

# Two-Phase Web Site Classification Based on Hidden Markov Tree Models

YongHong Tian<sup>1,2</sup>, TieJun Huang<sup>1,2</sup>, Wen Gao<sup>1,2,3</sup>, Jun Cheng<sup>2</sup>, PingBo Kang<sup>2</sup>

<sup>1</sup>(Digital Media Lab, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China);

<sup>2</sup>(Research Center for Digital Media, Graduate School of Chinese Academy of Sciences, Beijing 100039, China)

<sup>3</sup>(Dept. of Computer Science, Harbin Institute of Technology, Harbin 1500001, China)

E-mail: {yhtian,tjhuang,wgao,jcheng,pbkang}@jdl.ac.cn

## Abstract

*With the exponential growth of both the amount and diversity of the information that the web encompasses, automatic classification of topic-specific web sites is highly desirable. In this paper we propose a novel approach for web site classification based on the content, structure and context information of web sites. In our approach, the site structure is represented as a two-layered tree in which each page is modeled as a DOM (Document Object Model) tree and a site tree is used to hierarchically link all pages within the site. Two context models are presented to capture the topic dependences in the site. Then the Hidden Markov Tree (HMT) model is utilized as the statistical model of the site tree and the DOM tree, and an HMT-based classifier is presented for their classification. Moreover, for reducing the download size of web sites but still keeping high classification accuracy, an entropy-based approach is introduced to dynamically prune the site trees. On these bases, we employ the two-phase classification system for classifying web sites through a fine-to-coarse recursion. The experiments show our approach is able to offer high accuracy and efficient process performance.*

## 1 Introduction

In recent years web page or hypertext document categorization methods such as Bayesian classification [1], decision-tree induction [2] and SVM [3,4] have been widely studied, however, the classification of complete

web sites has not yet been investigated except in [5,6,7]. With the exponential growth of both the amount and diversity of the information that the web encompasses, automatic classification of topic-specific web sites is highly desirable because sites have some obvious advantages in content coherence [7], comparatively low dimension of analysis space and content stability [6]. Hence, this paper will focus on how to design a both effective and efficient web site classification algorithm. The goal of our work is to develop an intelligent topic-specific web resource analysis tool, *iExpert*, for Chinese Science Digital Library Project.

Existing web site classification algorithms [5,6,7] all treated whole pages as atomic indivisible nodes with no internal structure. But as a matter of fact, pages are more complex and have more topics, many of which are much *noisy* from the perspective of the query, such as in banners, navigation panels, and advertisements, etc [9]. Hence pages should be further divided into some finer-grained *logic snippets* [9], such as DOM (Document Object Model [8]) nodes. And the DOM nodes should be treated as the fundamental analysis units in web site classification.

Meanwhile, the content organization structure and local contexts of web sites are also important for web site classification. Typically, there are three representation models of web sites in web site classification algorithms: the superpage [5], topic vector [6], and the tree [6] or graph [7] representation. Comparatively, though difficult to be constructed, the tree representation is more suitable

for capturing the essence of the link structure. Hence this paper will propose a two-layered tree model of web sites based on the structure and context information.

Since many classification algorithms must be performed offline, the efficiency of web site classification crucially depends on the download amount of web pages. The experiment result shown in Figure 1 enlightens us that some sampling algorithm must be introduced into the web site classification procedure for downloading only a small part of a web site but still achieving high accuracy.

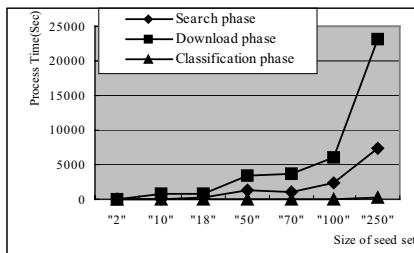


Figure 1. Comparison in process time between the search, download and classification phases on the baseline system given different sizes of seed sets and a fixed three levels of pages to be downloaded.

Base on the above considerations, this paper proposes a novel approach for web site classification based on the content, structure and context information of web sites. Our approach represents the site structure as a two-layered tree in which each page is modeled as a DOM tree and a site tree is used to hierarchically link all pages within the site. Two context models are presented to characterize the topic dependence relationships in the site, and the HMT model is utilized as the statistical model of the site tree and the DOM tree. In a fine-to-coarse recursion, the HMT-based classifier is applied iteratively to DOM trees and site tree to obtain the final classification result of the web site. Furthermore, for further improving the classification accuracy, we will also introduce an entropy- based approach to dynamically prune the web site trees and a text-based DOWN-TOP denoising procedure to remove the noisy DOM nodes or pages. The system design of our approach is illustrated in Figure 2.

