Webpage Classification based on Compound of Using HTML Features & URL Features and Features of Sibling Pages

Sara-Meshkizadeh [*1], Dr.Amir Masoud-Rahmani, [*2]
[*1, **Corresponding author**] *Department of Computer engineering, Science and Research branch,Islamic Azad University(IAU), Khouzestan,Iran*
[*2] *Department of Computer engineering, Science and Research branch,Islamic Azad University(IAU) , Tehran,Iran*
Sara_meshkizadeh@yahoo.com
, rahmani@srbiau.ac.ir
doi: 10.4156/ijact.vol2.issue4.4

Abstract

Webpage classification plays an important role in information organization and retrieval. It involves assignment of one webpage to one or more than one predetermined categories. The uncontrolled features of web content implies that more work is required for webpage classification compared with traditional text classification. The interconnected nature of hyper text, however, carries some features which contribute to the process, for example HTML Tags and URL features of a webpage. This study illustrates that using such features along with features of sibling pages, i.e. pages from the same sibling, as well as Bayesian algorithm for combining the results of these features, it would be possible to improve the accuracy of webpage classification based on this algorithm.

Keywords: *Classification, Hyper text, URL, HTML TAGS, Sibling pages, Bayesian algorithm*

1. Introduction

Traditional classification is supposed to be a supervised learning where a set of labeled data can be used for teaching a classifier to be employed for future classification. Webpage classification is distinct from traditional in a number of aspects. First, traditional textual classification is usually performed on the structured documents written based on a fixed style (e.g. news articles), whereas webpage content is far from such characteristics. Second, web pages are documents with HTML structure which might be translated for the user visually. Classification plays a significant role in information management and retrieval task. On the web, webpage classification is essential for focused crawling, web directories, topic specific web link, contextual advertising, and topical structure. It can be of great help in increasing search quality on the web. Using some changes in Bayesian algorithm , [1] has achieved 55% accuracy in classification. [2] uses user information as well as 6 HTML TAGS to provide a classification. [3] involves combination of parent page information with HTML features to arrive at classification. [4] reaches 80% accuracy through a combination of web mining information. [5]Combining some features of HTML and URL and in sum 9 features, classification is made with 80% accuracy. [6] has achieved 48% accuracy in classification. [7] uses URL information as well as webpage pictures to provide a classification with tree structure.This study introduces a novel method for webpage classification as it enhances existing algorithms. This method combines three different features of webpage URL with information extracted from 24 Tags of HTML Structure of a webpage, and it also employs information from Sibling pages with the same Parent to increase classification accuracy.

In this paper, in Section 2 related concepts are described. Section 3 introduces the proposed algorithm in detail. Section 4 evaluates the suggested algorithm, and Section 5 provides conclusions and insights towards future work.

2. Related Concepts

2. 1. Selecting Database

As webpage classification is usually taken as a supervised learning, there is a need for classified samples for learning. In addition, some samples are also required to test classification for classifier evaluation. Manual labeling involves something more than human, therefore, some available web directories are used for more research. One of them which is employed more is ODP [12]. In this database, 4519050 various websites are classified by 84430 classification editors. This research has been performed on universities, shopping, forums and FAQ (frequently asked questions) categories of ODP.

1

2. 2. Feature Extraction from Webpage

The textual content is the most salient feature that state on a web page and it can be used. Although with the care of different patterns of interference , the direct use of a bag of words that is presents to every phrase may decrease the access of optimal performance. Researchers have proposed various ways for optimal using of textual features. One popular way is feature selection. One apparent feature occurs in HTML and lack in traditional textual HTML is the existence of HTML Tags. It's substantiated that using of information derives from tags can strengthen the classifier. Furthermore, tags using can be the merits of constructed information in HTML files that is usually ignored in textual method. Though, since the most of HTML tags are used more in different presentations, web page designers have offered various versions of tags. In addition to features of HTML tags, a webpage can be classified based on its own URL. URLs are highly effective in classification. First, a URL is easily retrievable, and each URL is limited to one webpage, and each webpage has one special URL. Second, if this method is solely employed, classification of one webpage based on its URL causes download removal of the whole page. This tends to be an appropriate method for classifying pages which are not existing or their download is impossible, or time/space is critical for example in realtime classifications.

