Web Page Advertisement Classification

Kankana Shukla and Ben Choi

***Abstract*—In this paper we present an automatic genre-based web page classification system for determining whether a web page contains an advertisement or not. Due to the difficulties and subjectivities in defining a genre, its features and their categorization, genre based classification is still rudimentary. In this research, we identified key features and used those features to define the advertisement category. We then developed a genre-based classification system to automatically classify a web page into advertisement category, which is important for commerce and for web users who prefer to either view or skip advertisements. We implemented and tested the proposed system, which achieved an average accuracy of 82%. Furthermore, we incorporated this system with other genre and subject based system to create a comprehensive web page classification system.**

***Index Terms*—Information retrieval, knowledge classification, semantic web, web mining, web ontology.**

I. Introduction

Over the past decades, there has been a tremendous growth in the internet and the World Wide Web. The web users have increased exponentially resulting in the extreme commercialization of the web. Presently, online advertising has been a major source of revenue for businesses. With the growth of advertisements on the web, came the stage of it's over exploitation causing various problems to the users. Thus, classification of web pages into advertisement plays an important role, both, for users who are interested to view them, as well as for users who want to skip them. In this paper, we focus on classifying web pages into advertisement category.

Thus, we created a genre-based classification system to automatically classify web pages into advertisement category. In general, there are two types of web page classification systems: the subject-based and the genre-based systems. As the name suggests, in subject-based classification (also known as topic-based classification), web pages are classified according to their subjects or contents[1]-[5]Moreover, genre-based classification system[6]-[10]focus mainly on the structure or format of the document, the purpose of the web page, and the intended audience, which are collectively called as the genre related factors.

Genre based classifications are sporadically used, the main reason being that the genre of the Web pages is highly subjective. Divergence in comprehension of the genre of a Web page makes it challenging for an investigator to classify it into the correct genre. Typical viable concepts for genre–based classification are unsatisfactory to befit the search requirements of the user and are therefore still under development. An exemplification being [6], where only three categories are defined, making it complicated for users to determine an appropriate category for exacting their search requirements. While [10] concentrates primarily on organizational member's communication actions like the business or technical report. This approach is impertinent to cater every user's needs.

This paper defines "genre" as the web page's functional purpose provided to web users. Advertisement is the content of web pages whose purpose is to publish information for an advertiser and help attract consumers and increase productivity. The notion behind defining genre in this way is that we care about users' multi-facet search demands for the web pages.

Based upon the demands from businesses to identify advertisements, in this paper we focus on creating a genre-based classification system to automatically classify web pages into advertisement category. We extracted new key features from web pages and used those features to define the category. The entire feature set will be introduced in detail in Section III. In addition, we proposed new process to associate features to the category. For the results, we demonstrate our proposed system achieving an average classification accuracy of 82%, which ascertains that our system is viable and forms the background for future research.

Creating an advertisement category proves itself to be a challenging research problem. Prior works on subject-based [4] and in genre-based systems [11] have not addressed this issue. The two types of classification systems are presently not able to process the picture content on web pages, however many advertisements consist of pictures only. Thus, the classification systems, including the one proposed in this paper, have an equal likelihood of categorizing such web pages as advertisements. We also perceive that it is difficult to separate a product review from an advertisement, since both of which contain similar information about the product. The work presented in this paper forms an initial attempt to address these research problems and much research remain to be done.

The rest of this paper is organized into the following sections. Section II outlines the related research. Section III provides details of our genre classification system for advertisement, which provides methods for feature extraction from web pages, methods for association of features to the category, and methods for genre classification of web pages. Section IV provides implementation and test results. Finally, Section V gives the conclusion and outlines the future research.

II. Related Research

To identify the genre of a Web page, its structural information can play a key role. The Web pages having a

Manuscript received April 9, 2013; revised July 17, 2013.

The authors are with the Computer Science Department atthe Louisiana Tech University, Ruston, LA 71272, USA (e-mail: KankanaShukla.edu@gmail.com, pro@BenChoi.org).

