Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 55 (2013) 12825-12829

ISSN: 2229-712X

Classifying Web Pages using Support Vector Machine

Kavitha S[1] and Vijaya MS[2]

[1]PSGR Krishnammal College for Women, Coimbatore, India.

[2]GR Govindarajulu School of Applied Computer Technology, Coimbatore, India.

ARTICLE INFO

Article history:

Received: 10 December 2012;
Received in revised form:
26 January 2013;
Accepted: 1 February 2013;

Keywords

Categorization,
Knowledge,
Support vector machine,
Learning,
Web pages.

ABSTRACT

Web is an enormous warehouse of knowledge and frequent hyperlinks. Web also serves a wide diversity of user communities and worldwide information service centers. Every minutes the knowledge in web page upwards swiftly. Web page is used to transmit the knowledge to web users. Such voluminous size of web makes a complexity of web information retrieval, web content filtering, web usage mining and web structure mining. Hence, it is essential to perform proper categorization of web pages. This paper formulates the web page categorization problem as multi classification task and provides an appropriate solution using support vector machine. The classification model is generated by learning the features that have been extracted from different category of HTML structures and URLs of the web pages. The experimental results of support vector machine with various kernels have been evaluated and observed that accuracy of web page categorization model with RBF kernel (98.5%) performs well than linear and polynomial kernel.

Introduction and Background

With the rapid development of World Wide Web, the knowledge in web page grows explosively. Due to its swarm, the information overload and information unavailability are the tribulations in every web search engine. In addition the web pages are not consistently structured. Hence the web page classification is a vital task in every web search engine.

Web page categorization is a significant in numerous information retrieval tasks such as retrieval of scientific papers, e-books and digital library from the web. In web usage mining the web page classification consumes to build customized web services to individual web users. Web structure mining is concerned with discovering the model underlying the link structure on the web page, for example to envisage the links between terrorists in social networks. In web page filters such as e-mail filter, content filter, web content filtering determines the content that is to be blocked in a web page. Thus a web page categorization helps to reach competent web information retrieval, web content filtering, web structure mining and web usage mining.

A variety of rules based and machine learning techniques are currently in use for web page categorization. In [1], the various supervised learning techniques namely, decision tree, k-nearest neighbor, one r, multilayer perceptron and rbf kernel are adopted for web page categorization.

Web page categorization has been implemented using three feature selection techniques like filter model, wrapper model and hybrid model along with the page rank algorithm in order to decrease the redundant features in the web page [2].

In [3], the authors have used different features that are extracted from HTML source code and URL with a compound of HTML and URL along with its information sibling pages, for web page categorization. Naive Bayes algorithm is used as a classifier and it is compared with semi-supervised algorithm such as co-training and expectation maximization and inductive logic programming have been applied to increase the performance in weak learner for web page categorization in [4].

The research work offered in this paper syndicates the features of web pages declared in [1] [2] [3] [4] and identifies few innovative features which can contribute more in the ideal classification of web pages. The features such as strings between slashes and dots in the href attributes of all anchor tags, strings between underscores and minus symbols in the href attributes of all anchor tags, defined in HTML source code of web pages are additionally used. These features have been used to incorporate reference mechanisms available in web pages such as tables, footnotes and bibliographies. They also provide the interconnection between linked web pages, it consists of text, images, video and other multimedia contents.

The proposed web page categorization model also employs novel URL features such as, substring between underscores and minus symbols of URLs, substring between two different symbols of URLs, apart from those used in the existing work. These features have been used to provide additional resources to the URL. Hence these features are very much essential and guarantee to contribute more in web page categorization.

In most of the existing work, web page categorization was carried out to classify the web pages of similar domain. Here the web pages of different domains like arts, business, culture, education, entertainment, health and wellness have been considered for categorization.

This paper elucidates the implementation of support vector machine for classifying the web pages of six divergent domains. The features are extracted from HTML structures and URLs of a set of web pages in different categories. Feature extraction and the experiments carried out are described in rest of this paper.

Proposed Web Page Classification Model

The proposed web page categorization model decreases the convolutions in web mining. The different categories of web pages are composed arbitrarily from the search engines. The acquired web pages are preprocessed and features are extracted from HTML structure and URL using feature extraction methods. The training data set with instances associated to six domains such as arts, business, culture, education,

Tele:
E-mail addresses: s.kavithamphil2011@gmail.com, msvijaya@grgsact.com

entertainment, health and wellness are developed. The proposed web page categorization model employs support vector machine for learning the classification models. Finally the trained models are evaluated and used to classify the unknown category of web pages. The proposed web page categorization model is given in Fig. 1.

