Tree-Based Method for Classifying Websites Using Extended Hidden Markov Models

Majid Yazdani, Milad Eftekhar, and Hassan Abolhassani

Web Intelligence Laboratory, Computer Engineering Department,
Sharif University of Technology, Tehran, Iran
{yazdani,eftekhar}@ce.sharif.edu
abolhassani@sharif.edu

Abstract. One important problem proposed recently in the field of web mining is website classification problem. The complexity together with the necessity to have accurate and fast algorithms yield to many attempts in this field, but there is a long way to solve these problems efficiently, yet. The importance of the problem encouraged us to work on a new approach as a solution. We use the content of web pages together with the link structure between them to improve the accuracy of results. In this work we use Naïve-bayes models for each predefined webpage class and an extended version of Hidden Markov Model is used as website class models. A few sample websites are adopted as seeds to calculate models' parameters. For classifying the websites we represent them with tree structures and we modify the Viterbi algorithm to evaluate the probability of generating these tree structures by every website model. Because of the large amount of pages in a website, we use a sampling technique that not only reduces the running time of the algorithm but also improves the accuracy of the classification process. At the end of this paper, we provide some experimental results which show the performance of our algorithm compared to the previous ones.

Key words: Website classification, Extended Hidden Markov Model, Extended Viterbi algorithm, Naïve-Bayes approach, Class models.

1 Introduction

With the dramatically increasing number of sites and the huge size of the web which is in the order of hundreds of terabytes [5] and also with the wide variety of user groups with their different interests, probing for sites of specific interests in order to solve users' problems is really difficult. On the other hand, almost predominant section of the information which exists in the web is not practical for many users and this portion might interfere the results which are retrieved by users' queries. It is apparent that searching in the tiny relevant portion can provide us better information and lead us to more interesting sites and places on a specific topic.

At this time, there are a few directory services like DMOZ [3] and Yahoo [11] which provide us by useful information in several topics. However, as they

are constructed, managed and updated manually, most of the time they have incomplete old information. Not only webpages changes fast, but also linkage information and access records are updated day by day. These quick changes together with the extensive amount of information in the web necessitate automated subject-specific website classification.

This paper proposes an effective new method for website classification. Here, we use the content of pages together with the link structure of them to obtain more accuracy and better results in classification. Also among different models for representing websites, we choose tree structure for its efficiency. Tree model is useful because it displays the link structure of a given website clearly.

Before we begin to talk about ways of website classification, it is better to explain the problem more formally. Given a set of site classes C and a new website S consisting of a set of pages P, the task of website classification is to determine the element of C which best categorizes the site S based on a set of examples of preclassified websites. In other words, the task is to find a class C that website S is more likely to be its member.

The remaining of this paper is organized as follows. Related works are discussed in Section 2. Section 3 describes the models which we use for representing websites and website classes. In Section 4, we explain our method for classifying websites together with the extended version of Viterbi algorithm. Section 5 is about learning and sampling techniques. A performance study is reported in Section 6. Finally, Section 7 concludes the paper and discusses future works. A completer version of this paper is available through authors' homepages.

2 Related Works

Text classification has been an active area of research for many years. A significant number of these methods have been applied to classification of web pages but there was no special attention to hyperlinks. Apparently, a collection of web pages with a specific hyperlink structure conveys more information than a collection of text documents. A robust statistical model and a relaxation labeling technique are presented in [2] to use the class label and text of neighboring in addition to text of the page for hypertext and web page categorization. Categorization by context is proposed in [1], which instead of depending on a document alone, extracts useful information for classifying the document from the context of the page in which the hyperlink to the document exists. Empirical evidence is provided in [9] which shows that good web-page summaries generated by human editors can indeed improve the performance of web-page classification algorithms.