2. 3. Bayesian Algorithm

Bayesian inference is provides a probability method for inference. This method is built on the hypothesis that the considered values follow a probable distribution, and that optimal decisions can be made with an eye to inference on the probabilities as well as observed data. As this method is a quantitative one for weighing evidences which support different hypotheses, it is of great importance in machine learning. Bayesian inference provides a direct method for dealing with probabilities for learning algorithms, and it also creates a framework for analyzing performance of algorithms which are not directly related to probabilities. In many cases, the problem is finding the best hypothesis in hypothesis space H with available D learning data. One method to express the best hypothesis is that we claim we are looking for the most probable hypothesis with D data in addition to initial data on prior probabilities H. Bayesian theorem is also a direct method for calculating these probabilities.

To define Bayesian theorem, P(h) is used to express initial probability which maintains h hypothesis is true, earlier than observing learning data. P(h) is usually called prior probability, and it expresses any prior knowledge which states on the chance of correctness of hypothesis h. If there is no initial knowledge on hypotheses, we can assign a similar probability to the whole hypotheses space H. Likewise, P(h) is similarly used for expressing prior probability where D data are observed. In other words, probability of observing D in the case of there is no knowledge on correctness of hypotheses. P(D|h) is employed to express probability D in a space where hypothesis h is true. In machine learning we look for P(D|h), i.e. probability of correctness of hypothesis h in the case of observing D learning data. P(D|h) is called post probability, as it expresses our confidence of hypothesis h after observing D data.

Bayesian theorem is the main building block of Bayesian learning, as it provides a method for calculating post probability P(D|h) based on P(h) along with P(D) and P(D|h).

$$P(h|D) = \frac{P(D|h)\,P(h)}{P(D)} \quad (1)$$

As expected, it can be seen that P(h|D) increases with the increase of P(h), and P(D|h). Therefore, it is reasonable that P(D|h) decreases with the increase of P(D), because with higher probability of occurrence P(D) which is independent from h, fewer evidence of D are available to support h. In many learning scenarios, the learner considers a set of hypotheses H, and it is interested in finding the hypothesis h∈H which is the most probable one or at least one of the most probable ones. Any hypothesis which carries such feature is called Maximum a posteriori, MAP.

Using Bayesian theorem, it is possible to find MAP hypothesis to calculate posteriori probability of each candidate. In other words, HMAP is a hypothesis that

$$\begin{aligned} HMAP &= \text{argmax } hj \in H \; P(hj \mid Di) \\ &= \text{argmax } hj \in H(P(Di|hj) P(hj) /P(D)) \\ &= \text{argmax } hj \in H \, P(Di|hj) \, P(hj) \end{aligned} \quad (2)$$

Be noted that P(D) is deleted at the final step, as its calculation is independent from h, and it is always a constant. However, for all probable states on different features of a problem, this theorem can be generalized for all existing probabilities, as for all values of feature D1,D2…Dn:

$$H = \text{argmax } hj \in H \; P(hj) \, \Pi \, P(Di|hj) \quad (3)$$

And P(Di|hj) is calculated as follow:

$$P(Di|hj) = \frac{n\_c + mp}{n+m} \quad (4)$$

Where:
n= number of examples where h= hj
nc = number of examples where samples D= Di and h= hj
p= initial estimation for P(Di|hj)
m= size of sample space

2. 4. Using Features of Neighbor Webpage

Although the web pages include useful features, but in the special web page, these features may not exist, or lead to a mistake or under some reason may not be distinguished. For example, some web pages encompass large images or Flash objects, but they have less textual content. In these cases, it's difficult for classifiers to make a sound judgement based on the page features. To compound the problem, features can be extracted from adjacent pages that are connected to pages which must be classified in ways which information is gathered. The idea of using adjacent information can be used in instructor-free training. To sum up, on one side, the father, the child, the sibling and the spouse are useful pages in classification, and on the other side, because using of adjacent page information can lead to produce extra noise, we must use this information with great cautiousness. In this article, In order to reach a more accurate classification in pages that have same features, the brother pages are used. Because it seems that these pages have more similarity with certain page in shared subjects and therefore it's more than other adjacent available to use.