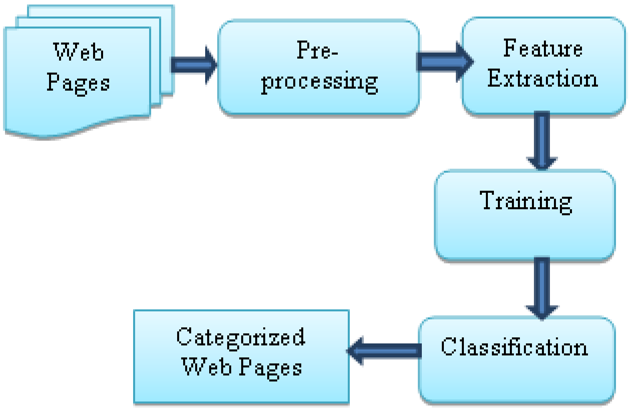


Fig. 1 Proposed web page categorization model

Preprocessing

Preprocessing can improve the computational effectiveness and value of the data. It is used to eliminate stop words, inadequate html tags and redundant symbols from the HTML structure of web pages.

Feature Extraction

Feature extraction plays an essential role in data mining. It is used to improve the classification effectiveness and computational efficiency. Two kinds of features i.e. HTML features and URL features are extracted. HTML features are obtained from the HTML structure of web pages using term frequency and structure oriented weighting technique. Term frequency is used for computing the weight of a term in a web page, which is nothing but the number of times the term occurs in a web page. Structure oriented weighting technique is used to assign most important terms that are more suitable for representing additional elements in HTML structure of the web pages. Features related to the URL are taken from the URL of the corresponding web page. Categorical values from 1 to 6 are assigned to features with reverence to arts, business, culture, education, entertainment, health and wellness respectively.

1) HTML Features

The HTML fundamentals defined in HTML structure namely, title tag, meta attribute tags, paragraphs, headings and links describes the contents of a web page. So the features pertaining to these elements are extracted to form the training dataset for web page categorization. The HTML features are scheduled below.

Title Tag

The title tag is obligatory in HTML structure, it states the title of the document. This tag plays an essential role in search engine optimization. The syntax of this tag is <Title> Title of a webpage </Title>. For example this tag <Title>Indian Cultural Websites</Title> returns the value of a feature as 'Culture' and the categorical value 3 is assigned to this feature.

Meta Description Tag

The Meta description tag is an HTML tag and it describes the contents of the web page. It can be used in the head section of a web page. It is used after the title tag and before the meta keyword tag. This tag provides the next importance to the search engine optimization and it consists of snippet information to the HTML structure. The format of this tag is <Meta name="description" content= "description of the web page">.

For example this tag <Meta name= "description" content="health web page may contain general health, men's health and women's health"> returns 'health' and the categorical value 6 is assigned to this feature.

META Keyword Tag

The Meta keywords are the list of terms and it is used to highlight knowledge to the web page. The keywords are alienated by comma. The syntax of this tag is <Meta name="keywords" content="list of terms">.

For example the meta keyword tag <Meta name="keywords" content="Best sports websites, Best sports teams"> returns 'sports' and the categorical value of this feature is 5.

Paragraph Tag

The goal of Paragraph tag in any web page is to illuminate disparate concepts associated to one topic. Paragraph tag is one consideration of an HTML element in which to divide one paragraph with more paragraphs on a web page. Paragraph element is basically represented with English alphabet p. The contents between the start tag and end tag forms a Paragraph. The format of Paragraph tag is <p> Contents of the paragraphs </p>. The value of this feature is determined as before.

Heading Tags

Heading tags are indicators and it defines the section headings and sub headings to the web page. This tag can be used to validate document structure and organization. It also represents to generate outlines and table of contents in a web page. Section headings at different levels namely, h1 tag represents the highest level heading, h2 tag is the next level down, h3 is a level below that, and so on to h6.

H1 tag represents the page title in a web page and h2 tag is used to provide the all major headings in a web page. For example these heading tags <h1>Special sports teams</h1> and <h2>Best sports</h2> returns 'sports' and the categorical values of these features are 5.