On the other hand, website classification has not been researched widely; the basic approach is superpage-based system that is proposed in [8]. Pierre discussed several issues related to automated text classification of Web sites, and described a superpage-based system for automatically classifying Web sites into industrial categories. In this work, we just generate a single feature vector counting the frequency of terms over all HTML-pages of the whole site, i.e. we represent a

web site as a single "superpage". Two new more natural and more expressive representation of web sites has been introduced in [6] in which, every webpage is assigned a topic out of a predefined set of topics. In the first one, Feature Vector of Topic Frequencies, each considered topic defines a dimension of the feature space. For each topic, the feature values represent the number of pages within the site having that particular topic. In the second one, Website Tree, the website is represented with a labeled tree and the k-order Markov tree classifier is employed for site categorization. The main difference between our work and this work is that in the latter, the topic (class) of each page is independently identified with a classifier and without considering other pages in the website tree, and then the topic of pages tree is used to compute the website class, but this independent topic identification will lower the accuracy of responses. Whereas in our work we calculate the website class without explicitly assigning a topic to each page and the topics of pages will be hidden to us. In [7] a website is represented with a set of feature vectors of terms. By choosing this representation, effort spent on deriving topic frequency vectors will be avoided. kNN-classification is employed to classify a given website according to training database of websites with known classes. In [10] the website structure is represented by a two layered tree: a DOM tree for each page and a page tree for the website. In [4] a new approach is presented for classifying multipage documents by naïve bayes HMM. However, to the best of our knowledge, there is no work on extending HMM for classifying data represented as tree.

3 Modeling Websites and Classes

Before modeling websites and website classes, defining some terminology is necessary.

Definition 1. *(The set of webSite class Label : SL) SL is the set of labels which can be assigned to websites as class labels, in other words, members of SL are category domains.*

Definition 2. *(The set of webPage class Labels : PL) For each label sl in SL, there is a set of known labels which can be assigned to individual pages of a website in that specified category. This labels can be seen as pages' topics.*

There are many choices to use as page class models like Naïve-Bayesian approach and Hidden Markov models. We prefer the former due to its simplicity. In this paper Naïve-Bayes model is adopted for modeling webpage classes. For modeling website classes, we extend Hidden Markov Model in order to satisfy the criteria of content and link structure within pages.

Definition 3. *(webSite Class Model : SCM) For each category domain $sl \in SL$, the model of sl is a directed graph $G(V, E)$ in which V, G's vertex set, is a set of webpage class models and for every two states $v_i, v_j \in V$ there is two directed edges $\{(i.j), (j, i)\} \in E$ that show the probability of moving from one to another. Also there is a loop for every state which shows the probability of transition between two pages of that class.*

In this model, the state-transition probability matrix A is an $n * n$ matrix in which n is the cardinality of SL and $a_{ij} = P(c_{t+1} = j|c_t = i),\ 0 \leq i, j \leq n$, which shows the probability of transition from state i to state j. Also the emission probability e is: $e_j(p) = P(p_t = p|c_t = j)$. $e_j(p)$ represents the probability of generating page p by webpage class model j.

As a result, we have a hybrid of Naïve-Bayes HMM model for all of websites classes. As we mentioned before, it was possible to model website classes by a hierarchial HMM, if HMM was used instead of Naïve-Bayes model for describing webpage classes. It is important to note that the input of website class models are websites which are modeled by trees. We describe website models in a more formal way, as follows.

Definition 4. *(webSite Model : SM) For each website ws, the ws's model is a page tree $T_{ws} = (V, E)$ in which V is a subset of ws's pages. The root of T_{ws} is the homepage of the website and there is a directed edge between two vertices $v_i, v_j \in V$ if and only if the corresponding page p_j is a child of p_i. We crawl the website by Breath-First-Search technique and ignore the links to previously visited pages in the time of model construction.*

4 Website Classification

In previous sections, we described a model for each website class. We assume SCM_i is the model of website class C_i. Also we consider that we want to classify website ws. To find the category of ws, we calculate $P(SCM_i|ws)$ for $1 \leq i \leq n$. By Considering Bayes rule, we can determine the class of ws as

$$C_{map} = argmax_i P(SCM_i|ws) = argmax_i P(SCM_i)P(ws|SCM_i)$$

$P(ws)$ is constant for all classes and we neglect $P(SCM_i)$. It can be considered later or even it can be equal for all classes if we use a fair distribution of websites over different classes. Therefore $C_{map} = argmax_i P(ws|SCM_i)$.