3. Proposed Algorithm

3. 1. Pre-processing of Web Content

In most studies, pre-processing is performed prior to feeding the web content to the classifier.first , useless html tags are removed. then,the content of usefull html tags and the URL address and page Title are pre-processed, i.e. words shorter than 3 characters, numbers, conjunctions and prepositions stored in a table called Stopwords are removed from this content. Using the function Porter Stemmer

[15], the useful words are changed to their stems (to avoid data redundancy), and then they are stored in the relevant table along with a value as the frequency of word in the data bank of the program.

3. 2. Extracting Features from webpages

3. 2. 1. Extracting Features from Html Tags

The textual content is the most salient feature that state on a web page and it can be used. Although with the care of different patterns of interference , the direct use of a bag of words that is presents to every phrase may decrease the access of optimal performance. Researchers have proposed various ways for optimal using of textual features. One popular way is *feature selection.* One apparent feature occurs in HTML and lack in traditional textual HTML is the existence of HTML Tags. It's substantiated that using of information derives from tags can strengthen the classifier. Furthermore, tags using can be the merits of constructed information in HTML files that is usually ignored in textual method. Though, since the most of HTML tags are used more in different presentations, web page designers have offered various versions of tags. In this case and in order to access the new features of web pages not reviewed up to present time, all HTML tags that are used in web pages are reviewed statistically on a 100 various web pages and among all, 24 items present in Table(1), are recognized useful for more accurate classification of web pages. At the time of reviewing these features one pay attention to various criteria such as : The selected features must be distinguishable, they must be useful for recognizing accurate category and they don't permeate with other selection features. It's emphasized that in researches conducted until now, maximum 5 or 6 items were reviewed and we could remarkably increase algorithm accuracy by evaluating more cases and using them in classification.

Table 1. HTML tags used in classification

HTML TAG	Remarks	Number
<Body> </ Body >	IT INCLUDES THE MAIN PASSAGE OF THE PAGE	1
<Title> </ Title >	IT ENCOMPASSES THE MAIN SUBJECT OF THE PAGE THAT IS SEEN IN BROWSER	2
<h1></ h1>	FOR IDENTIFYING HEADING OF A PAGE	3
** **	FOR IDENTIFYING THE CONTENT OF A PARAGRAGH	4
<q> </q>	FOR IDENTIFYING THE CONTENT OF A QUOTATION THAT MUST PUT IN	5
<a>	FOR CONSTRUCTING A LINK WITH OTHER PAGE OR THE SAME PAGE-THE TITLE CHARACTERISTIC OF THIS TAGS FOR PRESENTING OF POINTED LINK IS USEFUL FOR CLASSIFICATION	6
<Blockquote> </Blockquote >	TO ENCOMPASS THE INDENTED PASSAGE THAT IT TIMELY LONG SEVERAL LINES	7
<ImgSrc="Url">	FOR ATTACHING A PICTURE TO THE CONTENT OF THE WEB PAGE WHICH URL ADDRESS CAN BE USED.THE alt CHARACTERISTIC OF THIS TAG IS USEFUL FOR CLASSIFICATION WHEN THE PICTURE IS NOT OBSERVABLE BY BROWSER	8
<Map> </ Map >	FOR IDENTIFYING A MAP IN A PICTURE THAT INCLUDES VARIOUS LINKS TO PICTURE	9
<Table> </ Table >	FOR IDENTIFYING THE CONTENT OF A TABLE	10
<Tr> </ Tr >	FOR IDENTIFYING THE RANGES OF THE TABLE	11
<Th> </ Th >	FOR IDENTIFYIG THE HEADER CELLS OF THE TABLE	12
< Td > </ Td >	FOR IDENTIFYIG THE ELEMENTS OF A TABLE	13
<Caption>	FOR IDENTIFYING THE SUBJECT OF A TABLE	14
<Base>	FOR IDENTIFYIG THE MAIN URL OF A PAGE THAT ALL LINKES	15

