IBIMA Publishing
Communications of the IBIMA
http://www.ibimapublishing.com/journals/CIBIMA/cibima.html
Vol. 2010 (2010), Article ID 73408, 8 pages
DOI:10.5171/2010.734081

Joint Web-Feature (JFEAT): A Novel Web Page Classification Framework

Lim Wern Han and Saadat M. Alhashmi

Monash University, Kuala Lumpur, Malaysia

Abstract

With the increasing amount of web pages over the internet, it has been a major concern to obtain information on the internet accurately at a reasonable cost with decent performance. A potential solution is through the classification of web pages into meaningful categories. An effective classification of web pages is of benefit to various applications such as web mining and search engines. Unlike text documents, the nature of web pages limits the performance of successful traditional pure-text classification methods. Noises exist in the form of HTML tags, multimedia contents, dynamic contents and the network structure of web pages which requires a deeper look into effective feature selection of web pages. Often, these features are filtered out relying on the displayed texts of the web page for classification. This paper proposed a framework where web page features are taken into consideration during classification of the web page due to the potential valuable information that might be stored within each of the features. For this reason, this paper explores the potential of the universal Resource Locator (URL), web page title as well as the metadata for information to be used in classification with various categories defined by the users. The framework then explores suitable machine learning algorithms for individual classification of each web feature. The results would then be used for weighted voting to obtain the classification of that webpage. This approach showed improvements over pure-text as well as virtual-webpage classification approaches.

Keywords: web page classification, feature selection, machine learning

Introduction and Motivation

The objective of this research is to obtain an effective and efficient automated web page classification framework through the use of machine learning algorithms on selected web feature. The classification of webpage would help the users in obtaining information.

The World Wide Web is “the universe of network-accessible information, an embodiment of human knowledge” as advocated by Tim Berners-Lee. In a more technical aspect, it is a network of interlinked hypertext documents storing information connected via hyperlinks. This ‘web’ continues to expand day-by-day and as of July 2009, the web consists of 206,026,787 websites according to Netcraft (2010) [01] making it inefficient to obtain information. To help facilitate users in obtaining the information they need, various tools were introduced – net directories and search engines as suggested by Chekuriand Goldwasser (1996) [02].

Among the notable net directories on the web would be the DMOZ Open Directory Project and Yahoo! directory. Both constructed manually observed by Qi and Davison (2009) [03] to classify websites into their categories which allows the users to search for the set of desired websites within a specific category. The effectiveness of such taxonomies or directories in information retrieval has been demonstrated by

Contact author: Lim Wern Han e–mail: lim.wern.han@infotech.monash.edu.my

Dumais and Chen (2000) [04] to a great success. Accurate classification of web pages allows timely crawling of web pages which would vary based on their categorization such as the constant update of a news web page or blog. The classification of a web page often tells us its nature and structure which is highly valuable for effective and efficient data mining of web pages.

As these directories are manually classified and maintained for accuracy reasons, it is expensive, time consuming and would never achieve the complete coverage of the internet. This paved the way for machine learning methods in automatic classification of web pages as suggested by Tsukada et al (2001) [05]. Shen et al (2007) [06] stated that, unlike successful text-based documents through machine learning, the hypertext nature of web pages contain various noises that at times, would contain more irrelevant information as opposed to relevant ones.

First, we look into similar past researches conducted (section 1). We then propose that we extract information from web features with the lowest probability source of noises which lies with the URL, webpage title and the Metadata (section 2). These features are the core elements of a web page which is easily identified without being extremely dynamic and volatile.

In section 3, we present our framework – the Joint-Feature classification framework and related concepts which supports the framework. We then conducted experiments on the framework and analyze the results (section 4). We look into the possible future works to further extend this research in section5 and concluded our findings from this research in section 6.

1. Related Works

Over the years, researchers looked into automatic web page classification with high accuracy and efficient performance to replace the current manual classification of web pages.

Tate and (1996)[09] suggested that different type of web pages would require different evaluation criteria with the proposal of the 5 most commonly encountered web pages – advocacy, business, information, news and personal web pages. One of the approaches would be through the use of machine learning techniques as shown by Tsukadaet al (2001) [05] which obtained high accuracy in automatic classification in comparison with those of Yahoo! Japan. Boet al (2009) [15] compares the SU and Relief feature selection methods with respect to the performance of 6 classifiers, which 3 of those is selected to be explored in our paper. Motivated by this, we would look into various machine learning algorithms which would be trained to automatically classify web pages.