To find the probability of $P(ws|SCM_i)$ and also to solve the webpage classification problem, we should extend the Viterbi algorithm for our models.

4.1 Extended Viterbi Algorithm

Using dynamic programming techniques and therefore Viterbi algorithm ideas seems reasonable for solving these probabilities. However as we mentioned before, the inputs of our models are trees of pages, while Viterbi algorithm is provided to calculate the maximum probability of generating sequences. Therefore, we should modify the traditional Viterbi algorithm. Before introducing the new approach, it is necessary to present theorem 1 which is discussed in [10].

Theorem 1. *The probability of each node is only dependent to its parent and children. More formally, for every node n with p as its parent and $q_1, ...q_n$ as its children, if PL represents the webpage label set then*

$$P(PL_n|\{PL_k|k \in V\}) = P(PL_n|\{PL_p, PL_{q_1}, ..., PL_{q_n}\}) \quad (1)$$

Here, we propose a novel method in which the probability of a node and its siblings is computed simultaneously to avoid any possible inconsistency. Possible inconsistencies together with an illustrating example is provided in the completer version of the paper. Thus, in the nth level of the tree, we calculate the probability of the children of an $n-1$th level's node by equation 2.

Algorithm 1 *(Extended Viterbi Algorithm) For Classifying an indicated website ws which is modeled by a tree structure T_{ws} against n different classes $C_1, \ldots, C_n$ which are modeled by $SCM_1, \ldots, SCM_n$, if p is a node in level $n-1$ of T_{ws} and it has children $q_1, \ldots, q_n$ then*

$$P[(q_1, \ldots, q_n) \leftarrow (pl_{s_1}, \ldots, pls_n)] = \prod_{i=1}^{n} e_{pl_{s_i}}(q_i) * \max_{j=1}^{m}(P(p = pl_j) * \prod_{i=1}^{n} A_{jpl_{s_i}}) \tag{2}$$

where $pl_{s_i} \in PL$. In equation 2, $P[(q_1, \ldots, q_n) \leftarrow (pl_{s_1}, \ldots, pls_n)]$ is the maximum probability of generating the nodes of upper levels by the model as well as assigning pl_{s_1} to page q_1, pl_{s_2} to page q_2, $\cdots$, and pl_{s_n} to page q_n. To calculate equation 2 for lower levels, we should have the probability of each page individually. Therefore we calculate these probabilities by equation 3.

$$P(q_i = pl_j) = \sum P[(q_1, \ldots, q_n) \leftarrow (pl_{s_1}, \ldots, pl_{s_n})] \tag{3}$$

where $pl_{s_i} = pl_j$ and $\forall k \neq i : pl_{s_k} \in PL$. The probability of generating the tree model of the website ws from the model SCM_i when q represents a leaf of the tree is:

$$P(T_{ws}|SCM_i) = \prod_{q} max_{j=1}^{|PL|} P(q = pl_j) \tag{4}$$

Here we want to compute the complexity of our algorithm. We should compute $p(q_i = pl_j)$ for every node in each level. Suppose we have branching factor of maximum b in this tree, so for each set of siblings the computation of $P[(q_1, ..., q_b) \leftarrow (pl_{s_1}, ...pl_{s_b})]$ for every possible order of $< s_1, \ldots, s_b >$ has $O(n^{b+1})$ time complexity, in which n is the number of page labels. For each node in this set of siblings we can calculate $p(q_i = pl_j)$ from $P[(q_1, ..., q_b) \leftarrow (pl_{s_1}, ...pl_{s_b})]$ probabilities in $O(n^{b-1})$ time. So we compute $p(q_i = pl_j)$ for each node in $O(n^{b+1})$. Whereas we should compute this probability for every node in the tree, the total complexity will be $O(n^{b+1}b^{L-2})$, where L is the number of tree's levels.