</ Base >	REFER TO IT	
<Form> </ Form >	FOR EXPLAINING OF A FORM IN THE PAGE IS USED	16
<Button> </ Button >	FOR DEFINING OF A BUTTON THAT IT'S SUBJECT CAN BE USEFUL IS USED	17
<Legend> </ Legend >	FOR IDENTIFYIG THE DETAIL OF AN INFORMATION BOX	18
<Fieldset> </ Fieldset >	FOR IDENTIFYIG THE DETAIL OF AN INFORMATION BOX	19
<Input type=" ">	FOR IDENTIFYIG THE DETAIL OF AN INFORMATION BOX	20
<Select> </ Select >	FOR DESIGNING A COMBO BOX IS USED	21
<Option Value=" "> </ Option >	FOR IDENTIFYING THE CONTENTS OF A COMBO BOX IS USED	22
<Optgroup Label> </ Optgroup >	FOR IDENTIFYING THE CONTENTS OF A GROUP IN COMBO BOX	23
<Meta>	FOR INCLUDING THE FEATURES OF A PAGE IS USED(DEFENITION OF THE PAGE METADATA) AND METADATA AND KEYBOARD FEATURES THAT ARE USEFUL IN CLASSIFICATION	24

3. 2. 2. Extracting Features from URL

a) First Feature

URL do have appropriate features for classification. Two sets of features which are easily extracted from URL pages are Postfix and Directory. The general format of a postfix is usually as Abbreviation or Abbreviation.Abbreviation . For example, .edu or ac.ir or .ac.uk indicate pages related to universities or academic websites. The general format of a Directory feature is mostly like Word(Abbreviation)slash. For example, a directory named FAQ or Forum represented as /Faq/ or /Forum/ can represent the relation between the current page to the relevant class. These features are put in Table 2.

Table 2. Features of URL Addresses

URL feature	Specification	number
Postfix : .edu,.ac.ir	To find class of pages based on Stem URL Address	1
Directory : Forums/,Faq/	To find class of pages based on rest of Stem URL Address	2

b) Second Feature

The second feature important in webpage URL is attention to the domains which have been observed by the system and further their correct class is identified. To do so, once the class of a webpage is determined at the last step and its accuracy is confirmed by the user, the page's address is registered in the URL table along with the correct class. Afterwards, if a webpage with the same domain is given to the system, the system can recognize more simply and quickly based on similarity in addresses. Take for example the address www.aut.ac.ir/sites/e-shopping/raja.ir whose domain is recognized as shopping, thereby it is registered under shopping class at the URL table.
If a webpage such as www.aut.ac.ir/sites/e-shopping/iranair.ir is given to the system as a test, due to the domain www.aut.ac.ir/sites/e-shopping/ in the table, the system quickly and simply assigns shopping class to the second address. It should be noted that using domain similarity of webpage URL is a new idea which has not been taken into account in webpage classification.

c) Third Feature

To enhance the efficiency of the proposed algorithm another method based on URL address can also be employed. This method involves expanding URL of a webpage to use existing elements better. Forexample,considering http://www.washington.edu/news/nytimes , and also page title, i.e. NewYorkTimes it is easily understood that the address refers to NewYorkTimes news database. The machine, however, misses the point. To solve the problem and to employ human guesses in the procedure, with an eye to the function used in [1] but with some changes in definition and usage, a likelihood function is presented. Table 3 illustrates the method. The likelihood is used to compare URL's token letter by letter, and similar word in the page Title.

Table 3. Likelihood Function

Rank	Condition	number
2	First letter of URL token is the same as first letter of similar word at page Title	1
1	First letter of URL token is the same another letter of similar word at page Title	2
1	A letter of URL token is the same as a letter of similar word at page Title	3
2	Last letter of URL token is the same as last letter of similar word at page Title	4
0	Ignoring a character of URL token	5

For example consider the news website Newyorktimes with the address http://www.nytimes.com. Now take the title "The New York Times-Breaking News, World News & Multimedia". The above-mentioned function calculates likelihood for a condition where URL token, Nytimes and similar word at page Title is Newyorktimes as bellow:

(Condition1)→N(Condition5)→E(Condition5)→W(Condition1)→Y(Condition5)→O(Condition5)→R (Condition5)→K(Condition1)→T(Condition3)→I(Condition 3)→M(Condition 3)→E(Condition 4)→S

In this case, the word Newyorktimes receives score 11 from words at the page Title which is the highest score out of other tokens, and according to the likelihood function, thereby in future if Nytimes is seen in test samples, it is replaced with Newyorktimes. To work with the third feature, the system runs the likelihood function on tokens of URL and the page Title. In the case of expanding existing token in URL, instead of the former token, an expanded token is used for classification.