Due to the nature of webpages and noises,Shenet al (2007)[06] use text summarization techniques via an ensemble method to remove the noise in web pages in the attempt to improve the performance of web page classification. Using this method, they are able to achieve more than 12% improvement over the pure text method traditionally used for web pages. Similarly, Choi and Yao (2005)[07] remove noises in the form of HTML tags, stop words as well as rare words. It identifies and uses important information stored in the HTML structure of the web page such as title, heading and metadata for classification weighted using a Structure-oriented Weighting Technique (SWT). Kan and Thi (2005)[10] explore the use of significant values on various HTML tags on each keyword which provided improvement on accuracy.

Besides information found within the web pages itself, the importance of the uniform resource location (URL) feature of a web page in classification was identified in classification of webpages by Kan and Thi (2005) [10].Shih and Karger (2004)[11] found success in an automatic web page classification system that uses URL and their location on the web page in a tree-like structure to identify web advertisements. Thus, we would look into the URL of web pages in our attempt for an effective web page classifier for various classes instead of only advertisements.

Chen and Choi (2008)[16] extracted 31 features from the web page ranging from URL, texts and various HTML tags to classify web pages into 5 genres with high success. As the authors stated however, it requires considerable manual tuning during the training phase. As the increase in feature increases the complexity and the overhead of the classifier, Indraet al (2008)[08] proposed the use of feature selected during pre-processing. Currently, we selected 3 features of a webpage to help classify the webpage.

Qi and Davison (2009)[3] summarizes the various concepts used for automatic web page classification with respect to recent works. The distinct difference with our work is how we classify each feature individually and then joint to give the classification of the web page.

2. Web page features

Currently, we are interested in 3 distinct feature of a webpage what we believe stores information which are of importance in the classification of a web page through machine learning. In the development process, each of these features would be processed individually and provide a classification to the web page based on their accuracy with the machine learning algorithm of choice. Jointly, these features would ultimately decide the classification of the web page.

2.1. Universal Resource Locator (URL)

The universal resource locator of a web page could be separated into 2 parts – the host and the path. The host would be the domain name of the web page which carries the name of the webpage providing a hint of the classification of the web page. The extension for the host as well could possibly help out on the classification of the web page such as ".gov" for government websites though the ".com" extension is widely used even though it was meant for commercial web pages. Similarly, the path of the web page stores information about the web page due to their organization within the website through folders (for example product folders).

In the development process, the URL would be split through into its host and path. Individually, these elements would be parsed via n-gram to obtain tokens which would be crossed referenced with a dictionary and weighted. For each subdirectory in the path, there would be a decay of 10% in weights.

2.2. Title

The web page title is one of the most important features of a web page due to the information it conveys about the web page in a short piece of text. Often, the web page title is not left out from web pages and is easily found under the <title> tag which could be extracted and processed through a text based approach.

2.3. Metadata

The metadata of a web page stores information about the web page which would not be displayed by the browser found within Meta tags. In this research, we look into Meta descriptions and Meta keywords. The Meta description contains a short description about the web page that could be processed using full text methods. The Meta keywords contain the keywords that provide short and accurate information about the web page. Large number of keywords would be given penalty for overextending.

3. Joint-Feature Classification

Through our understanding of the nature of web pages and identifying the features to be used for our classification (see Section 2); we now propose an approach to web page classification called Joint-Feature classification.

3.1. Categorization/ Genre

The category or genre which web pages would be classified into is flexible to the need of the user via machine learning classification algorithms. Thus, the framework is flexible for any classification classes that are required by the user.

3.2. Training Sources

The classifier would need to be trained using web pages obtained manually by the user to suit his/ her classification from the web such as taxonomies (DMOZ, Yahoo!) or search engines (Google). It is these training web pages and their associated class that would decide the performance of the classifier with respect to the user's perception of the web page. It is subjective to the user's perception and interpretation on what category/ genre these training web pages fall into. Thus, the classifier would be attuned to that user's need automatically via machine learning.

3.3. Preprocessing

Each web page would be preprocessed individually to extract information from the selected feature of the web pages (see Section 3). Each feature of the web page is processed separately to suit their nature. The processing of each feature would be based on the same bag of words obtained from a suggested java dictionary (over 83,700 words) to ensure consistency. Tokens which do not belong to that bag would not be accepted to be used during training and classification. Of course, that bag of words could be edited to meet the requirements of the users.

The result from the preprocessing of these web pages would be stored into an ARFF file (to be compatible with WEKA) for each web page feature to be used for training or classification.

