International Journal of Computer Engineering
and Technology (IJCET), ISSN 0976 – 6367(Print)
ISSN 0976 – 6375(Online) Volume 1
Number 1, May - June (2010), pp. 112-122

IJCET
© I A E M E

RECENT RESEARCH IN WEB PAGE CLASSIFICATION – A REVIEW

Alamelu Mangai J & Santhosh Kumar V
Faculty of Computer Science and Engineering
BITS, Pilani, Dubai, U. A. E
E-mail id: mangaivpm@yahoo.com

Sugumaran V
Department of Mechatronics, SRM University
Kattankulathur, Kancheepuram Dt.
E-mail id: v_sugu@yahoo.com

ABSTRACT

As the volume of web pages given by a search engine in response to a user query is huge, automatic classification of web pages into relevant categories has become the state of the art research topic. Web page classification algorithms have additional challenges since web pages have structured, semi-structured and also unstructured data. In this article, a survey of the different approaches to web page classification, its applications and some of the associated issues are presented.

Key Terms – Web page classification, Web mining, Feature selection

1. INTRODUCTION

Over the past decade, the world has witnessed an explosive growth on the internet, with millions of web pages on every topic easily accessible through the web. The internet is a powerful medium for communication between computers and for accessing online documents all over the world; however, it is not a tool for locating or organizing the massive information. Tools like search engines assist the users in locating information on the internet. The performance of most of the search engines in locating the information is excellent. However, its ability to organize the web pages are limited. Internet users are now confronted with thousands of web pages returned by a search engine using simple keyword search. Searching through these web pages itself becoming

a more difficult task for users [1]. Hence, users are interested in tools that can help make a relevant and quick selection of information that is needed. It is also believed that the actual size of the web is at least several times bigger than what search engines currently cover. Describing and organizing the vast amount of content is essential for realizing the webs full potential as an information resource. Hence the automation of web page classification is necessary. This helps in focused crawling, to the assisted development of web directories, to topic specific web link analysis, and to analysis of the topical structure of the web. It also improves the quality of web search.

2. WEB PAGE CLASSIFICATION

Web page classification (WPC) also known as web page categorization, is the process of assigning a web page to one or more predefined category labels. Classification is often proposed as a supervised learning problem, in which a set of labeled data is used to train a classifier which can be applied to label future unseen examples.

2.1 ARCHITECTURE OF A WEB PAGE CLASSIFICATION SYSTEM

A WPC system involves the different modules as shown in Figure 1. As the web pages contain many irrelevant words and stop words that reduce the performance of the classifier, extracting relevant features and selecting the representative features from the web page is an essential pre-processing step.

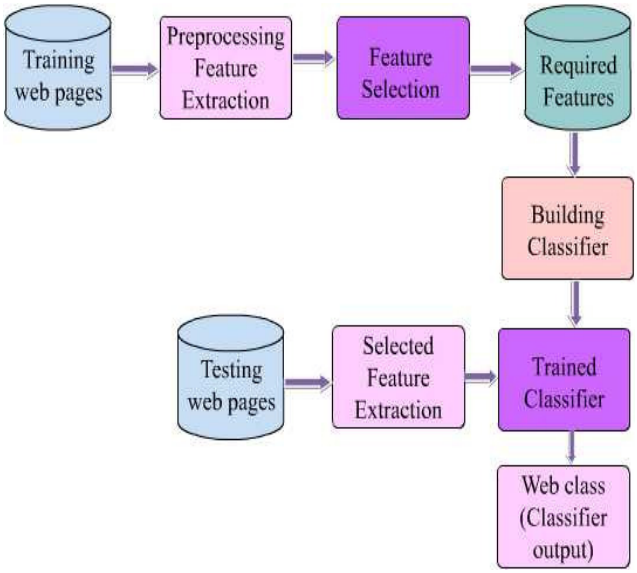

Figure 1 Architecture of a Web Page Classification System

Feature selection methods are of two categories namely, filter and wrapper methods. The wrapper methods use a learning algorithm to determine the best features. The filter models use the general characteristics of the training data to select the best features. At times, two or more feature selection techniques are combined to give better performance. For instance, the cfssubset evaluator combined with term frequency gives minimal qualitative features to attain considerable classification accuracy [2]. Chen *et. al* [3] has proposed a fuzzy ranking analysis paradigm together with a novel relevance measure, discriminating power measure(DPM), to effectively reduce the input dimensionality (number of web features) from tens of thousands to a few thousands with zero rejection rate and small decrease in accuracy. A web page classifier faces a huge scale dimensionality problem, with tens of thousands of features and hundreds of categories. Hence, reducing the dimensionality is a critically important challenge for web page classifiers.