Base URL in HTML Structure

Base URL specifies the all relative href and other links in a web page. It represents an external resource to the web page. HTML structure permits only one base element for each web page. The base URL has attributes, but it does not have the contents of a web page.

This Base URL tag, http://dir.yahoo.com/Health/Health_Care/Universal_Health_Care returns 'Health' and so its categorical value is 6.

Anchor Tag

An anchor element is called an anchor because the web designers can use it to anchor a URL to some text in a web page. When users view the web pages in a system, they can click the text to activate the link and visit the page that the URL has the link. In HTML structure an anchor tag may be either the origin or target of a hyperlink. In the href attribute the anchor becomes a hyperlink to the web page. Alternatively with the name or id attribute set, the element becomes a target. The URL can link to this target via a fragment identifier.

In anchor tag the title represents a brief knowledge about the link. The format of the anchor tag is <a href = "URL" title="additional information"> linked web page contents </a>. The anchor tag features are enumerated below.

Strings between slashes

This feature specifies the link between slashes in the href attribute of all anchor tags in the HTML structure. For example this href attribute http://dir.yahoo.com/health/mental_health returns ‘health’ and the categorical value 6 is assigned to this feature.

Strings between minus symbols

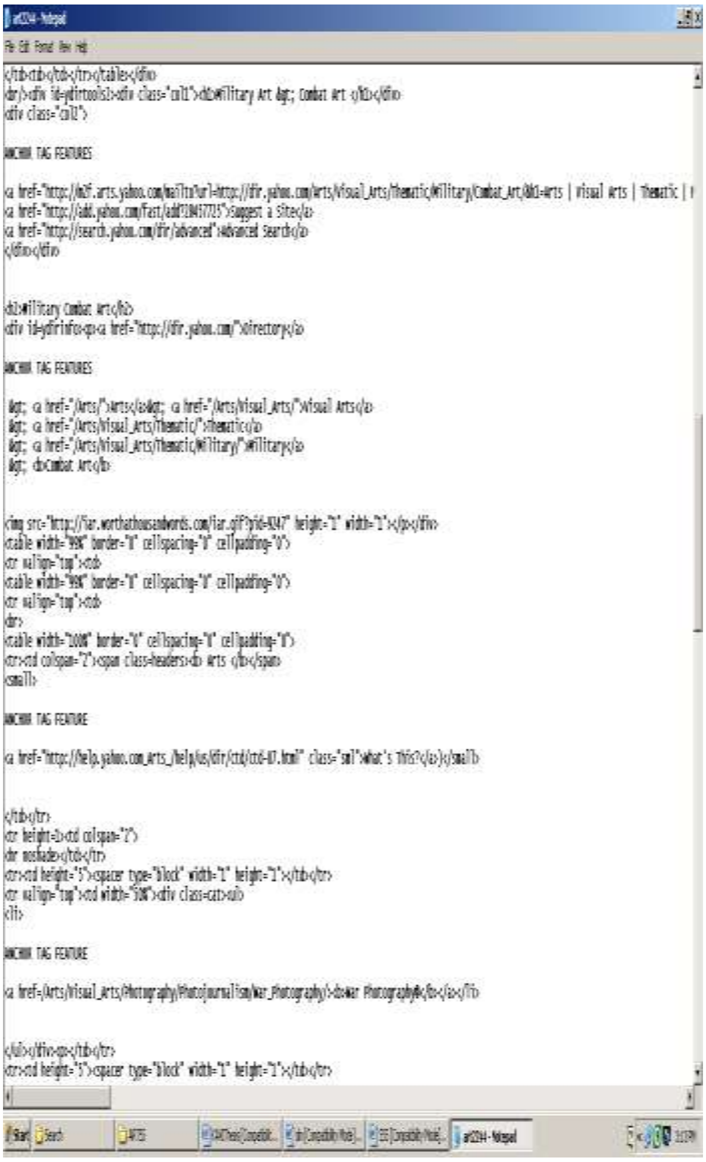
It represents the link between minus symbols in the href attribute of all anchor tags in the HTML structure. The value of this feature is determined as before.

Strings between underscores

It can be used in the states the link between underscores in the URL of all anchor tags in HTML. For example this attribute http://dir.yahoo.com/Arts/Visual_Arts/Forensic_Art returns ‘Art’ and its categorical value is 1.