3.4. Training

Instead of building a virtual web page with all of the extracted features and their information, we suggest that these features are classified separately. Each having feature having its own suitable classifier (though it might vary based on the training data) with the highest accuracy for the feature. Thus, one classifier for each feature would be trained based on the training data to build the relationship between the feature and the web page classification.

3.5. Machine Learning Algorithms

We took the machine learning approach to automatic web page classification where through machine learning, our solution would learn and classify web pages into their appropriate category or genre. There is a wide range of machine learning algorithms out there. For our work, we would be using the machine learning tools found within the Waikato Environment for Knowledge Analysis (WEKA). Among the considered machine learning algorithms would be: -

- Naïve Bayes (NB) which applies probabilistic statistical Bayes's Theorem in assigning the most likely class to a given example described by its feature vector [12] with the assumption that these features are independent of each other in the classification (conditional independent).
- Support Vector Machine (SVM), relatively new machine learning classification algorithm based on the non-linear mapping of input vectors to a high-dimension feature space [13] where a linear decision surface (hyperplane) is constructed for classification (binary). For our work, we would be splitting multi class problems into multiple binary classifications using the SMO class in WEKA.
- C4.5 Tree algorithm by Quinlan being one of the most popular decision tree learning algorithm where a tree data structure is created and used to classify new cases based on a set of training cases where each case could be described by a set of attributes [14].

3.6. Looking as a Whole

Given a particular sentence or body of words, we would look into this body as a combination of the words rather than just looking at each of the word in this body and how would they go in hand with classifiers. For example given a body text of "selling a book" we would look at how 'selling' and 'book' as a whole help to decide that the body belongs to a business categorization and not looking at 'selling' as a business classification and 'book' as an education classification thus removing the need for a manually define dictionary. This helps the classification of the webpage as words has different meaning based on the context that they are used in.

This could be trained automatically using the classifier instead of the users defining their own dictionary to associate each word to a particular classification (such as a book as an education word). Thus, it would require less human effort to build and maintain the classifying model which is one of the goals of the problem (automatic web page classification).

3.7. Joint-Feature

The result of the training would be evaluated (0-10 folds depending on the memory load) for each feature to obtain the accuracy of each feature in classification. The result from this evaluation would be used to assign weights to each feature.

When a web page is to be classified, each feature is classified individually. The results of these classifications are then multiplied by the weights given for each feature and then joined together to decide the classification of the web page based on their weights.

4. Experimental Analysis

4.1. Training Data

For our testing, we selected 337 web pages from the Yahoo! directory spanning over 6 categories found within the 'popular websites' recommendations to be used as our training data. These web pages were selected based on both our judgment with regard to the accompanied explanation texts for each link. The categorization and the number of selected web pages are shown in Table 1. The unbalance number of web pages is due to the varying suitable "popular websites" for each category.

Table 1. Training data

Category	Number of Web Pages
BusinessEconomy	79
Entertainment	38
Government	43
Health	88
News	40
Sports	49

4.2. Training Evaluation

We recorded 325 web pages that were successfully parsed and processed for training the classifier. Thus, these 325 were processed successfully to be used to train the classifier.

From our various evaluation runs (10-folds) of smaller data size due to memory constraint, we concluded that SVM would be the most suitable machine learning algorithm for this set of test data. The result on the evaluation done for the training data using SVM with the evaluation tool from WEKA is available in Table 2. The accuracy of each feature would be used as the weight of that feature rounded to the nearest whole number.

Table 2. Training data evaluation (SVM)

Feature	Correctly Classified Instances	Weight
URL	93.8462 %	94
Title	77.2308 %	77
Meta keywords	68.0000 %	68
Meta descriptions	72.6154 %	73

4.3. Evaluation Data

We selected 75 suitable web pages from the Yahoo! directory of the same categories under the "popular websites' of each category as the training data but are not found within the training data to test our proposed solution. The unbalanced numbers for each category is due to the number of web pages in each category found within the "popular websites" of each category within the Yahoo! directory with majority of them selected web pages to be used in training. Web pages with technical issues such as dead links, containing foreign language, being under maintenance or time-out error are omitted from the evaluation.

4.4. Evaluation Results

The result from our evaluation testing is shown in Table 3 looking at the actual number of web pages for each category and the number of correctly/ incorrectly classified web pages for each category.