3. 3. Webpage Classification

3. 3. 1. Training Step

In this step about 500 pages of pointed classes of ODP database that are downloaded before, are used for teaching system. The HTML file of pointed page accompanies to selected class is supplied with system. The useful words are extracted from features of HTML file of above page and saved in database of relevant class with repetition of every word in every feature on the ground of the amount of that word in that feature of selected class .moreover, about 2000 pages of considered classes in ODP database are used for system learning. As such, the webpage URL address, the page's Title along with the class is fed to the system. In this step, useful words from considered features of the URL are extracted, and are stored in the table related to the considered class of the database along with the frequency of each word in each feature as the value of that word in that feature. In this step, information of pages in each class is stored in the table related to the same class. So far we have been dealing with learning the system and feeding the useful information.

3. 3. 2. Test Step

In this step for testing of the system, a collection of totally various pages from training step is used. To test of the system a three-level procedure is performed. At first , by paying attention to the features relevant to address of URL page, a class is nominated as the selected class based on these features. Then, the procedure pertinent to HTML features is performed and the class based on these features is candidate as the selected class. In conclusion, the last class is selected with comparison between scores of two former steps and this is promulgated to user.

a) Test Step Based on Url features

In this step, to test the system a number of different pages in Training step is used. To do so, URL address and the page Title are given to the system. The system, based on the first feature analyses the possibility of recognizing the category according to the address and checking features of Postfix and Directory. If it fails to find a previously observed Postfix or Directory, it refers to the second feature, and compares the page address with all addresses observed so far. After this step, it is obvious that accuracy of this algorithm increases as the number of learned pages considerably increases. If the address of this page is not similar to existing addresses of the bank, the algorithm analyzes the third feature. It must be stated that the first priority of classification system is the first feature, followed by the second feature and finally the third one. To analyze the third feature, the system performs the likelihood function on existing tokens in URL and page Title. If the existing tokens in URL are expanded, instead of former token an expanded expression is employed for classification. Afterwards, the frequency of each word is calculated as frequency feature. Then using Bayesian rule, posteriori probability P(h|D) from priori probability P(h) along with P(D) and P(D|h) according to frequency of each word on the page is calculated in order to predict the possibility of belonging the page to the category where the highest number of features has been occurred. After calculating the probability for all extracted words of one page for all four categories, the category which has the highest probability for all words is recognized as the main category.

b) Test Step Based on HTML features

In this step, at first, pointed HTML file page is supplied to system. Then pointed features of above HTML page are extracted and the number of repetition of any word of above feature is calculated and seemed as frequency feature of pointed word. Then, by using of Bayesian law, the latest probability P(h|D) from the former probability P(h) together with P(D) and P(D|h) on frequency feature of every word of above page, in order to predicting the probability of belonging of pointed page to the class that the most numbers of above mentioned features occur is calculated. After computing the above probability for all extracted words from a page of 4 pointed classs, the class with the highest score based on Bayesian probability is specified as the selected class of this step.

3. 3. 3. Neighbor Effect Step

In another research that has conducted, all attention went to review of role of four pages such as father, child, spouse , and sibling, in order to specify that In what manner and help of what pages class can help to reach the best accuracy of classification. At the end of the process , we can conclude that when pages with the same father , brother pages or siblings , as adjacent to help in classification are used, the best obtained conclusion of these pages can help the classification to a great extent. For this reason , in order to reach the higher accuracy of classification , the existent information on siblings is used. In this step the conclusion of former two test steps with relevant probabilities are compared to the conclusion obtained from the influence of siblings on classification. The class that has the highest probability among above three probabilities is selected as the main class and announced to user.