2.2 APPROACHES TO WPC

The general problem of web page classification can be divided into multiple sub problems viz., subject classification, functional classification, sentiment classification and other types of classification [4]. Subject classification is concerned about the subject or topic of a web page. For example, judging whether a page is about arts, business or sports is an instance of subject classification. Functional classification cares about the role that the web page plays. For example, deciding a page to be a personal home page, course page or admission page is an instance of functional classification. Sentiment classification focuses on the opinion that is presented in a web page, *i.e.,* the authors attitude about some particular topic. Other types of classification include genre classification [5], search engine spam classification [6] and so on.

Based on the number of classes in the problem, classification can be divided into binary and multi-class classification, where binary classification categorizes instances into exactly one of two classes; multi-class classification deals with more than two classes. Based on the number of classes that can be assigned to an instance, classification can be divided into single-label classification and multi-label classification. In single-label classification, one and only one class label is to be assigned to each instance, while

more than one class labels can be assigned to an instance in multi-label classification. Based on the type of class assignment, classification can be divided into hard classification and soft classification. In hard classification, an instance can be assigned to only one class and there is no intermediate state. While in soft classification, an instance can be predicted to be in some class with some likelihood.

Based on the organization of categories, web page classification can also be divided into flat classification and hierarchical classification. In flat classification, categories are considered parallel, *i.e.*, one category does not supersede another. While in hierarchical classification, the categories are organized in a hierarchical tree-like structure, in which each category may have a number of subcategories.

2.3 WEB PAGE CLASSIFICATION ALGORITHMS

Different algorithms have been adopted to classify web pages. Shanks *et. al* [7] have used the first fragment of each document to classify news articles. This approach is based on the assumption that a summary is present at the beginning of each document, which is true only for news articles, this approach was later applied to hierarchical classification of web pages by Wibowo *et.al* [8].

Since web pages can be considered as instances which are connected by hyperlink relations, web page classification is solved as a relational learning problem. Relaxation labeling is one of the algorithms that work well in web page classification. Chakrabarti *et. al* [9], states that "In the context of hypertext classification, the relaxation labeling algorithm first uses a text classifier to assign the class probabilities to each node(page). Then it considers each page in turn and reevaluates its class probabilities in the light of the latest estimates of the class probabilities of its neighbors". Another variation was proposed by Angelova *et.al* [10] where not all neighbors are considered. Instead, only neighbors that are similar enough in content are used.

Besides relaxation labeling, other relational learning algorithms are also applied to web page classification. Sen and Getoor [11] compared and analyzed relaxation labeling along with two other popular link-based classification algorithms: loopy belief propagation and iterative classification. Their performance on a web collection is better than text classifiers. Macskassy and Provost [12] implemented a tool kit for classifying

networked data, which utilized a collective inference procedure for a relational classifier and demonstrated its powerful performance on several data sets including web collections.

It is also observed in literature that efforts are made to tweak the traditional algorithms, such as k-Nearest Neighbor (kNN) and Support Vector Machines (SVM), in the context of web classification. k-Nearest Neighbor classifiers require a document dissimilarity measure to quantify the distance between a test document and each training document. Most existing kNN classifiers use cosine similarity or inner product. Based on the observation that such measures cannot take advantage of the association between terms, Lee [13], developed an improved similarity measure that takes into account the term co-occurrence in documents. The intuition is that frequently co-occuring terms constrain the semantic concept of each other. The more co-occurred terms two documents have in common, the stronger the relationship between the two documents. Their experiments proved performance improvements over cosine similarity and inner product measures.

Co-training, introduced by Blum and Mitchell [14], is an approach that makes use of both labeled and unlabeled data to achieve better accuracy. In a binary classification scenario, two classifiers that are trained on different sets of features are used to classify the unlabeled instances. The prediction of each classifier is used to train the other. Compared with the approach which uses only the labeled data, this co-training approach is able to cut the error rate by half. Ghani [15] generalized this approach to multi-class problems. The results showed no improvement in accuracy when there are large number of categories. They proposed a method which combines error-correcting output coding (a technique to improve multi-class classification performance by using more than enough classifiers) with co-training which was able to boost the performance. Park and Zhang [16] also applied co-training in web page classification which considers both content and syntactic information. Classification usually requires manually labeled positive and negative examples. Yu *et al*. [17], devised an one class SVM to eliminate the need for manual collection of negative examples while still retaining similar classification accuracy. Given positive data and unlabeled data, their algorithm is able to identify the most important positive features. Using these positive features, it filters out possible

positive examples from the unlabeled data, which leaves only negative examples. An SVM classifier is then trained on the labeled positive examples and the filtered negative examples.

Based on classical "divide and conquer", Dumais and Chen [18] suggested the use of hierarchical structure for web page classification. It is demonstrated in their paper that splitting the classification problem into a number of sub-problems at each level of the hierarchy is more efficient and accurate than classifying in the non-hierarchical way.