Table 3. Evaluation data result

Category	Total Number of Web Pages	Number of Web Pages Classified as	
		Correct	Incorrect
BusinessEconomy	20	20	15
Entertainment	11	7	0
Government	7	4	0
Health	20	16	0
News	9	5	1
Sports	8	7	0
Total	75	59	16

We measured the effectiveness of our solution using precision, recall and the f-measure values for each category in the context of classification. These values are least affected by the varying number of web pages for each category unlike that of accuracy (where high number of correctly classified web pages of one class would boost up the accuracy of the evaluation). High precision denotes smaller classification error of webpages whereas high recall denotes smaller leak of classification from positive web pages. The result is shown as in Table 4 and Table 5.

$$Precision = \frac{\begin{array}{c} numberofwebpagesclassified \\ correctlyaspositive \end{array}}{\begin{array}{c} numberofwebpagesclassified \\ aspositive \end{array}}$$

$$Recall = \frac{\begin{array}{c} numberofwebpagescalssified \\ correctlyaspositive \end{array}}{numberofpositivewebpages}$$

$$F - measure = \frac{2 \times recall \times precision}{recall + precision}$$

Table 4. Effectiveness of solution

Category	Precision	Recall	F-Measure
BusinessEconomy	57.1429 %	100 %	72.7273 %
Entertainment	100 %	63.6364 %	77.7778 %
Government	100 %	57.1429 %	72.7273 %
Health	100 %	80 %	88.8889 %
News	83.3333 %	55.5556 %	66.6667 %
Sports	100 %	87.5 %	93.3333 %
Average	90.0794 %	73.9725 %	78.6869 %

4.5. Discussion

From the evaluation results, it could be seen that the “Business and Economy” category is of high recall but low precision. In other words, the business category is overpowering the other categories. Upon inspection, we note that this is partly due to the overlapping of the business category with other categories such as entertainment business (entertainment), business organization and trade bodies (government); health business (health), business and economy news (news) as well as sport business (sports) over the diversity of training data based on the Yahoo! classification. The same could be said for the news category where a government website (based on Yahoo!) was classified by our solution as a news website where the website focuses on governmental news.

We believe that with better definition of classification categories or genre, the precision of our solution could be increased as the amount of overlapping between categories is reduced. When overlapping could not been prevented, it is suggested that we allow overlapping of categories as the classification of a web page is highly subjective based on the perspective of the users. If the users would personally train the solution based on their perspective, the solution would be highly precise to him or her.

Due to the low precision of the classification for the “Business and Economy” category, the recall of the other categories is affected particularly where web pages of those categories are classified as business web pages over the ones suggested by the Yahoo! directory’s evaluators. This is particularly obvious where the business category is highly dominant over the government and news categories in overlapping cases. Another possible explanation could also be attributed to the high number of training web pages within the business category which increases the amount of category overlapping web pages to be trained as business web pages. The solution to this would be the clear definition of categories.

The advantage of our solution where each feature is classified individually and joint by weight is that we could assign more than one classes to a web page by allowing the classification which come in second by weights to define the sub category of a particular web page. For example, a web page about business news would be a main-class news web page as well as an off-class business web page. This way, we could increase the effectiveness of information retrieval through our solution.

In the investigation, we perform a comparison of results with that of related work [05] which also uses Yahoo! (though Yahoo! Japan may differ a little) top level categories for the evaluation of their solution over 5 classes against ours of 6 classes. We selected [5] to benchmark our work against due to the similar application of our classifier and real-world source of testing. The comparison took their best results with the lowest error rate over 5 classes is shown in Table 5.

From Table 5, it could be seen that our solution on average outperforms the solution from research [05] that applies decision tress on “basket analysis” conducted on web pages.

Table 5. Comparison of results with [05]

	Results [05]	Our Results	Difference
Highest Precision	87.20 %	100 %	+ 22.80
Lowest Precision	67.60 %	57.14 %	- 10.46
Average Precision	79.36 %	90.08 %	+ 10.72
Highest Recall	69.00 %	100 %	+ 31.00
Lowest Recall	45.50 %	55.56 %	+ 10.06
Average Recall	57.30 %	73.97 %	+ 16.67

5. Future Work

There is still room for various improvements to the framework. One of the possible extensions to the framework would be applying machine learning classifier again on the classification results of each web page feature instead of weights obtained from evaluation of the training classifier. This we believe would have the potential to increase the performance of our classifier when dealing with a large set of classes/ categories/ genres to be classified or when we look further into other features of web pages.

Besides the web page features mentioned in this paper, we would explore other possible features that could provide valuable information and hints on the classification of the web page. One of the identified features would be through headings of the web page. If a feature is deemed to be of low significant, it would be reflected by its weights and on cases where 2 classifications is close for a web page, it could provide the small deciding push to the classification of the web page.