5 Learning and Sampling

As mentioned above, first, we determine website class and webpage class labels. Then a set of sample websites for learning phase are assigned by an expert or from sites like DMOZ as seeds for constructing class models. One of the fundamental steps in this phase is assigning a label to each web page. There are two types of pages: One that has predetermined labels (i.e. form a constant part in their

URL) and the other which has no specified label. We have to assign labels to the second type. We assign labels to about %2 of pages in the Training set manually. The remaining %98 of the pages will be labeled by Naive Bayes based upon this 2 percent. To construct website class models, we have to compute the state-transition matrix A and emission probability e. A can be easily computed as follows.

$$a_{ij} = \frac{N(c_i, c_j) + 1}{\sum_{k=1}^{n} N(c_i, c_k) + n}$$

in which $N(c_i, c_j)$ is the number of times that a page of type c_i links to a page of type c_j in the training set. As we use naïve-bayes, we can also easily compute $e_{c_j}(p)$. Therefore, for each page model c_j and word w_l

$$P(w_l|c_j) = \frac{N(w_l, c_j) + 1}{\sum_{i=1}^{|V|} N(w_i, c_j) + |V|}$$

$N(w_l, c_j)$ is the number of occurrence of word w_l in webpages of class c_j and V represents the set of selected keywords. Therefore, if p is formed from $w_1 \dots w_n$ then $e_{c_j}(p) = P(w_1|c_j) \dots P(w_n|c_j)$.

There are two general reasons for our motivation to use page pruning algorithm. First, downloading web pages in contrast with operations that take place in memory is very expensive and time consuming. In a typical website there are too many pages that cannot convey useful information for website classification, if we can prune these pages, website classification performance improves significantly. The second reason that leads us to use pruning algorithm is that in a typical website, there are pages that affect classification in an undesirable direction, so pruning these unrelated or unspecific pages can improve our accuracy in addition to performance.

We use the pruning measures used in [6] for its efficiency and we modify the pruning algorithm for our method. To compute measures for a partial tree which we have downloaded up to now, we should be able to compute membership of this partial tree website for each website class. Our model is suitable for this computation because we compute the probability of each partial tree incrementally and when we add new level to previous partial tree we just compute the $P[(q_1, ..., q_b) \leftarrow (pl_{s_1}, ...pl_{s_b})]$ probabilities for every possible set of $\{s_1, \dots, s_b\}$ for all the new siblings by using previous probabilities and according to our algorithm. Then by using these probabilities we calculate $p(q_i = pl_j)$ for each node in the new level. For each node q, we define $P(q|sl_i) = \max_{pl_j \in PL} p(q = pl_j)$ where $sl_i \in SL$.

For each node q of partial website tree t: $weight(q) = \sigma^2_{sl_i \in SL}(P(q|sl_i)^{\frac{1}{depth(q)}})$. By adding a new node q to a partial tree t we obtain a new partial tree t_2. We stop descending the tree at q if and only if $weight(q) < weight(parent(q)) * \frac{depth(q)}{\omega}$. By means of this pruning procedure and saving probabilities, we perform classification in the same time we use pruning algorithm, and then we can determine the most similar class of the given website. We examine different ω to find appropriate one for our data set. Choosing proper ω can help us to achieve even higher accuracy than complete website download.

6 Experimental Results

In this section we demonstrate the results of some experimental evaluation on our approach and compare it with other existing algorithms, mainly extracted from [6]. We use these methods to classify scientific websites into ten following classes: Agriculture, Astronomy, Biology, Chemistry, Computer Science, Earth Sciences, Environment, Math, Physics, Other. We use DMOZ[3] directory to obtain websites for the first 9 classes. We downloaded 25 website for each class and a total of 86842 web pages. At this point we downloaded a website almost completely (Limited number of pages at most 400). For the "other" class we randomly chose 50 websites from Yahoo! Directory classes that were not in the first nine classes which had 18658 web pages. We downloaded all websites to local computer and saved them locally. To use our algorithm first we should prepare our initial data seed and then we build a model for each class and compute its parameter as stated in the learning phase. To classify the pages of a website class, we labeled about %2 of them in each web site class manually then The remaining %98 of the pages were labeled by Naive Bayes based upon this labeled pages. At the end of this process we have a naïve-bayes model for each page class of each website category. By means of these naïve-base models we classified web pages for other methods. For testing our methods we randomly downloaded 15 new website almost complete (limiting to 400 pages)for each class.