Due to the concise nature of most addresses and also Title of many webpages, the classification based on just features extracted from a webpage has been performed with 72.8% accuracy as shown at Table 4. To improve classification accuracy, it was decided to use extracted information from neighbor pages. As such, URL address and Title of 500 neighbor pages related to Test step pages have been employed to increase classification accuracy. As illustrated in Table 4, the accuracy average raised to 85.5% that implies the emphasis on neighbor pages is helpful.

In case of using the HTML features, by paying attention to existent information stated on Table 4 and without using of sibling pages information, accuracy of 87.2% in classification was obtained and after blending this information with sibling pages information, the accuracy of algorithm up to 90.4% . At the end of the process, for categorizing these pages, we utilized the blending of the extracted information both in URL and HTML features that it led to increase of algorithm accuracy up to 91.3% .

In this case when we used of sibling-page features to improve the classification accuracy, the average accuracy was 94.7% that is perfectly indicated the effect of extracting information in sibling pages to improve algorithm accuracy. The Figure 1 represents the effect of this blending in increasing of classification accuracy.

Table 4. Comparison of Accuracy before and after using the information of Sibling pages

parameter	Average Accuracy without using the information of sibling pages	Average Accuracy with using the information of sibling pages
Using URL features	%72.88	%85.8
Using HTMLfeatures	%87.2	%90.4
Using compound of URL and HTML features	%91.3	%94.7

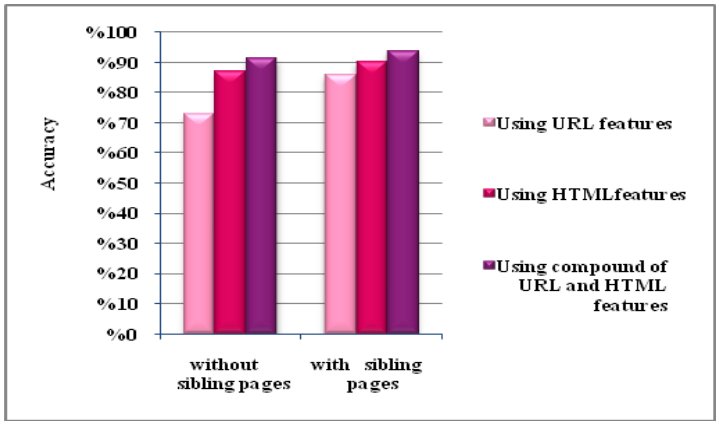

Figure1. the effect of using sibling pages to improve Accuracy

4. Evaluation of Proposed Algorithm

In this study, to evaluate the proposed algorithm, 500 different pages with training and test steps, and categories downloaded from ODP such as universities, shopping, forums, and FAQ have been employed. To implement the algorithm, Vb.net 2005 programming environment as well as Sqlserver 2000 database on a computer with specifications cpu core 2,2.13 GHz, and 2 GB Ram, and 300 GB Hard disk which run the operating system Windows XP service pack 3 is used. After pre-processing downloaded pages and extraction of useful features from URL address and Title of pages, and calculating the probability of belonging of the webpage to each one of categories using Bayesian algorithm, and then calculating the above probability based on neighbor pages information, the category with the highest probability is recognized as main page category and it is announced to the user. If the category is recognized correctly, all its extracted features are added to the table of database, and is added as a correct case to the statistics of correct recognition, parameter a from calculation criterion of algorithm's general accuracy is added too. Otherwise, the correct category is recognized by the user, and the extracted information is added from the page to the correct table. Furthermore, one case to wrongly recognized cases from the first category, i.e. parameter b from calculation criterion of algorithm's general accuracy, and also one case to cases which is wrongly ignored in the correct category, i.e. parameter c from calculation criterion of algorithm's general accuracy are added. To recognize the algorithm accuracy, algorithm accuracy evaluation criterion has been used as follows:

$$\text{Precision} = a / (a+b) \quad (5)$$

$$\text{Recall} = a / (a+c) \quad (6)$$

$$\text{F-measure} = 2 * \text{Recall} * \text{Precision} / (\text{Recall} + \text{Precision}) \quad (7)$$

where a= number of pages of sample examples which are correctly classified, b= number of pages of sample examples which are wrongly classifies, and c= number of sample examples which are wrongly ignored.

