligence, Text mining, Machine learning techniques and so on(Anagnostopoulos I et al.,2002,. Information filtering and retrieval techniques usually build either a thesauri or indices by analyzing a corpus of already classified texts with specific algorithms. Vector representation of the corpus texts may be used instead of building thesauri and indices. Natural Language parsing techniques may be used for the classification of web pages. Many researchers propose the use of text-mining techniques to do web mining/ web page classification. As the HTML pages are semi-structured documents containing tags, frames, etc., some preprocessing is required to be done in the web pages before applying text-mining techniques. Moreover applying text mining techniques for web page classification has the major drawback that it does not utilize the contextual features like URL, Links, Structure, META , TITLE tags, Tables, Frames and Visual layout of HTML pages which are very much useful for web page classification. In the late 90’s machine learning comes into picture. It is a fully automated process.(Yanmin Sun, Yang Wang, and Andrew K.C. Wong 2006)

Automatic classification of web pages is needed for the following reasons. (a) Large amount of information available in the internet makes it difficult for the human experts to classify them manually (b) The amount of Expertise needed is high (c) Web pages are dynamic and volatile in nature (e) More time and effort are required for classification. (f) Same type of classification scheme may not be applied to all pages (g) More experts needed for classification.(Hsin-Chang Yang, Chung-Hong Lee 2003)

The paper is organized as follows. Section 2 focuses on the Literature Survey. Section 3 presents what is meant by feature selection and their issues. Section 4 deals with our proposed work. Section 5 focuses on the Experimental results. Section 6 ends with the conclusion and section 7 describes the references.

2. Related work

Chandra Chekuri, Michael H. Goldwasser, Prabhakar Raghavan, Eli Upfal (1999) propose a new architecture for web search using automatic classification. They describe a search interface that combines context-free syntactic search with context-sensitive search guided by classification. The classification process is statistical, and is based on term-frequency analysis.

Arul Prakash Asirvatham and Kiranthi Kumar Ravi implemented a structure based categorization system to categorize the web page into three broad categories. (i) Information pages (ii) Research pages (iii) Personal Home pages. They have taken into account the images and the structure of the page for the automatic classification of web pages into these three categories.

Rachel Aires et al., use the stylistic features of web texts in Portuguese to classify the web apges based on their user needs. In their paper they had some seven needs. Based on the user satisfaction, they tested with 45 classifiers in WEKA workbench.(Ian and Witten)

3. Feature Selection

Web page classification needs a lot of preprocessing work because of the presence of hyperlinks and large number of HTML tags. It is estimated that 80% of the preprocessing is needed before the classification of web pages.

Feature extraction or selection is one of the most important pre-processing steps in pattern recognition or pattern classification, data mining, machine learning and so on. It is also an effective dimensionality reduction technique and an essential preprocessing method to remove noise features (Balaji et al., 2004). The basic idea of feature selection algorithms is searching through all possible combinations of features in the data to find which subset of features works best for prediction. The selection is done by reducing the number of features of the feature vectors, keeping the most meaningful discriminating ones, and removing the irrelevant or redundant ones.(Hsin-Chang Yang, Chung-Hong Lee 2003, Yiming Yan and Jan O. Pederson 1997). During generation and evaluation of subsets of features increasing feature brings disadvantages for classification problem.

3.1. Issues in Feature Selection

On one hand, feature increased gives difficulties to calculate, because the more data occupy amount of memory space and computerization time, on the other hand, a lot of features include certainly many correlation factors respectively, which results to information repeat and waste. It is necessary to take measures to decrease the feature dimension under not decreasing recognition effect; this is called the problems of feature optimum extraction or selection (Yiming Yan and Jan O. Pederson 1997). The characteristics of good features should be simple, moderate, less redundancy and unambiguous.(Shou-Bin Dong and Yi-Ming Yang 2002, Rudy Setiono, Huan Liu 1997)

4. Proposed approach

The goal of this paper is to find the best combination of feature selection techniques for web page categorization problem. It also overcomes the issues in feature selection. We also made a comparison of the learners. This process contains 3 stages: a) the extraction of representative features, to describe content – the initial set, b) the selection of the best features from initial set by applying another feature selection technique (minimizing the number of features and maximizing the discriminative information carried by them) and c) the training and classification using the resulting features in the different classifiers to determine the quality of features.

```
Algorithm combined Feature Selection and classification
{
   Initial Feature selection (set of keywords)
      {
        //The keywords may be more for each and every category. Apply term frequency approach.
      Term_frequency(keywords)
     {
       Calculate how many times the terms get repeated. It shows the frequency of each
```

```
            term and how relevance the term is used for classification
               return (relevant words);
          }
    }
    The output will be the "initial set of features"
    Final Feature Selection (initial set of features)
    {
      Use the elevators and search methods for finding "final set of features".
       The final set of features is considered as most relevant features.
    }
    Then perform classification using respective learners
    }
```

Our proposed method works like follows:

1. Initial Feature Selection

From the keywords of the particular category, apply the term-frequency approach.(The term-frequency approach calculates the number of occurrences of each term. The more relevant are chosen). It will result in "initial set of features".