DOI: 10.7763/IJCEE.2014.V6.793

"common feature" are classified into the respective category. The features affecting the structure include but are not limited to textual features, such as the number and placement of links and logos, image features, such as the distinct colors of the image, and other features, such as a flash video or other multimedia contents[6].Research in [6]uses a structure based approach which consists of only three categories defined in the paper: information pages, research pages, and personal home pages. The technique in [12] is profoundly similar to [6], both classify Web pages are according to their structural characteristics. The categories in [12] are online shopping, product catalog, advertisement; call for paper, links, frequently asked questions, glossary, home page, and bulletin board. The features used in the paper include URL, keyword, image, link, OCR, structure, and plug in [12].

Conjointly, another denomination of genre used is the communicative actions of Web pages. Paper [10] defines genre as "a type of communication recognized and enacted by organizational members". They presented six questions: why(purpose of communicative action),what(contents of a genre or genre system),who/whom(participants in genre or genre system),when(timing of genre or genre system use),where (place of communicative action),and how (the form of genre and genre system)[10]. They describe a category by answering these questions.

In [7], a genre is defined in terms of dimensions such as the degree of expertise, the amount of detail presented and whether it reports facts and/or opinions. The expertise dimension is estimated as a function of the frequency and length of words in the document [7]. The estimation of detailed dimension is performed as a function of the document's physical size, number of lines, and the frequency of long words [7]. Superficial linguistic features such as the part-of-speech tags are used to determine the subjective dimension. When users care about the degree of expertise, these dimension shave proved to be very advantageous.

Furthermore, some hybrid techniques have also been developed that conjoin multiple concepts. For instance, [13]classifies Web pages based upon the characteristics such as purpose or function of the page, its intended audience, its surface content or format (e.g. words, tables, sounds, tools, etc.), the type of links it contains, and its relationship to the pages to which it provides the links (e.g. cover page, index, etc.) [13].The categories defined in the paper accommodate organizational pages, documentation, text, homepage, multimedia, database entry, and tools.

III. Our Genre Classification System for Advertisement

Internet and the World Wide Web are emerging and ever-expanding fields aiding immediate publishing of content through Online Advertising from any part of the world to the other in virtually no time. It has proved to be an effective source to bring in customers for the advertisers in the recent past. On the other hand, it has also provided the authority to the Web users to examine the advertisement or product or not. The function of an online advertisement is very clear- to show product to consumer, ask them if they want to check out the product further, provide them more information (if unless desired) and to let them place the order if they desire. Though advertisements come in disparate formats, to achieve all the functions, typical online advertisement consists of several common properties. For instance, many of them may have price information and clickable images or flash videos which can link to a specific product page; some advertisements may be contained within pictures with no textual information. There frequently exist some clickable buttons which enables the users to be redirected to another website that provides more information about the product in the advertisement. We utilize these common traits to distinguish a normal Web page from one with advertisements.

In this section, we propose a new classification system based on web page genre and focused on advertisement category. We define features that are to be extracted from Web pages. Then, we describe how to use these features to specify the advertisement category. We finally describe how to assign a category to a Web page.

A. Feature Extraction from Web Pages

In order to extract features pertaining to online advertisements from Web pages to identify the category, we analyze not only the contents of the Web pages, but also the URLs, HTML tags, Java scripts, and VB scripts. The principles used to select features include, but are not limited to: (1) The features should positively contribute to the classifier model. (2) The features should be detectable; (3) The computational cost of detecting each feature should be modest. (4)The features should follow the idea of Web page genre classification, and not involve the features that used for subject-based approach. This principle proves to be more favorable when we combine the concepts of genre classification with subject-based approach.