We evaluate our approach on physics web site mining tasks given different seed sets. The seed sites are also used as training samples. The baseline system is based on the



Figure 2. System design of our approach

superpage classification approach with a fixed download depth. Experiments show that our approach has average 12.7% improvement in classification accuracy and 38.2% reduction in process time over the baseline system.

Compared with the previous web site classification algorithms presented in [5,6,7], the two-layered tree representation and two-phase HMT-based classification are the main features of our approach. Experiments also show our approach obtains marked improvements both in classification accuracy and process performance over these existing algorithms.

This paper is organized as follows. Section 2 proposes the two-layered tree model of web sites. Section 3 simply reviews the HMT model and then presents the HMT-based classification algorithm. In section 4 we discuss the text-based denoising procedure and entropy-base pruning strategy. Thereafter, experiments and their results will be described. Finally, Section 6 concludes this paper.

## 2 The Two-Layered Tree Model of Web Sites

The representation model of web sites directly affects the efficiency of the web site classification algorithm. The superpage method just represents a web site as a set of terms or keywords [5]. As discussed in [6], this method performs poorly and is only suitable for constructing baseline systems. Analogously, the topic vector approach [6] that represents a web site as a topic vector (where each topic is defined as a keyword vector), is essentially a two-phase keyword-based classification. On the other hand, the tree-based representation model [6] can effectually utilize the semantic structure of sites and local contexts, and more importantly, transform the sampling size problem of web sites to the pruning problem of site trees. However, the previous tree-based web site classification method [6] still employed the keyword-based text classifiers for the pre-classification of pages and thus the noises within pages would affect the final classification accuracy of web sites. Therefore, this paper

proposes a two-layered tree representation model for web sites. The model is based on the following propositions:

**Proposition 1 (Tree Structure Assumption):** *The structure of most web sites is more hierarchic than network-like [6]. Furthermore, each HTML/XML page can be represented as a DOM tree [8].*

Thus, we define the web site as follows:

**Definition 1 (Site Structure Model):** A web site can be represented as a site tree  $T(P, E)$ , where  $P = \{p_1, \dots, p_n\}$ , root  $p_1$  is the starting page of the site, and  $\forall p_i \in P$  is a HTML/XML page within the site. Furthermore,  $p_i$  can be represented as a DOM tree, i.e.,  $p_i = DOM_i(DN, DE)$ . A link between  $p_i$  and  $p_j$  is represented by the directed edge  $(p_i, p_j) \in E$ , where  $p_i$  is the parent node of  $p_j$  and  $p_j$  is one of the children of  $p_i$ .

To build the web site tree, a *breadth-first search* will be performed. In our application, we only sample the pages located “below” the starting page. For example, if the URL of the starting page is <http://phys.cts.nthu.edu.tw/en/index.php>, we only sample the pages shared the base URL <http://phys.cts.nthu.edu.tw/en/>.

In hyperlink environment, links contain high-quality semantic clues to a page’s topic, and such semantic information can help achieve even better accuracy than that possible with pure keyword-based classification [10]. In this paper we refer to the semantic information surrounding links in a page as the *context* of the page, and generalize this concept to the classification of DOM nodes. Therefore, according to Markov Random Field (MRF) theory [11], we have the following proposition:

**Proposition 2 (Context Assumption):** *In web site classification, context is treated as node set and environment information topically related to the analysis node. Context information is helpful to improve the classification accuracy of analysis nodes [10,11,15].*

Thus we define the following two kinds of context models used in web site classification:

**Definition 2 (Page Context Model PC):** Each web page or hypertext is topically related to its in-linked and out-linked pages [1, 11]. In site tree  $T(P, E)$ ,  $p_i$ ’s parent node  $p_j$  is one of its in-link page,  $p_i$ ’s children

$(p_{c_1}, \dots, p_{c_{n_i}})$  are its out-link pages, then the page context

model of  $p_i$  is  $PC(p_i) = \{p_j, p_{c_1}, \dots, p_{c_{n_i}}\}$ , where

$$P(C_{p_i} | \{C_{p'} | p' \neq p_i\}) = P\{C_{p_i} | C_{p_j}, C_{p_{c_1}}, \dots, C_{p_{c_{n_i}}}\} \quad (1)$$