2. Choose the nominal attribute from the initial set of features. The nominal attribute should be selected in such a way that selected attribute should coincide with the category in which we are classifying.

For ex: If we want to classify the web pages as "Course" category means then "iscourse{yes,no}" can be taken as the nominal attribute.

3. Use the nominal attribute as the evaluator combination with the search method performs feature selection again. This will result in the "final set of features". These final set of features are considered as the most relevant features for classifying the web page.

4. Use those final set of features for classification using respective learners.

}

4.1. Feature Selection phase

Feature selection is normally done by searching the space of attribute subsets, evaluating each one. This is achieved by combining attribute subset evaluator with a search method. In this paper we choose seven attribute evaluators with five search methods to find the best feature set.

For the feature selection phase, two objects must be set up: a feature evaluator and a search method. The evaluator determines what method is used to assign a worth to each subset of features. The search method determines what style of search is performed.

The feature selection can be done two ways: 1) using full training set (the worth of the feature subset is determined using the full set of training data), or 2) by cross-validation (the worth of the feature subset is determined by a process of cross-validation). In addition, the classifying time grow dramatically with the number of features, rendering the algorithm impractical for problems with a large number of features.

In practice, the choice of a learning scheme (the next phase) is usually far less important than coming up with a suitable set of features.

We experimented with several evaluators and search methods:

4.1.1 Evaluators:

- CfsSubsetEval - Evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them; subsets of features that are highly correlated with the class while having low inter-correlation are preferred.
- ConsistencySubsetEval - Evaluates the worth of a subset of features by the level of consistency in the class values when the training instances are projected onto the subset of features.
- PCA - Performs a principal components analysis and transformation of the data.
- Wrapper Subset Eval- Wrapper attributes subset evaluator.
- 4.1.2 Search methods:

- BestFirst - Searches the space of feature subsets by greedy hill-climbing augmented with a backtracking facility.
- GeneticSearch - Performs a search using the simple genetic algorithm
- Ranker - Ranks features by their individual evaluations. Use in conjunction with feature evaluators (ReliefF, GainRatio, Entropy etc).
- Exhaustive search –Performs an exhaustive search over all the features.
- Forward Selection –Performs selection of an attribute one by one.

4.2. Classification phase

For the classification phase we used four machine learning classifiers (CV Parameter Selection, Logit Boost, Random Committee, VFI). For some of our experiments we use WEKA workbench. (Ian and H.Witten)

5. Experimental Setup

For our experiment the database used is WEBKB data set and is downloaded from the UCI repository. It is a benchmarking dataset for machine learning problems. This is the university database having seven categories of web pages: Course, Project, Student, Faculty, Department, Staff and others. We select all the pages in the course category (930 pages) and non course category (66 pages). The initial set of features for "course" category is listed in Table 1.

Course, class, syllabus, handout, homework, cs, lecture, notes, slides, solution, problem, program, instructor, information, project, paper, guide, study, prelim, professional, activities, resume, publications, language, research, teaching, contact, projects, professor, interests, department, personal, office, advisor, home, page, links, phone, iscourse (yes,no)

Table 1: Initial set of features for the Course category

5.1. Results and analysis

The initial set of features is fed into the classification phase (CV Parameter Selection, Logit Boost, Random Committee, VFI). The results are shown in table 2.

S.No	CV Parameter Slection		Logit Boost		Random committee		VFI	
	CCI	Macro F	CCI	Macro F	CCI	Macro F	CCI	Macro F
1.	981	0.99	981	0.99	997	0.998	959	0.978

Table 2: Classification accuracy using Initial set of features

5.2. Combined Feature Selection results

5.2.1 Feature Selection phase

From the initial set of features, the feature selection technique is applied. The feature selection evaluators and their search methods and the list of features selected are shown in Table 3.

5.2.2 Results and Comparisons

The results for classification using final set of features are listed in Tables 5 and 6.

We compare our method with the Rachel Aires work. The classification accuracy obtained by him for the classifiers CV Parameter Selection, Logit Boost, Random Committee and VFI are given in table 5.