TABLE I. Weights of Each HTML File Feature in the Category

No	Feature Name	Search Key	Weights
1	Get Free	free	0.2
2	Save	save	1
3	Graphical Interchange	.gif	0.5
4	Learn More	more	0.4
5	Price Off	off	0.5
6	Play	play	0.2
7	Image	img	0.3
8	Click here	click	0.5
9	Now	now	0.5
10	Google	google	0.5
11	Right box	right	0.5
12	Information	info	0.6
13	Price in $	$	0.7
14	Pop up	pop	0.7
15	Off in %	%	0.2
16	Company Logo	logo	0.9
17	Try or Trial	try	0.9
18	Flashy Color	color	0.4
19	Price Only	only	0.9
20	Buttons	button	0.8
21	Do Online	online	0.2
22	Advertisement	ad	0.8
23	Do here	here	0.6
24	Start	start	0.5
25	Enter	enter	0.7
26	Contents	content	0.3
27	Search	search	0.4
28	Find	fund	0.8

Following these principles, we determined the whole

feature set by gathering hundreds of Web pages of discrete categories and analyzing the HTML files, and the embedded scripts. We developed the model and extracted28 features as provided in Table I, the details of which will be provided in the following subsections.

B. Association of Features to the Category

To build a classification model for genre-based system, we first need to determine how to specify each category using the given features. There are two major stages: (1) estimation of feature weights and and (2) fine tuning the feature weights that define the category.

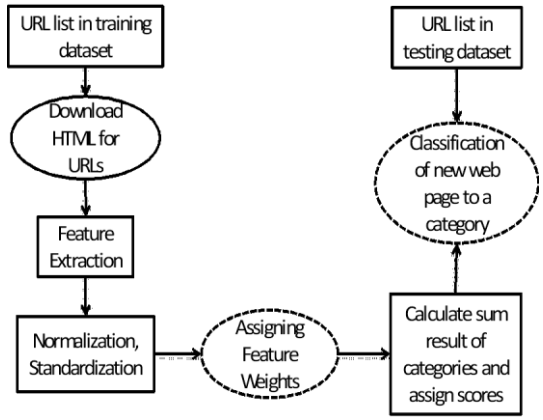

Fig. 1. Estimation of feature weights

The first stage consists of finding the approximate value of parameters including feature weights. The process in this stage is shown in Fig. 1. First, the URLs of the Web pages to be used as the training dataset are selected and stored in a file. Then we download the HTML code for the Web page from the URL content. The HTML files are then passed through a HTML feature extraction module, by which each feature's occurrence frequency in each category is calculated. The initial value of the feature weight in the advertisement category was defined by the frequency of occurrence of the feature in that category. The formula to calculate the initial weights is:

$$Weight(C_i, F_i) = \frac{2 \times Freq\ (C_i, F_i) - N\ (C_i)}{N(C_i)}$$

- Weight (C_i, F_j) is the feature Fj'sweight in category C_i.
- Freq (C_i, F_j) is the occurrence frequency of the feature F_j in the category C_i.
- N (C_i) is the number of Web pages of category C_i in the training set.

The fine tuning of the initial values of threshold and features weights produced from the first stage is performed to associate features to the category of advertisements. Fig. 2 shows the process. The URLs of Web pages in training dataset are first provided to both the Web page HTML download module and the feature extraction module. The Web page download module takes the URL and downloads the HTML file and then passes the file to a HTML feature extraction module, which analyzes HTML file and detects the frequencies of occurrence of each feature. The HTML feature frequencies are then passed through the sum generation module, which receives the feature weights corresponding to the advertisement category. Two tasks are performed in this module- to calculate the sums of the feature weights in each category, and to transmit it to the result processing module. In the course of result processing, the data is normalized, and is compared to the normalized score to the threshold. If the score is not higher the threshold, then the Web page will be considered not belonging advertisement categories and labeled as "other"; otherwise the page will be classified to the advertisement category.