**Definition 3 (DOM Context Model DC):** Each DOM node is topically related to its parent and children nodes in the DOM tree. Namely, the DOM context model of  $DN_i$  is  $DC(DN_i) = \{p(DN_i), DN_{c_1}, \dots, DN_{c_{n_i}}\}$ , where

$$P(C_{DN_i} | \{C_{DN'} | DN' \neq DN_i\}) = P\{C_{DN_i} | C_{p(DN_i)}, C_{DN_{c_1}}, \dots, C_{DN_{c_{n_i}}}\} \quad (2)$$

The above site structure model and context models constitute our web site representation model. The model represents the structure of web sites as a two-layered tree  $T(\{DOM_i(DN, DE)\}, E)$ , and captures the topic dependencies (NOT link structure) between nodes according to context models  $PCs$  and  $DCs$ , as shown in Figure 3.

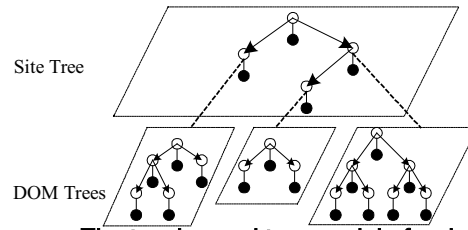


Figure 3. **The two-layered tree model of web sites. Black and white nodes represent nodes and their topical class attributes, respectively.**

Based on this representation, we can apply the two-phase classification architecture for web site classification. The two-phase classification exploits the topic dependencies between nodes within DOM trees and site trees to obtain the final classification results of web sites. Meanwhile, this representation model enables us to purify the content of web sites at both the DOM node and page levels so as to further reduce the affect of noise information on the classification results. The following sections will further discuss the details.

### 3 The HMT-Based Classification Algorithm



construct HMT-based classifiers. Here we don't discuss it further.

Thus we may apply the HMT-based classifiers to the web site classification. The classification procedures are shown in Figure 5 when the whole page or the DOM node is treated respectively as the atomic analysis node.

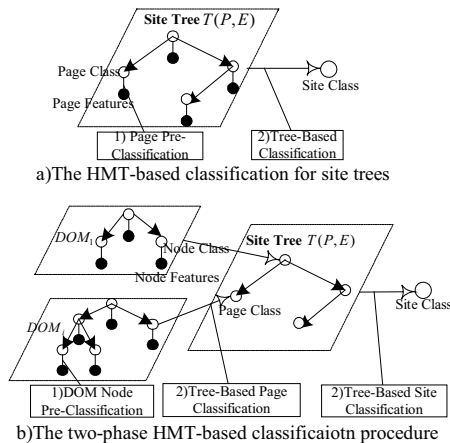


Figure 5. Applying the HMT-based classifiers to the web site classification.

#### 4 Denoising and Pruning

As discussed above, to obtain high classification accuracy in the web site classification, two tasks should be performed, i.e., denoising and pruning.

The task of denoising is to remove the DOM nodes or pages that are irrelevant to the query topics or cannot be identified by text-based classifiers, including animated introductions and frames, banners, navigation panels, and advertisements, etc. Hence, it is natural to utilize text-based pruning method at the DOM node and page levels. In this application, we use the DOWN-TOP procedure to purifying the content of web sites, i.e., text-based denoising method is performed in the DOM node level at first, and if a majority of DOM nodes within a page is marked for removing, then the page is removed in whole. In our experiments, the thesaurus-based method is used in the text-based denoising procedures, which determines the pertinence of a DOM node to a given topic by analyzing the occurrence frequency of topic-specific keywords and terms in that node. Experiments show the thesaurus-based

text denoising method can effectively reduce the noises within sites and thus improve the classification accuracy.

On the other hand, the pruning process for reducing the sampling size of web sites is more complex. The literature [6] exploited the variance of the conditional probability over the set of all web site classes to measure the importance of a path for site classification and then proposed a pruning algorithm based the variance and the path length. To capture data structure beyond second order (variance) statistics, in this paper we employ the *relative entropy* or *Kullback Leibler distance* [18] to model the 'distance' between the distribution embodied by the original model and that of the pruned model. Besides the tree structure assumption, our pruning approach is based on the following propositions:

**Proposition 3 (Assumption on pruning necessity):** *In web site classification, download of a remote web page is much more expensive than in-memory operations [6].*

This proposition has been verified by our experiment shown in Figure 1. It not only shows the importance of sampling in web site classification, but also enlightens us that we might reduce dramatically the download time by increasing somewhat local in-memory operations so as to optimize the total process time.