From the results we have analyzed that VFI achieves higher accuracy (0.979) with the number of features as 12 rather than the initial set of features(0.978 with 40 features).Logit Boost attains the maximum accuracy (0.99) with 32 number of features.CV Parameter Selection achieves maximum accuracy (0.99) with 30 features. (consistency Subset Eval +forward selection). For Random Committee it attains optimal accuracy (0.979) using 11 features(cfs subset eval +forward selection). By analyzing the tables 4, 5 and 6 it effectively shows our proposed method works better.

Method Name	**Search Name**	**No. Of features selected**	**Selected Features**
Principal Components	Ranker	36	course, class, syllabus, handout, homework, cs, lecture, notes, slides, solution, problem, program, instructor, information, project, paper, guide, study, prelim, professional, activities, resume, publications, language, research, teaching, contact, projects, professor, interests, department, personal, office, advisor, home, page
Consistency Subset Eval	Best First	32	course, class, syllabus, handout, homework, cs, lecture, notes, problem, program, instructor, office, information, project, paper, study, resume, publications, language, research, teaching, contact, projects, associate, professor, interests, department, personal, advisor, home, page, phone
	Exhaustive search	36	course, class, syllabus, handout, homework, cs, lecture, notes, slides, solution, problem, program, instructor, information, project, paper, guide, study, prelim, professional, activities, resume, publications, language, research, teaching, contact, projects, professor, interests, department, personal, office, advisor, home, page
	Forward selection	30	course, class, syllabus, homework, cs, lecture, notes, problem, program, instructor, office, information, project, paper, study, resume, publications, language, research, contact, projects, associate, professor, interests, department, personal, advisor, home, page, phone
Cfs Subset Eval	Genetic Search	12	course, class, syllabus, resume, publications, language, research, associate, interests, department, personal, advisor
	Forward Selection	11	course, class, syllabus, homework, instructor, resume, publications, research, interests, department, advisor
	Rank Search	10	course, class, syllabus, instructor, resume, publications, research, interests, department, advisor
	Best First	11	course, class, syllabus, homework, instructor, resume, publications, research, interests, department, advisor
	Exhaustive Search	35	course, class, syllabus, handout, homework, cs, lecture, notes, slides, solution, problem, program, instructor, office, information, project, paper, guide, study, prelim, professional, activities, resume, publications, language, research, teaching, contact, projects, associate, professor, interests, department, personal, advisor
Wrapper Subset Eval	Rank Search	2	interests, advisor

Table 3: Feature selection phase from the initial set of features

6. Conclusion and Future work

We have concluded that the features selected by the search methods such as best first, Rank search and Forward Selection with the evaluator cfs subset Eval yields better results. We conclude that applying more than one feature selection technique (Increemental feature selection) is essential for the effective performance.It is also found that Logit Boost and CV Parameter Selection performs little bit better than VFI and Random Committee. Our proposed method attains the optimal accuracy (0.987) as in case of initial set of features with the number of attributes as 11 (cfs subset eval +best first).As in case of VFI it attains 0.979 accuracy with the number of attributes as 12.(cfs subset eval +Genetic search). Depending on resources available we can choose any one of the feature selection technique like Wrapper Subset Eval with rank search for limited resources. As a future work we have to combine all these feature selection methods to improve the classification accuracy further.

S.no	**Name of the classifier**	**Classification accuracy (%)**
1.	CV Parameter Selection	13.7
2.	Logit Boost	55.22
3.	Random Committee	50.16
4.	VFI	47.8

Table 4 : Experimental results for Rachel Aires work

S,No	**Classifier name**	**Search name**	**Logit Boost**				**CV Parameter Selection**			
			Precision	Recall	CCI	Macro F	Precision	Recall	CCI	Macro F
1.	cfssubseteval	Bestfirst	0.981	0.996	978	0.988	0.975	0.989	966	0.982
2	Cfssubseteval	Forward selection	0.981	0.996	978	0.988	0.975	0.989	966	0.982
3	Cfssubseteval	Genetic Search	0.979	0.995	975	0.987	0.971	0.989	962	0.98
4	Cfssubseteval	Exhaustive search	0.981	0.996	978	0.988	0.977	0.987	966	0.982
5	cfssubseteval	Rank search	0.982	0.994	977	0.988	0.977	0.988	967	0.982
6	Wrapper subseteval	Rank search	0.961	0.987	951	0.974	0.961	0.987	951	0.974
7	Consistency subseteval	Exhaustive search	0.977	0.999	977	0.988	0.982	0.998	981	0.99
8	Consistency subseteval	Bestfirst	0.982	0.999	982	0.99	0.978	0.987	967	0.982
9	Consistency subseteval	Forward selection	0.979	0.995	975	0.987	0.982	0.998	981	0.99
10	Principal components	Ranker	0.977	0.999	977	0.988	0.982	0.998	981	0.99