Strings between dots

This feature stipulates the link between dots in the href attribute of all anchor tags. The value of this feature is determined as before. Some of the anchor tag features are shown in Fig. 2.



```
</td><td></td></tr></table></div>
<br/><div id=ydirtools1><div class="col1"><h2>Military Art &gt; Combat Art </h2></div>
<div class="col2">

ANCHOR TAG FEATURES

<a href="http://m2f.arts.yahoo.com/mailto?url=http://dir.yahoo.com/Arts/Visual_Arts/Thematic/Military/Combat_Art/&t=Arts | Visual Arts | Thematic | 
<a href="http://add.yahoo.com/fast/add?18457725">Suggest a Site</a>
<a href="http://search.yahoo.com/dir/advanced">Advanced Search</a>
</div></div>


<h2>Military Combat Art</h2>
<div id=ydirinfo><p><a href="http://dir.yahoo.com/">Directory</a>

ANCHOR TAG FEATURES

 &gt; <a href="/Arts/">Arts</a>&gt; <a href="/Arts/Visual_Arts/">Visual Arts</a>
 &gt; <a href="/Arts/Visual_Arts/Thematic/">Thematic</a>
 &gt; <a href="/Arts/Visual_Arts/Thematic/Military/">Military</a>
 &gt; <b>Combat Art</b>


<img src="http://far.worthathousandwords.com/far.gif?pid=9347" height="1" width="1"></p></div>
<table width="99%" border="0" cellspacing="0" cellpadding="0">
<tr valign="top"><td>
<table width="99%" border="0" cellspacing="0" cellpadding="0">
<tr valign="top"><td>
<br>
<table width="100%" border="0" cellspacing="0" cellpadding="0">
<tr><td colspan="2"><span class=headers><b> Arts </b></span>
<small>

ANCHOR TAG FEATURE

<a href="http://help.yahoo.com_Arts_/help/us/dir/ctd/ctd-07.html" class="sml">What's This?</a>)</small>


</td></tr>
<tr height=1><td colspan="2">
<hr noshade></td></tr>
<tr><td height="5"><spacer type="block" width="1" height="1"></td></tr>
<tr valign="top"><td width="50%"><div class=cat><ul>
<li>

ANCHOR TAG FEATURE

<a href=/Arts/Visual_Arts/Photography/Photojournalism/War_Photography/><b>War Photography</b></a></li>


</ul></div><p></td></tr>
<tr><td height="5"><spacer type="block" width="1" height="1"></td></tr>
```

Fig. 2 Anchor tag features

2) URL Features

Uniform Resource Locator is an exact character string that constitutes a reference to an internet resource and it is a type of uniform resource identifier. The URL features used in this research work are enumerated below.

Substring between dots

This feature specifies the substring between dots in the URL. For example, this URL http://rakings.business.com, returns ‘business’ and so its categorical value is 2

Substring between slashes

It states the substring between slashes in the URL. For example, this http://dir.yahoo.com/Arts/ returns ‘Arts’ and so it's categorical value is 1.

Substring between underscores

This feature used to specify the substring between underscores in the URL. For example this URL http://yahoo.com/Health/General_Health returns ‘Health’ and the categorical value of this feature is 6.

Substring between minus symbols

This feature represents the substring between minus symbols in the URL. This URL http://www.magfact.com/sports/best-sports-websites/ returns ‘sports’ and so its categorical value is 5.

Substring between two different symbols

This feature specifies the substring between two different symbols in the URL. This URL http://dir.yahoo.com/Society_and_Culture/Cultures_and_Groups returns ‘Culture’ and the categorical value is 3.

This a list of 16 features is extracted from the HTML structures and the URLs of the web pages. The categorical values for these 16 features are obtained through a PHP code. This list of 16 categorical values can form a feature vector. Since support vector machine finds a nonlinear decision function in the input space by mapping the data into a higher dimensional feature space, the computational complexity does not depend on the dimension of the feature vector. So the training dataset with 16 dimensions is used for training SVM. The feature vectors corresponding to web pages of six different categories are generated and the training data set is developed.

Support Vector Machine

Support vector machine is one of the most actively developed classification methodology in data mining and machine learning. SVM represents a new approach to supervised pattern classification that has been successfully applied to a wide range of pattern recognition problems. SVM is very suitable for working accurately and efficiently with high dimensional feature space and it provides most prominent properties such as the margin maximization and nonlinear classification via kernel tricks and confirmed to be effective in various real world applications. A classification task usually involves separating data into training and testing sets. All instances in the training set contain one target value as the class labels and several attributes as the features or observed variables. The target of SVM is to create a model based on the training data that predicts the target values of the test data given only the test data attributes.