We compare our algorithm to 4 other methods: 0-order Markov tree, C4.5, Naïve-bayes, classification of superpage. In superpage, classifying a web site is to extend the methods used for page classification to our definition of web sites. We just generate a single feature vector counting the frequency of terms over all HTML-pages of the whole site, i.e. we represent a web site as a single "super-page". For C4.5 and naïve-bayes, first we build Feature vector of topic frequencies and then apply naïve-bayes and C4.5 algorithms on them. For the 0-order Markov tree we used the method described in [6]. You can find the accuracy of tested methods on testing dataset in table 1.

Table 1. Comparison of accuracy between different classification methods

Classifier	Accuracy
Super Page	57.7
Naïve-Bayes	71.8
C4.5	78.5
0-Order Markov Tree	81.4
Our algorithm	85.9

As it can be seen the accuracy of our method is better than other methods. It is more accurate compared to 0-order Markov tree because page classes are hidden here and we calculate probability of the whole website that is generated from a model.

At last we examine the impact of different ω values on the sampling algorithm in our training set. With an appropriate ω, the accuracy increases in comparison to complete website. To find appropriate ω, we increased ω gradually and when the overall accuracy stopped to increase, we choose ω. In our data set the appropriate ω was 6 , but this can change in respect to data set.

7 Conclusions and Future Works

In the growing world of web, taking advantage of different methods to classify websites seems to be very necessary. Website classification algorithms for discovery of interesting information leads many users to retrieve their desirable data more accurately and more quickly. This paper proposes a novel method for solving this problem. With extending Hidden Markov Model, we described models for website classes and looked for the most similar class for any website. Experimental Results show the efficiency of this new method for classification.

In the ongoing work, we are seeking for new methods to improve the efficiency and accuracy of our website classification method. Demonstrating websites with stronger models like website graphs can bring us more accuracy.

References

1. G. Attardi, A. Gullí, and F. Sebastiani: Automatic Web page categorization by link and context analysis. In: *Proc. of THAI-99, European Symposium on Telematics, Hypermedia and Artificial Intelligence*, 105–119, Varese, IT (1999).
2. S. Chakrabarti, B. E. Dom, and P. Indyk: Enhanced hypertext categorization using hyperlinks. In: *Proc. ACM SIGMOD*, 307–318, Seattle, US (1998)
3. DMOZ. open directory project.
4. P. Frasconi, G. Soda, and A. Vullo: Text categorization for multi-page documents: A hybrid naïve bayes hmm approach. In: *Proc. 1st ACM-IEEE Joint Conference on Digital Libraries* (2001)
5. J. Han and M. Kamber: *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publisher, San Francisco, California (2006)
6. M.Ester, H. Kriegel, and M.Schubert: Web site mining: A new way to spot competitors, customers and suppliers in the world wide web. In: *Proc. of SIGKDD'02, 249–258,* Edmonton, Alberta, Canada (2002)
7. HP. Kriegel, M. Schubert: Classification of Websites as Sets of Feature Vectors. In: Proc. IASTED DBA (2004)
8. J. M. Pierre: On the automated classification of web sites. In: *Linkö ping Electronic Article in Computer and Information Science*, Sweden 6(001)(2001).
9. D. Shen, Z. Chen, H.-J. Zeng, B. Zhang, Q. Yang, W.-Y. Ma, and Y. Lu: Web-page classification through summarization. In: *Proc. of 27th Annual International ACM SIGIR Conference* (2004)
10. Y.-H. Tian, T.-J. Huang, and W. Gao: Two-phase web site classification based on hidden markov tree models. *Web Intelli. and Agent Sys.*, 2(4):249–264 (2004)
11. Yahoo! Directory service.