Table 5: Classification using Final Set of features

S,No	**Classifier name**	**Search name**	**Random committee**				**VFI**			
			Precision	Recall	CCI	Macro F	Precision	Recall	CCI	Macro F
1.	cfssubseteval	Bestfirst	0.979	0.995	975	0.987	0.955	0.947	947	0.971
2	Cfssubseteval	Forward selection	0.979	0.995	975	0.987	0.955	0.947	947	0.971
3	Cfssubseteval	Genetic Search	0.979	0.995	975	0.987	0.99	0.961	955	0.979
4	Cfssubseteval	Exhaustive search	0.959	0.991	952	0.975	0.995	0.928	929	0.961
5	cfssubseteval	Rank search	0.98	0.992	974	0.986	0.995	0.942	942	0.968
6	Wrapper subseteval	Rank search	0.961	0.987	951	0.974	0.961	0.987	951	0.974
7	Consistency subseteval	Exhaustive search	0.963	0.992	957	0.976	0.995	0.928	929	0.961
8	Consistency subseteval	Bestfirst	0.968	0.992	962	0.98	0.994	0.954	952	0.974
9	Consistency subseteval	Forward selection	0.967	0.991	960	0.978	0.961	0.987	951	0.974
10	Principal components	Ranker	0.963	0.992	957	0.976	0.995	0.928	929	0.961

Table 6: Classification using Final Set of features

7. References

[1] Anagnostopoulos I., Kouzas G. Anagnostopoulos C., Vergados D., Papaleonidopoulos I., Generalis A., Loumos V. and Kayafas E., (2002) "Automatic Web Site Classification in a Large Repository under Information Filtering and Retrieval Techniques", IEEE MELECON 2002, May 7-9,2002, Cairo, EGYPT. Pp 279 – 283

[2] Anagnostopoulos, C. Anagnostopoulos, V. Loumos and E. Kayafas (2004) "Classifying Web pages employing a probabilistic neural network", IEE Proc.-Softw., Vol. 151, No. 3, June 2004,pp 139 – 150

[3] Arul Prakash Asirvatham,and Kiranthi Kumar Ravi, "Web Page Categorization based on document structure"

[4] Balaji Krishnapuram, Alexander J. Hartemink, Lawrence Carin and Mario A.T. Figueiredo (2004) 'A Bayesian Approach to Joint Feature Selection and Classifier Design' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No. 9, Pp1105 – 1111

[5] Chandra Chekuri, Michael H. Goldwasser, Prabhakar Raghavan, Eli Upfal (1999) "Web Search Using Automatic Classification", Sixth International World Wide Web Conference Poster Presentations

[6] Hsin-Chang Yang, Chung-Hong Lee (2003) "A Text Mining Approach on Automatic Generation of Web Directories and Hierarchies", Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03), 2003 IEEE

[7] Ian H.Witten and Eibe Franh "Data Mining-Practical Machine Learning Tools and techniques" –) book (second edition)

[8] John M.Pierre (2001) "On the Automated Classification of Web Sites", Link¨oping Electronic Articles in Computer and Information Science, Vol. 6(2001): nr 0. http://www.ep.liu.se/ea/cis/2001/000/. February 4, 2001

[9] Rachel Aires, Aline Manfrin, Sandra Aluisio, Diana Santos (2002) 'Which classification algorithm works best with stylistic features of Portuguese in order to classify web texts according to the user's needs'

[10] Rudy Setiono, Huan Liu(1997) 'Feature Selection via Discretization' IEEE Transactions on Knowledge and Data Engineering, Vol 9,Issue 4, 642-645

[11] Shou-Bin Dong and Yi-Ming Yang, " Hierarchical Web Image Classification By Multi-Level Features", Proceedings of the first international conference on Machine Learning and Cybernetics, Beijing, 4-5 November 2002, pp 663 – 668

[12] Susan Dumais, Hao Chen, "Hierarchical Classification of Web Content", SIGIR 2000 , ACM, pp 256 –263

[13] Tom M.Mitchell (1999) 'The Role of Unlabeled data in Supervised Learning' Proceedings of the Sixth International Colloquium on Cognitive Science

[14] Yanmin Sun, Yang Wang, and Andrew K.C. Wong, " Boosting an Associative Classifier", IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 7, July 2006, Pp 988 – 992

[15] Yiming Yan and Jan O. Pederson (1997) 'Comparative Study of feature selection in Text Categorization', Proceedings on Fourteenth International Conference on Machine Learning (ICML'97) pp 412-420