The machine is offered with a set of training examples, (xi,yi) where the xi is the real world data instances and the yi are the labels signifying which class the instance belongs to. The two class pattern recognition problem, yi = +1 or yi = -1. A training example (xi,yi) is called positive if yi = +1, otherwise it is negative. SVM constructs a hyperplane that separates two classes and tries to attain maximum separation between the classes. Separating the classes by a large margin minimizes a bound on the projected simplification fault. The simplest representation of SVM is a maximal margin classifier, constructs a linear separator or optimal hyperplane is given by w T x - γ = 0 involving two classes of examples. The gratis parameters are a vector of weights w, which it is orthogonal to the hyperplane and a threshold rate γ. These parameters are obtained by solving the subsequent optimization complexity using Lagrangian duality.

$$\text{Minimize } = \frac{1}{2}\|W\|^2$$

$$\text{Subject to } D_{ii}\left(W^T X_i - \gamma\right) \geq 1, i = 1,.....,l.$$

Where D_{ii} corresponds to class labels +1 and –1. The instances with non-null weights are called support vectors. In the occurrence of outliers and wrongly classified training examples it may be useful to permit some training errors in order to avoid over fitting. A vector of slack variables ξi that compute the quantity of destruction of the constraints is introduced and the optimization problem referred to as soft margin is given below.

$$\underset{W,\gamma}{Minimize} = c\sum_{i=1}^{l}\varepsilon_i + \frac{1}{2}\|W\|^2$$

$$\text{Subject to } \underset{\varepsilon_i \geq 0}{D_{ii}} \left(W^T X_i - \gamma\right) + y_i \geq 1, i = 1,.....,l.$$

In this formulation the contribution to the objective function of margin maximization and training errors can be unbiased through the use of regularization parameter c. The next decision rule is used to correctly visualize the class of new instance with a minimum error.

$$f(X) = \text{sgn}\left[W^T x - \gamma\right]$$

The advantage of the dual formulation is that it permits an efficient learning of non–linear SVM separators, by introducing kernel functions. Specifically, the kernel functions compute a dot product between two vectors that have been mapped non-linearly into a high dimensional feature space. Since there is no necessity to complete this mapping obviously, the training is still sufficient while the dimension of the real feature space can be very elevated or even endless. The parameters are obtained by solving the following non-linear SVM formulation (in matrix form),

$$\text{Minimize } L_D(u) = \frac{1}{2} u^T Qu - e^T u$$

$$d^T u = 0 , 0 \leq u \leq Ce$$

Where ***Q = DKD***and ***K*** - the Kernel Matrix. The kernel function K (AAT) (polynomial or Gaussian) is used to construct hyperplane in the feature space which separates two modules linearly by performing computations in input space. The decision function is given by,

$$f(X) = \text{sgn}\left(K\left(x, x_i^T\right) * u - \gamma\right)$$

Where, u - the Lagrangian multipliers.

When the number of class labels is more than two, the binary SVM can be extended to the multi class SVM. One of the indirect methods for multiclass SVM is the one versus rest method. For each class a binary SVM classifier is constructed, selected data points of that class against the rest. Thus in case of N classes, N binary SVM classifiers are built. During testing, each classifier yields a decisive value for the test data point and the classifier with the highest positive decision value assigns its label to the data position. The comparison between the decision values formed by dissimilar SVMs is still valid because the training parameters and the dataset remain the same.

Experiment and Results

The web page categorization model is implemented using SVM^{light}. The data set used in the experiment is developed by collecting the web pages randomly from the search engines. Web pages relevant to six categories such as arts, business, culture, education, entertainment, health and wellness are downloaded from yahoo, google etc. From each category 50 web pages have been downloaded and totally 300 web pages are used in experimenting web page categorization.

The features narrating distinguishing characteristics of a web page are extracted and the feature vector of size 16 is generated for all the 300 web pages as described earlier. Class labels 1 to 6 are assigned to all the instances relevant to arts, business, culture, education, entertainment, health and wellness respectively and the training data set is developed.