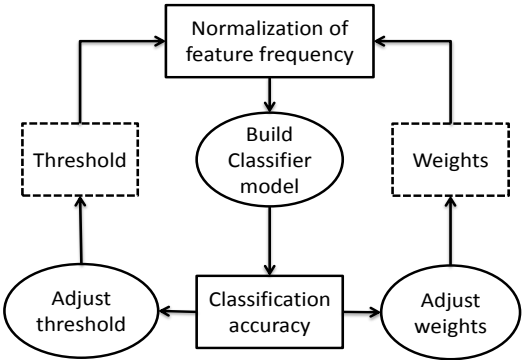

Fig. 2. Fine tuning feature weights and threshold

We then evaluate the final classifier performance and adjust the threshold and features weights to optimize the classification system. The weights initially are all ranging from 0 to 1.For the fine tuning, we choose one key feature from each category and increase the weight to 0.2. This is based on an assumption that most of the categories have their unique feature that may represent themselves best and distinguish from other categories. Then we ran the classification system on the training dataset to check the likely value. The classification results were checked, and the errors were analyzed and used as the feedback to modify the weights again. It is necessary to repeat the whole fine tuning process many times in order to achieve the best performance. The resulting weights for one our experiments are shown in Table I.

Besides fine tuning the feature weights, we also need to fine tune the threshold, which is used to distinguish whether a Web page belongs to the defined genre categories or not. The value of the threshold is very important since it will significantly influence the performance of the classification system. We use the precision, recall, and F-measure to describe the performance [4]:

$$Precision = \frac{a}{a+b}$$

$$Recall = \frac{a}{a+c}$$

a: the number of testing examples correctly assigned to thecategory; b: the number of testing examples incorrectly assigned to thecategory; c: the number of testing examples incorrectly rejected to the category

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

For our experiments, when the value of the threshold

increases, the precision will increase, but the recall will decrease. To keep a balance between the precision and recall, the threshold cannot be too high or too low. F-measure combines precision and recall, and allow us to keep a balance between them by adjusting the threshold to maximize the value of F-measure. For our training dataset, the best performance happens when the threshold is set to 0.55, using whichthe average of the F-measure is high, and the distributing of the F-measure does not result in any category having a significantly low F-measure.

C. Genre Classification of Web Pages

After features are associated to the advertisement category and the feature weights and the threshold are fine-tuned, our system presents itself to be equipped to classify new Web pages. The classification process is shown in Fig. 3. This process is similar to the fine tuning process except we do not modify the weights and the threshold. The URL of the Web page to be classified is first given. The features of the Web page is extracted and weighted. The total weight for each category is calculated and the highest one is selected. The highest weight is then normalized by the size of the HTML file of the Web page. If this normalized weight is larger than the threshold, then the Web page is classified to the category that has the high weight; otherwise it is considered not belonging to any defined genre categories.

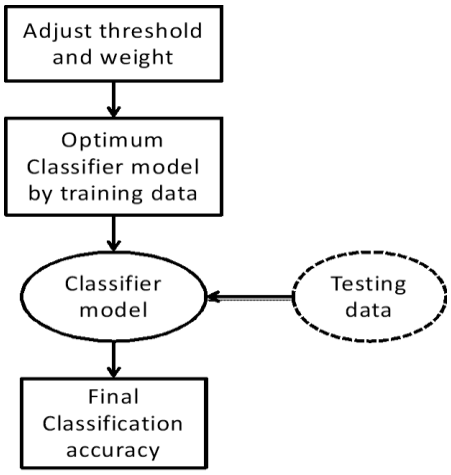

Fig. 3. Process for classifying new web pages

IV. Implementation and Test Results

We implemented and tested our proposed genre-based classification system for advertisement using Mathematica. First, we collected a set of training web pages from the Internet. These training web pages represent a diversify ranges of web pages that contain advertisements. We analyzed the training web pages in details, and extracted 28 key features that signify advertisements.

Having the set of features, we then create a feature vector that defines the advertisement category. This is done by (1) retrieve all HTML codes of the training web page, for which we use the Mathematica function: Import[url, "Source"]. (2) We analyzed the HTML codes to determine the feature counts, for which we use the following Mathematica code:

$$StringCount[html, \#, IgnoreCase \rightarrow True]\&/@features$$

where, "html" is the codes for one of the training web pages, "features" is a set of 28 search key (see Table 1). The above code creates a feature vector for a training web page. After processing feature vectors for all the training web pages, we then normalized all the vectors to create a single feature vector that represent the advertisement category.