**Proposition 4 (Assumption on sample size):** *Web site classification need to download web pages within the site. However, after the downloaded pages are more than a fixed quantity, to download more pages further cannot improve the classification accuracy.*

It is not clear how to measure how many pages are sufficient to web site classification in order to keep a comparatively high accuracy. Intuitively, there are cases where a whole subtree does not show any clear class membership at all [6], hence features extracted from these pages cannot be helpful to improve any classification accuracy. In this paper we use Kullback Leibler distance to measure whether adding a page to web site tree will result in a reduction in the uncertainty of classification results or not. Therefore, we propose the following dynamic pruning strategy for site trees (See Figure 6):

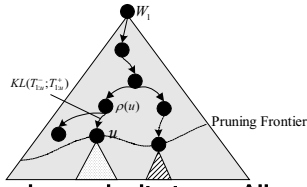


Figure 6. Pruning web site trees. All pages below the pruning frontier will not be downloaded.

Let  $u$  be the current downloading page,  $T_{1u}^-$  and  $T_{1u}^+$  denote the web site trees before and after downloading  $u$ , and  $P(C_i | T_{1u}^-, \lambda_p)$  and  $P(C_i | T_{1u}^+, \lambda_p)$  be the likelihood functions corresponding to  $T_{1u}^-$  and  $T_{1u}^+$  given the HMT model  $\lambda_p$ . Now we see if adding  $u$  to the web site tree will result in a reduction in the uncertainty of classification results, if so we add  $u$  to the site tree, and continue until no new page to be downloaded. The KL distance from  $T_{1u}^-$  to  $T_{1u}^+$  is given by

$$KL(T_{1u}^-; T_{1u}^+) = \sum_i P(C_i | T_{1u}^-, \lambda_p) \log \frac{P(C_i | T_{1u}^-, \lambda_p)}{P(C_i | T_{1u}^+, \lambda_p)} \quad (7)$$

Then the dynamic pruning strategy can be defined:

If  $(KL(T_{1u}^-; T_{1u}^+) \geq \text{depth}(T_{1u}^+) \cdot \Delta)$  then add  $u$  to the web site tree, else suppress the growth of the tree to node  $u$ .

Where  $\text{depth}(T_{1u}^+)$  is the depth of the site tree  $T_{1u}^+$ , and  $\Delta$  is the convergence parameter used in the HMT training of site trees. Similar to the pruning strategy presented in [6], this pruning approach becomes less sensitive with increasing  $\text{depth}(T_{1u}^+)$ . Our experiment will show the entropy-based approach can improve classification accuracy by downloading only a small part of a web site.

## 5 Experiments and Results

Currently, we evaluate our approach on physics web site mining tasks given different seed sets. We use *web site mining* to denote the task of seeking out a list of web sites that it either considers the most authoritative for a given topic,

or has the same topics with a given seed sites, and then grouping these sites under the predefined topic categories. The seed sites are meantime used as training samples. Currently, the sizes of seed sets in which we run all experiments are 2, 10, 18, 50, 70, 100, 250, respectively (Total 528 physics web sites are available). They have been downloaded to local server completely, and labeled by domain experts according to the *Physics Subject Classification*, which composes of 10 classes and 71 subclasses. It should be noted that the little distinguish-ability between some classes in the class hierarchy increases the difficulty of classification tasks. A *Physics Subject Thesaurus* with 9181 terms is also used for text-based denoising. The baseline system is based on bilingual kernel-weighted KNN classifier, using super-page classification approach with a fixed download depth of web sites. A hyperlink analysis program and a web focused-crawler are used for both baseline system and our HMT-based classification algorithms. All experiments are run in the following environments: 800MHz CPU, 256M RAM and shared 2M LAN bandwidth. And the evaluation metrics are *classification accuracy* and *process time*.