The dataset is trained in SVM with linear, polynomial and RBF kernels with disparate parameter value for C where C is the regularization parameter. In linear kernel the value for t is given as 0. In polynomial kernel the value t is given as 1 and the value for d is assigned as 1 and 2. In RBF kernel the value of t is assigned as 2 and values for gamma is assigned as 1 and 2. The performance of trained models is evaluated using 10-fold cross validation for its classification accuracy. The classification accuracy is dignified as the ratio of the number of correctly classified instances in the test dataset and the total number of test cases. The performances of SVM classifiers are evaluated based on the classification accuracy and learning time. Regularization parameter C assigns different values in the range of 1 to 5 and found that the model performs better and attain a stable state for the value C = 3. The result of the classification model based on SVM with linear kernel is shown in Table 1.

Table 1. Results of SVM with Linear Kernel

Linear SVM	**C=1**	**C=2**	**C=3**	**C=4**	**C=5**
Accuracy (%)	70	75	76.7	73.4	75
Time taken (Sec)	0.00	0.00	0.00	0.01	0.01
Number of support vectors	8	9	10	11	10

The results of the classification model based on SVM with polynomial kernel with parameters d and C are shown in Table 2.

Table 2. Results of SVM with Polynomial Kernel

Parameters	**C=1**		**C=2**		**C=3**	
d	***1***	***2***	***1***	***2***	***1***	***2***
Accuracy (%)	76.6	73.4	73.4	93.3	75	96.7
Time taken (Sec)	0.02	0.04	0.02	0.03	0.02	0.04
Number of support vectors	8	58	10	48	11	55

The results of the classification model based on SVM with RBF kernel with parameters C and g are shown in Table 3.

Table 3. Results of SVM with RBF Kernel

Parameters	**C=1**		**C=2**		**C=3**	
g	***1***	***2***	***1***	***2***	***1***	***2***
Accuracy (%)	97	96	97.8	98.1	98	98.5
Time taken (Sec)	0.01	0.01	0.01	0.01	0.01	0.01
Number of support vectors	1	1	1	1	1	1

The average and comparative performance of various kernels in SVM is given in Table 4.

Table 4. Average Performance of Three Kernels

Kernels	**Classification accuracy (%)**	**Learning time (Sec)**
Linear	76.7	0.00
Polynomial	96.7	0.04
RBF	98.5	0.01

The average and comparative performance of each kernel in SVM is given in the Fig. 3 and Fig. 4.

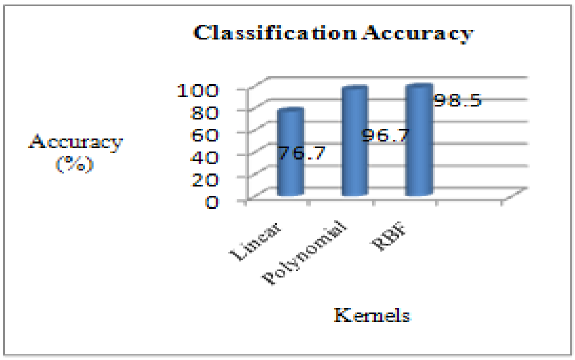


Fig. 3 Classification accuracy of various kernels

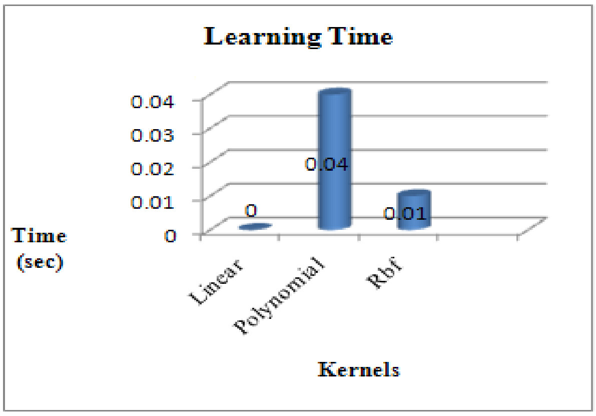


Fig. 4 Learning time of various kernels

From the above comparative analysis, it is found that the classification accuracy shown by SVM with RBF kernel (98.5%) is higher than the linear (76.7%) and polynomial kernel (96.7%). The time taken to construct the model using SVM with polynomial kernel is more, than linear and RBF kernels. Also the number of support vectors is less in case of RBF kernel, which proves that the proposed web page categorization model based on SVM with RBF kernel, is more proficient. Hence it is concluded that SVM with RBF kernel based web page categorization model performs well than SVM with linear and polynomial kernel.