Research in blog classification can be divided into three types: blog identification (to determine whether a web document is a blog), mood classification and genre classification. Research in first category aims at identifying blog pages from a collection of web pages, which is essentially a binary classification of blog and non-blog. Nano *et al*. [19] presented a system that automatically collects and monitors blog collections, identifying blog pages based on a number of simple heuristics. Elgersma and Rijke [20] examined the effectiveness of common classification algorithms on blog identification tasks. Using a number of human –selected features (some of which are blog-specific, *e.g*., whether characteristic terms are present, such as "comments" and "archives"), they found that many off-the-shelf machine learning algorithms can yield satisfactory classification accuracy.

Research in mood classification includes identification of the mood or sentiment of blogs. Michalcea and Liu [21] showed that blog entries expressing the two polarities of moods, happiness and sadness, are separable by their linguistic content. A naïve Bayes classifier trained on unigram features achieved 79% accuracy over 10,000 mood-annotated blog posts. Similarly Chesley *et al*. [22] demonstrated encouraging performance in categorizing blog posts into three sentiment classes (Objective, Positive and Negative). However, real world blog posts indicate moods much more complicated than merely happiness and sadness (or positive and negative). Classifying blog posts into a more comprehensive set of moods is a challenging task.

Research in genre classification focuses on the genre of blogs and is usually done at blog level. Nowson [23] discussed the distinction of three types of blogs: news, commentary and journal. Qu *et al*. [24] proposed an approach to automatic classification of blogs into four genres: personal diary, news, political and sports. Using unigram pdf

document representation and Naïve Bayes classification, Qu *et al*'s approach has achieved an accuracy of 84%.

Using swarm intelligence to classify web pages is a new trend in this direction. Moayed *et. al* [25] investigated usage of a swarm intelligence algorithm in the field of the web page classification. Ant Miner II is the proposed algorithm focusing on Persian web pages. A simple text preprocessing technique to reduce the large numbers of attributes associated with web content mining, without dealing linguistic complications is also proposed. The results have shown that Ant Miner II and their proposed preprocessing technique are efficient in the field of web page classification.

Contextual advertising seeks to place relevant advertisements to generic web pages based on their contents. Recently, it is observed that classifying web pages into a well–organized taxonomy of topics is promising for matching topically relevant ads to web pages. Following the observation, Lee *et al*. [26] proposed two methods to increase classification accuracy for web pages in the context of contextual advertising. Their strategy is to enhance the baseline classifier by reflecting unique features of web pages and the taxonomy. In particular, category tags extracted from web pages are utilized to augment term weights, and the hierarchical structure of the taxonomy is taken into account to categorize web pages with high confidence.

3. APPLICATIONS OF WPC

The content of a web page is useful for many information retrieval tasks. WPC is mainly used in constructing, maintaining or expanding web directories. It is used to improve the quality of the search results, since query ambiguity is among the problems that undermine the quality of search results [27]. WPC also helps in question answering systems. A question answering system uses a classification technique to improve its quality of answers. For building focused crawlers WPC is used. By estimating the relevance of a retrieved web page to the given topic, it is possible to fix the crawl boundary.

4. CONCLUSION

Classification of web page content is essential to many tasks in web information retrieval such as maintaining web directories and focused crawling. Web pages are

dynamic and volatile in nature. There is no unique format for the web pages. Some web pages may be unstructured (full text), some pages may be semi-structured (HTML pages) and some pages may be structured (databases). Hence, the uncontrolled nature of web content presents additional challenges to web page classification than traditional text classification. Most web classification work found in literature focuses on hard classification with a single label per document. However, multi-label and soft classification better represent the real world documents which are rarely represented by a single predefined topic. Also, complexity of evaluation and a lack of appropriate data sets (especially for mood and genre classification) has prevented straightforward progress in these areas. These issues have motivated further research in this area.

5. REFERENCES

[1] Daume III, H., & Brill, E., Web search intent induction via search results partitioning. Proceedings of HLT, 2004.

[2] Indra Devi M., Rajaram R., Selvakuberan K. Generating best features for web page classification, Webology,. March 2008. Vol 5, No: 1.

[3] Chih-Ming Chen, Hahn-Ming Lee and Yu-Jung Chang, Two novel feature selection approaches for web page classification, Expert Systems with Applications, ScienceDirect, 2009 vol 36, Issue 1, pp: 260-272.

[4] Xiaoguang Qi and Brian D.Davison. Web Page Classification: Features and Algorithms, Technical Report. Dept. of Computer Science and Engineering, Lehigh University, June 2007.

[5] zu Eissen, S. M. and B. Stein ,Genre classification of web pages, In Proceedings of the 27th German Conference on Artificial Intelligence, Berlin, 2004 ,Sep 20-24, Springer, 2004,Volume 3238 of LNCS, pp. 256–269.