The offered algorithm in this Article was evaluated in four phases : phase one evaluates and performed algorithm just on URL features without using of HTML features that this caused the algorithm accuracy(f-measure) up to 72.8% that this improving could obtained from several reason. The first one is the use of blending several features of URL that is not utilized hitherto. For instance, the second feature is selection of URL pages that means utilizing of resemblance scope in URL web page address , this new feature is not used hitherto. The third one means utilizing resemblance function in classifying web pages which is caused with [6] by using this feature the accuracy up to 43% , blending of this method with Bayesian algorithm for categorizing web pages lead to accuracy of 72.8% to be obtained which indicates that this approach need to review more and more. The second phase of evaluating is to perform of Bayesian algorithm on 24 selected features of HTML page which caused in comparison with [2] that it used just 6 features and reached to accuracy of 84% ,algorithm accuracy up to 87.2% that indicated the improvement until this stage. The third phase of evaluating is to compound of using both HTML and URL web page features which caused the classification accuracy of the web pages in comparison with [5] which used some features of URL pages and using of Bayesian algorithm with reach to care of 80% and or [4] which performing Bayesian on above two features that is reached to accuracy of 63% in classification improve to 91.3% . Phase four of evaluating, however, is to blend using of both URL and HTML web page features and using of these features on adjacent pages in classification that in comparison to [3] which used whole passage of adjacent pages reached to care of 86%, the care of 94.7% is reached in classification that is substantiated that in this method value continues and improve more and more. These conclusions in Table 5 and figure 2 is observable too.

Table 5. Comparison of average Accuracy in all classes with various features

Feature	Shopping	University	Forums	Faq	Average F-measure
Using URL Features	76.2%	74.5%	70.7%	69.8%	72.8%
Using HTML Features	88.5%	93.2%	84.6%	82.5%	87.2%
Using compound of URL and HTML features	92.4%	93.3%	91.2%	88.3%	91.3%
Using compound of URL and HTML features and Sibling pages	93.4%	97.7%	92.8%	94.9%	94.7%

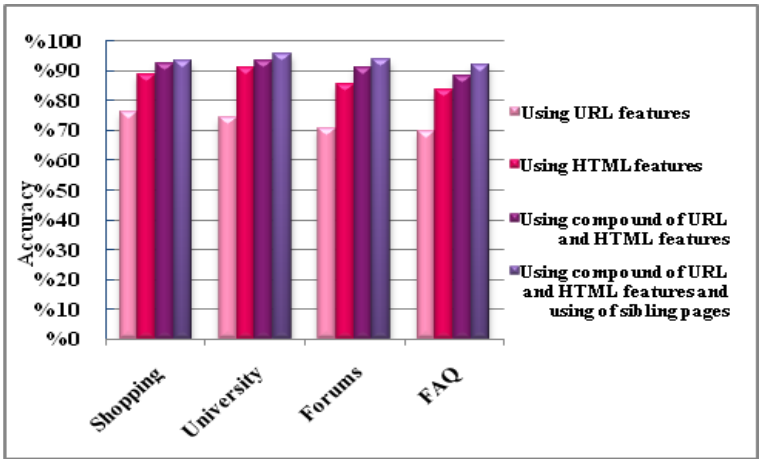

Figure 2. Final Results of Evaluation in various classes

5. Conclusion and Future Works

In this Article the new version of classifying web pages was offered that caused improving available algorithms. At first , this method by blending three various feature of URL pages and then, by utilizing more features of HTML, to decide for categorizing of web pages and at the end of the process

the information of brother adjacent pages is used to help in classification. Finally, the offered algorithm reached to accuracy of 94.7% . The average accuracy of above algorithm is observable in various clusters which is distinguished under the feature of pointed purpose in figure 3. To reach a more improvement in this field, perhaps using of other features of HTML page tags and paying attention to other features based on URL web pages can leave an impression on it.