Conclusions and Future Work

This paper formulates the web page categorization problem as the multi classification task. The proposed web page categorization model is implemented using support vector machine. Features are extracted from HTML structure and URL of different categories of web pages and learned using SVM under different parameter settings. The outcome of the experiments indicates that the SVM with RBF kernel classifies the web pages more accurately than the other models. As a scope for future work, the web page classification can be extended to the web community mining.

References

[1] Alamelu Mangai J and Santhosh Kumar V, "A Novel Approach for Web Page Classification using Optimum features", in proceedings of International Journal of Computer Science and Network Security, Vol.11, No.5, May 2011.

[2] SiniShibu, Aishwarya Vishwakarma and Niket Bhargava, "A Combination Approach for Web Page Classification using Page Rank and Feature Selection Technique ", in proceedings of International Journal of Computer Theory and Engineering,Vol.2, No.6, Dec 2010.

[3] Sara Meshkizadeh and Amir Mason Rahmani, "Web Page Classification based on Compound of Using HTML Features and URL Features and Features of Sibling Pages", in proceedings of the International Journal of Advancements in Computing Technology, Vol.2, No.4, Oct 2010.

[4] Nuanwan Soonthornphisaj and Boonserm Kijsirikul, "Combining ILP With Semi-supervised Learning for Web Page Categorization", in proceedings of International Journal of Information and Mathematical Sciences, Vol.1, No.4, 2005.

[5] Santhana Lakshmi V and Vijaya M.S, "The SVM Based Interactive Tool for Predicting Phishing Websites", in proceedings of the International Journal of Computer Science and Information Security, Vol.9 No.10, Oct 2011.

[6] Rekha Jain and Purohit G.N, "Page Ranking Algorithms for Web Mining", in proceedings of International Journal of Computer Application, Vol.13, Jan 2011.

[7] Ting S.L, W.H.IP, Albert H.C.T, "Is Naïve Bayes a Good Classifier for Document Classification", in proceedings of International Journal of Software Engineering and its Applications", Vol.5, No.3, July 2011

[8] Zhihua Wei, Hongyun Zhang_, Zhifei Zhang, Wen Li, Duoqian Miao, "A Naïve Bayesian Multi-label Classification Algorithm with Application to Visualize Text Search Results", in proceedings of International Journal of Advanced Intelligence", Vol.3, No.2, pp. 173-188, July 2011

[9] BinduMadhuri CH, AnandChandulal J, Ramya K and Phanidra M, "Analysis of Users Web Navigation Behavior using GRPA With Variable Length Markov Chains", in proceedings of International Journal of Data Mining and Knowledge Management Process ", Vol.1, No.2, March 2011

[10] Pooja Sharma and Pawan Bhadana, "Weighted Page Content Rank for Ordering Web Search Result ", in proceedings of International Journal of Engineering Science and Technology, Vol.2, 2010.

[11] Wongkot Sriurai, Phayung Meesad and Choochart Haruechaiyasak, "Hierarchical Web page Classification based on a Topic Model and Neighboring Pages Integration", in proceedings of International Journal of Computer Science and Information Security, Vol.7, No.2, 2010.

[12] Selvakuberan K, Indradevi M and Rajaram R, "Combined Feature Selection and Classification-A Noval Approach for the Categorization of Web Pages", in Proceedings of International Journal of Information and Computing Science, Vol.3, No.2, Pp.083-089, 2008.

[13] Brown EN, Kass RE and Mitra PP "Multiple neural spike train data analysis: state-of-the-art and future challenges", Nature Neuroscience, 7 (5): 456–61, 2004.

[14] Arabib and Michael A, "The Handbook of Brain Theory and Neural Networks.

[15] Russell and Ingrid, "Neural Networks Module", 2012.

[16] Yogendra kumar jain and Sandeep wadekar "Classification based Retrieval Methods to Enhance Information Discovery on the Web", in proceedings of International Journal of Managing Information Technology, Vol. 3, No. 1, Feb 2011.

[17] Shiqun Yin, Yuhui Qiu, Chengwen, Zhong, Jifu Zhou, "Study of Web Information Extraction and Classification Method", IEEE International Conference on Wireless Communications, Networking and Mobile Computing, Wicom, PP.5548-5552, 2007.

[18] Lilac A.E Al-safadi, "Auto Classification for Search Intelligence", in proceedings of World Academy of Science, Engineering and Technology, 2009.

[19] Caruana R and Niculescu Mizila, "An empirical Comparison of Supervised Learning Algorithms", in proceedings of 23rd International Conference on Machine Learning, 2006.