Alternatively, we could explore the network nature of the web and how can the classification of a web page be affected by both its inward and outward links. This would include hyperlinks and hyperlink texts which would be easily found via their appropriate tags.

We would also work on the dimension reduction of our solution as currently we are having over 80,000 attributes (or words in our bag of words) to reduce the complexity of our solution and increase the performance of our solution. The complexity freed up could be used on other areas to increase the performance such as increasing web page features or training web pages.

Furthermore, we would look at the time performance as well as the complexity of our solution which was not measured at the moment. Currently, we believe that our solution is computably feasible within reasonable range from the parsing of web pages to the training and classification of the solution.

6. Conclusion

In summary, the proposed framework of automatic web page classification is highly promising from the results of our evaluation in comparison with that of manual classification from editors. With an average accuracy of 90.08%, the automatic classification of webpages could be used to build taxonomies automatically that would help users locate their required information on the web.

The selected web page features do store valuable information about the web page in the classification of web pages. Classification of these features separately and jointly deciding the classification of the web page is highly successful as opposed to the approach of text summarization or buildinga virtual web page.

We believe that the research into web page classification is highly promising with the advent of the web and web 2.0. Successful classification of web pages especially through automation is highly rewarding in various fields especially those of data mining, search engines and so forth.

7. References

[01] Netcraft (2010), “May 2010 Web Server Survey”. [Online].Netcraft Ltd [Retrieved May 17, 2010], http://news.netcraft.com/archives/2010/05/14/may_2010_web_server_survey.html

[02] Chekuri, C. and Goldwasser, M.H. (1996). “Web Search Using Automatic Classification” [Online].Stanford University [Retrieved July 22 2009], http://theory.stanford.edu/people/wass/publications/Web_Search/Web_Search.html.

[03] Qi, X. and B. D. Davison (2009). "Web page classification: Features and algorithms." ACM Computing Surveys (CSUR) **41**(2): Article No.: 12.

[04] Dumais, S. and H. Chen (2000). "Hierarchical classification of Web content." Annual ACM Conference on Research and Development in Information Retrieval in the Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval: 256 - 263 .

[05] Tsukada, M., T. Washio, et al. (2001). "Automatic Web-Page Classification by Using Machine Learning Methods." Lecture Notes In Computer Science in the Proceedings of the First Asia-Pacific Conference on Web Intelligence: Research and Development **2198**: 303 - 313

[06] Shen, D., Q. Yang, et al. (2007). "Noise reduction through summarization for Web-page classification." Information Processing and Management: an International Journal **43**(6): 1735-1747

[07] Choi, B. and Yao, Z. (2005). “Web Page Classification: Features and Algorithms”, Book chapter in “Foundations and Advances in Data Mining”, Springer-Verlag.

[08] Indra Devi, M., Rajaram, R., &Selvakuberan, K. (2008), “Generating beat features for web page classification”, Webology, 5(1), Article 52. Available at: http://www.webology.ir/2008/v5n1/a52.html.

[09] Tate, M. and J. Alexander (1996). "Teaching critical evaluation skills for World Wide Web resources." Computers in Libraries **16**(10): 49 – 55

[10] Kan, M.-Y. and H. O. N. Thi (2005). "Fast webpage classification using URL features." Conference on Information and Knowledge Management in the Proceedings of the 14th ACM international conference on Information and knowledge management: 325 - 326

[11] Shih, L. K. and D. R. Karger (2004). "Using urls and table layout for web classification tasks." International World Wide Web Conference in the Proceedings of the 13th international conference on World Wide Web: 193 - 202

[12] Rish, I. (2001). "An Empirical Study of the Naïve bayes Classifier". In: Proceedings of IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence

[13] Cortes, C. and V. Vapnik (1995). "Support-Vector Networks." Machine Learning **20**(3): 273 - 297.

[14] Salzberg, S. L. (1994). "Book Review: C4.5: Programs for Machine Learning by J. Ross Quinlan, Morgan Kaufmann Publisersinc. 1993". Kluwer AvademicPublisers.

[15] Bo, S., S. Qiurui, et al. (2009). "A Study on Automatic Web Pages Categorization." 1423 - 1427.

[16] Chen, G. and B. Choi (2008). "Web Page Genre Classification." Symposium on Applied Computing in the Proceedings of the 2008 ACM symposium on Applied computing: 2353-2357