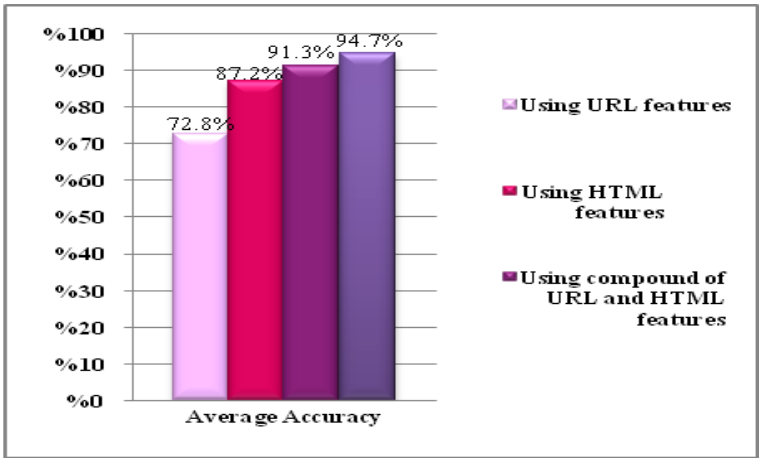

Figure 3. Final Results of Average Accuracy in various classes

6. Acknowledgement

The authors thank National Iranian oil company (NIOC) and National Iranian South oil company (Nisoc) for their help and financial support.

7. References

[1] G.S. Tomar, Shekhar Verma, Ashish Jha, 2006," Web Page Classification using Modified NaïveBayesian Approach", TENCON 2006. 2006 IEEE Region 10 Conference,on ,14-17 Nov 2006, On page(s):1–4.

[2] j.li,W.xue,N.dang, 2007,"Application of the Naïve Bayesian Method with User Current Usage and hierarchy from website in chinese webpage classification", Automation and Logistics, 2007 IEEE International Conference on,18-21 Aug. 2007,On page(s): 1364-1367.

[3] X.qi,B.D.Davison, 2008,"classifiers without borders : incorporating fielded text from neighboring webpages", In Proceedings of the 31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval, Singapore, July 2008, On page(s):643-650.

[4] S.Morales,H.Fandino,J.rodriguez, 2009,"Hypertext Classification to filtrate information on the web", Proceedings of the 2009 Euro American Conference on Telematics and Information Systems: New Opportunities to increase Digital Citizenship 2009, Prague, Czech Republic June 03-05, 2009, Article No.1.

[5] Ch.Lindemann,L.littig, 2006,"Coarse-grained classification of websites by their structural properties", Proceedings of the 8th annual ACM international workshop on Web information and data management table of contents,Arlington, Virginia, USA .OnPage(s): 35-42.

[6] M.kan,2004,"Web page categorization without the web page", Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, ACM,OnPages : 262-263.

[7] L.k.shih,D.r.karger, 2004,"Using URLs and Table Layout for Web Classification Task", Proceedings of the 13th international conference on World Wide Web ,ACM, OnPages : 193-202.

[8] Sebastiani, F.,2002," Machine learning in automated text categorization". ACM Computing Surveys (CSUR) archive,Volume 34 , Issue 1 (March 2002),On Page(s): 1-47.

[9] Chakrabarti, S, Morgan Kaufmann. 2003," Mining the Web: Discovering Knowledge from Hypertext Data",San Francisco, CA. Morgan-Kaufmann Publishers.

[10] Mladenic, D, 1999," Text-learning and related intelligent agents: A survey", Intelligent Systems and their Applications, IEEE**,** Jul/Aug1999, Volume: 14, Issue: 4, On page(s): 44-54.

[11] Getoor, L, Diehl, C. 2005" Link mining: A survey", ACM SIGKDD Explorations Newsletter archive (Special Issue on LinkMining), December 2005, Volume 7 , Issue 2 ,On Page(s): 3–12.

[12] Furnkrunz, J. 2005," Web mining. In The Data Mining and Knowledge Discovery Handbook", O. Maimon

and L. Rokach, Eds. Springer, Berlin, Germany, On Page(s): 899–920
[13] Choi, B. , Yao. Z, 2005," Web page classification. In Foundations and Advances in Data Mining" W. Chu and T. Y. Lin, Eds. Studies in Fuzziness and Soft Computing, vol. 180. Springer-Verlag, Berlin, Germany, On Page(s):221–274.
[14] www.dmoz.org
[15] http://tartarus.org/~martin/PorterStemmer/