[6] Castillo, C., D. Donato, A. Gionis, V. Murdock, and F. Silvestri .Know your neighbors: Web spam detection using the web topology. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, 2007, July 23-27, ACM 2007 .pp: 423-430.

[7] Shanks, V. and H. E. Williams. Fast categorizations of large document collections. In Proceedings of Eighth International Symposium on String Processing and Information Retrieval (SPIRE), Chile, 2001, Nov 13-15, pp. 194–204.

[8] Wibowo, W. and H. E.Williams. Simple and accurate feature selection for hierarchical categorization. In DocEng '02: Proceedings of the 2002 ACM Symposium on Document Engineering, Virginia, 2002, Nov 8-9, ACM 2002, pp. 111–118.

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

[10] Angelova, R. and G. Weikum Graph-based text classification: Learn from your neighbors. In SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington, 2006, Aug 6-11, ACM 2006. pp. 485–492.

[11] Sen, P. and L. Getoor Link-based classification. Technical Report CS-TR-4858, University of Maryland. 2007.

[12] Macskassy, S. A. and F. Provost, Classification in networked data: A toolkit and a univariate case study. Journal of Machine Learning Research, May 2007., Vol 8 pp: 935–983.

[13] Press. Kwon, O.-W. and J.-H. Lee, Web page classification based on k-nearest neighbor approach. In IRAL '00: Proceedings of the 5th International Workshop on Information Retrieval with Asian languages, Hong Kong 2000, Sep 30 – Oct 2, ACM 2000, pp. 9–15.

[14] Blum, A. and T. Mitchell, Combining labeled and unlabeled data with co-training. In COLT' 98: Proceedings of the 11th Annual Conference on Computational Learning Theory, New York, 1998, July 24-26, ACM Press, 1998. pp. 92–100.

[15] Ghani, R. Combining labeled and unlabeled data for multiclass text categorization. In ICML '02: Proceedings of the 19th International Conference on Machine Learning, Sydney, 2002, July 8-12, Morgan Kaufmann 2002. pp. 187–194.

[16] Park, S.-B. and B.-T. Zhang (2003, April). Large scale unstructured document classification using unlabeled data and syntactic information. In Advances in

Knowledge Discovery and Data Mining: 7th Pacific-Asia Conference (PAKDD), Korea, 2003, Apr 30 – May 2, Springer 2003. Volume 2637 of LNCS, pp. 88–99.

[17] Yu, H., J. Han, and K. C.-C. Chang (2004). PEBL: Web page classification without negative examples. IEEE Transactions on Knowledge and Data Engineering 2004. Vol 16 (1), pp: 70–81.

[18] Dumais, S. and H. Chen, Hierarchical classification of web content. In SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Greece, 2000, July 2, ACM pp. 256–263.

[19] Nanno, T., T. Fujiki, Y. Suzuki, and M. Okumura, Automatically collecting, monitoring, and mining Japanese weblogs. In WWW Alt. '04: Proceedings of the 13th international World Wide Web Conference on Alternate Track Papers & Posters, New York, 2004, May 17-20, ACM 2004. pp. 320–321.

[20] Elgersma, E. and M. de Rijke. Learning to recognize blogs: A preliminary exploration. In EACL 2006 Workshop: New Text - Wikis and blogs and other dynamic text sources., Italy, April 4 , 2006.

[21] Mihalcea, R. and H. Liu. A corpus-based approach to finding happiness. In N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin (Eds.), Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium, Menlo Park, CA, AAAI Press. Technical Report SS-06-03. March 2006. , pp. 139–144

[22] Chesley, P., B. Vincent, L. Xu, and R. K. Srihari Using verbs and adjectives to automatically classify blog sentiment. In N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin (Eds.), Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium, Menlo Park, CA, pp. 27–29. AAAI Press. Technical Report SS-06-03. March 2006.

[23] Nowson, S. The Language of Weblogs: A study of genre and individual differences. Ph. D. thesis, University of Edinburgh, College of Science and Engineering. June 2006.

[24] Qu, H., A. L. Pietra, and S. Poon, Automated blog classification: Challenges and pitfalls. In N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin (Eds.),

Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium, Menlo Park, CA, AAAI Press. Technical Report SS-06-03. March 2006. pp. 184–186.

[25] Moayed M. J., Sabery A. H, Khanteymoory. A., Ant-colony algorithm for web page classification, Proceedings of ITSim 2008, Int'l Symposium on Information Technology, Kaula Lumpur, 2008, Aug 26-28, 2008, Vol 3, pp: 1-8.

[26] Jung-Jin Lee, Jung-Hyun Lee, Jongwoo_Ha, Sang Keun Lee, Novel web page classification techniques in contextual advertising, Proceedings of the 11th International Workshop on Web Information and data management, China, 2009, Nov 02, ACM 2009, pp: 39-47.