Fig. 7 shows the comparison of classification accuracy between different classifiers given the above seed sizes. Besides the baseline system and the two-phase HMT-based classifier, we also employed the HMT-based site tree classifier (as depicted by Figure 5a), and the 0-order Markov Tree classifier (as presented in [6]) to classify the new discovering web sites. Not surprisingly, although performed poorly in small size of training set, the two-phase HMT-based classifier had the best accuracy, with nearly 12.7% improvement over the baseline super-page approach. We also noticed that since the two-phase HMT-based classifier carried out denoising operations at both the DOM node and the page levels, they clearly outperformed than the HMT-based site tree classifier and the 0-order Markov Tree classifier that only denoised at the page level by about 7% and 2.6% respectively. This conclusion confirms that the noise information in pages and web sites is one of the main factors to cause the comparatively low accuracy for the web classification.

On the other hand, the 0-order Markov Tree classifier provided only an accuracy of about 69.3%, which was 4.4% more than the HMT-based site tree classifier. It means that the 0-order Markov Tree classifier has better performance because of fewer model parameters though they utilize the same site representation model. We also noticed that in our experiments the accuracy of the Markov Tree classifier was much less than 87% described in [6]. A possible reason is that in our class hierarchy the little distinguishability between some classes more easily causes the classification errors.

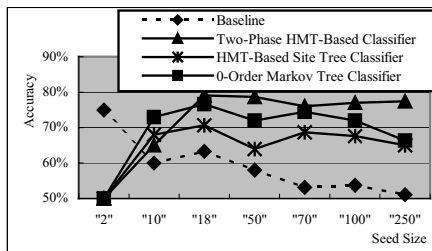


Figure 7. Comparison of the accuracy between the baseline superpage classifier, 0-Markov tree classifier, HMT-based site tree classifier and Two-phase HMT-based classifier given different sizes of seed set

We also compared the process time performance between the baseline system and the two-phase HMT-based classifier, and found that on the average, the two-phase classifier approach had saved 46.7% (See Figure 8) download time but spent more 28.7% classification time compared with the baseline system, which always downloaded a fixed three levels of pages from web sites. Totally, the two-phase HMT-based classifier saved 38.2% process time in the whole experiment. This result confirms the proposition 3, i.e., we can obtain dramatic reduction of total process time at the cost of increasing somewhat local in-memory operations. Furthermore, comparison of the HMT-based classifiers with the pruning step and the fixed download depth showed that pruning a web site tree would yield about 6% accuracy improvement while did not increase dramatically the downloaded data (See Figure 9). This is because the pruning strategy has purposefully imposed on somewhat

controls on the sampling process. These results show the entropy-based pruning algorithm is efficient.

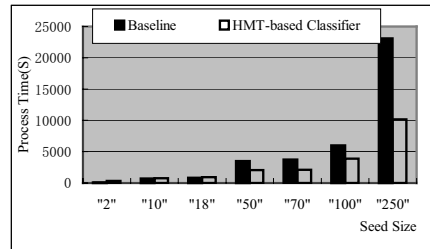


Figure 8. Comparison of the performance between the baseline and HMT-based classifier on download phase.

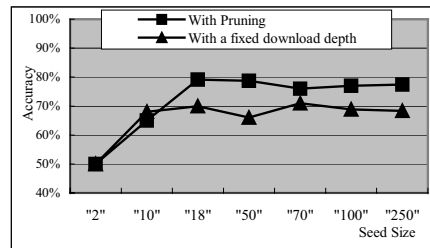


Figure 9. Comparison of the classification accuracy between with and without using entropy-based pruning method.

To sum up, the experiments show the two-phase HMT-based web site classification algorithm is able to offer high accuracy and efficient process performance. Comparatively, it has the following two deficiencies: Firstly, due to few features extracted from DOM nodes and lack of local context, the poor raw classification results of DOM nodes would affect the final classification accuracy; Secondly, the performance should be further improved when limited training samples are available.

## 6 Conclusion and Future Work

In this paper, we investigated the content organization structure of web sites and two context models used in web site classification, and proposed a two-layered tree representation model for web sites. Furthermore, this paper exploited the Hidden Markov Tree model to capture the statistical topic dependencies between nodes in site trees and DOM trees, and presented an HMT-based classification algorithm. To remove the irrelevant DOM

nodes or pages, a DOWN-TOP text-based denoising procedure was presented; to reduce the download size, an entropy-base dynamic pruning strategy was introduced. The experiments demonstrated our algorithm was effective and efficient.

In ongoing work we are seeking a further improvement of our web site classification algorithm. We need to investigate the more comprehensive context models to refine the poor raw classification results of DOM nodes so as to obtain both reliable and accurate classification.

## 7 Acknowledgement

This work was supported by Chinese Science Digital Library under Grant No CSDL 2002-18. The Conference Participation is Supported by Nokia Bridging the World Program

## 8 References

- [1] YiMing Yang, Sean Slattery, and Rayid Ghani. "A Study of Approaches to Hypertext Categorization." *Journal of Intelligent Information Systems*, 2002, Vol.18,pp. 219-241
- [2] Wen-Chen Hu. "WebClass: Web Document Classification Using Modified Decision Trees", In *Proc. of the Fifth International Conference on Computer Science and Informatics*, Atlantic City, NJ, USA, 2000
- [3] Susan Dumais and Hao Chen. "Hierarchical classification of Web content". In *Proc. of SIGIR-00*, Athens, Greece, 2000,pp. 256-263.
- [4] Aixin Sun, Ee-Peng Lim, Wee-Keong Ng. "Web Classification Using Support Vector Machine." In *Proc. of the fourth international workshop on Web information and data management*, McLean, Virginia, USA, 2002,pp. 96-99
- [5] John M. Pierre. "On the Automated Classification of Web Sites". *Computer and Information Sciences*, Vol. 6, 2001
- [6] Martin Ester, Hans-Peter Kriegel, Matthias Schubert. "Web Site Mining: A new way to spot Competitors, Customers and Suppliers in the World Wide Web", In *Proc. of SIGKDD02* Edmonton, Alberta, Canada, 2002,pp. 248-259
- [7] Loren Terveen, Will Hill, and Brian Amento. "Constructing, Organizing, and Visualizing Collections of Topically Related Web Resources". *ACM Trans. On Computer-Human Interaction*, 1999, pp. 67-94.
- [8] Johnny Stenback, Philippe Le Hégarret, Arnaud Le Hors: Document Object Model (DOM) Level 2 HTML Specification (Version 1.0), W3C Tech Report, <http://www.w3.org/TR/2003/REC-DOM-Level-2-HTML-2-0030109>, 2003
- [9] Chakrabarti, S., Joshi M., and Tawde V. "Enhanced Topic Distillation using Text, Markup Tags, and Hyperlinks", In *Proc. of the ACM SIGIR '01*, New Orleans, Louisiana, USA, 2001, pp. 208-216
- [10] Jiawei Han, Kevin Chen-Chuan Chang, "Data Mining for Web Intelligence", *IEEE Computer*, 35(11), Nov. 2002, pp. 64-70.
- [11] Soumen Chakrabarti, Byron Dom, Piotr Indyk, "Enhanced Hypertext Categorization Using hyperlinks", In *Proc. of SIGMOD '98* Seattle, Washington, 1998, pp. 307-318
- [12] M.S. Crouse, R.D. Nowak, and R.G. Baraniuk. "Wavelet-Based Statistical Signal Processing using Hidden Markov Models". *IEEE Trans on Signal Processing*, Vol. 46, 1998, pp. 886-902
- [13] M. Diligenti, P. Frasconi, and M. Gori. "Image Document Categorization Using Hidden Tree Markov Models and Structured Representation" In Singh, S, Murshed, N. & Kropatsch, W. (Eds.) *Advances in Pattern recognition - ICAPR 2001*. Lecture Notes in Computer Science, 2001,pp. 147-156.
- [14] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian Tree-Structured Image Modeling Using Wavelet-Domain Hidden Markov Models". *IEEE Trans. On Image Processing*, Vol. 10, No. 7, Jul 2001, pp. 31-44.
- [15] J. Li, R. M. Gray, "Context-based Multiscale Classification of Document Images Using Wavelet Coefficient Distributions", *IEEE Trans. on Image Processing*, vol. 9, no. 9, Sep 2000, pp. 1604-1616.
- [16] Gotoh, Y. Hochberg, M. and Silverman, H. "Efficient Training Algorithms for HMMs Using Incremental Estimation". *IEEE Trans. on Speech and Audio Processing*, 1998, Vol.6, No.6, pp. 539-548.
- [17] M. Diligenti and M. Gori and M. Maggini and F. Scarselli, Classification of HTML documents by Hidden Tree-Markov Models, In *Proc. of the International Conference on Document Analysis and Recognition (ICDAR)*, Seattle, WA (USA), 2001, pp. 849-853
- [18] Minh N. Do. "Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Model", *IEEE Signal Processing Letters*, Apr. 2003, pp. 115-118