During the training process, a weight vector is created and multiplied with the feature vector to create a weighted feature vector, which represent the advertisement category as the result of fine turning the weight vector, as introduced in the last section. The resultant weight vector is show in the "Weights" column in Table 1.

During our testing, we collected a new set of web pages of diversify subjects. For each of the test web pages, we generated a test feature vector using the Mathematica code as discussed above. This test feature vector is compared with the weighted feature vector of the category using the cosine similarity measure, which is defined as:

$$\text{cosSim} := \text{Function}[\{u, v\},$$

$$\sum_{i=1}^{28} (u[[i]] \times v[[i]]) \Big/ \left(\sqrt{\sum_{i=1}^{28} (u[[i]])\text{^}2} \times \sqrt{\sum_{i=1}^{28} (v[[i]])\text{^}2} \right)]$$

If the value of the cosine similarity is greater than a predefined threshold, then the corresponding test web page is classified into the category. The threshold is determined (as discussed in the last section) during the training process and is 0.55 for our experiments.

The classification result for each test web page is checked by visiting the web page to verify its contents. The overall result of our test can be summarized as achieving an average accuracy of 82%. Pertaining to the results, we formulate the conclusion that the proposed approach is viable and has a scope for future developments.

V. Conclusion and Future Research

This paper proposes a new automatic genre-based web page classification system, which focuses on the advertisement category. New features to identify advertisement web page and new methods to define the category are introduced. The proposed system can achieve reasonable average classification accuracy, which provides the groundwork for future research. We incorporated this system with our other genre [11] and subject based [1]-[5]system to create a comprehensive web page classification system[2], [4], [14].

To achieve the high accuracy, the system currently requires considerable manual fine tuning of the feature weights and the threshold during the training phase. Since the web is dynamically changing, new features will emerge and need to be identified. When new features are added into the system, in addition to determining the weights of the new features, the existing feature weights and the threshold should also be modified. To keep up the changes, future research should seek to develop more automated training and fine tuning process.

To achieve better accuracy for classifying web page containing advertisements, future research should address the problem of processing the contents of pictures incorporating pattern recognition, since numerous advertisements are contained within pictures.

REFERENCES

[1] B. Choi and Q. Guo, "Applying Semantic Links for Classifying Web Pages," *Developments in Applied Artificial Intelligence, IEA/AIE 2003, Lecture Notes in Artificial Intelligence*, vol. 2718, pp. 148-153, 2003.

[2] B. Choi and X. Peng, "Dynamic and Hierarchical Classification of Web Pages," *Online Information Review*, vol. 28, no. 2, pp. 139-147, 2004.

[3] B. Choi and Z. Yao, "Web Mining by Automatically Organizing Web Pages into Categories," *Distributed Artificial Intelligence, Agent Technology, and Collaborative Applications*, Idea Group Inc, ch. XII, pp. 214-231, 2008.

[4] B. Choi and Z. Yao, "Web page classification," *Foundations and Advances in Data Mining*, Springer-Verag, 2005.

[5] X. Peng and B. Choi, "Automatic Web Page Classification in a Dynamic and Hierarchical Way," in *Proc. IEEE International Conference on Data Mining*, pp. 386-393, 2002.

[6] A. Asirvatham and K. Ravi, "Web page classification based on document structure," *IEEE National Convention*, Dec. 2001.

[7] M. Dimitrova, A. Finn, N. Kushmerick, and B. Smyth, "Web genre visualization," *Smart Media Institute*, in *Proc. conference on Human Factors in Computing Systems*, Minneapolis, 2002.

[8] B. Kwanik and K. Crowston, "Genres of digital documents: Introduction to the Special Issue," *Information, Technology & People*, vol. 18, no. 2, pp. 76-88, 2005.

[9] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic text categorization in terms of genre and author," *Computational Linguistics*, vol. 26, Issue 4, pp. 471-495, Dec. 2000.

[10] T. Yoshioka and G. Herman, "Genre taxonomy: a knowledge repository of communicative actions," *ACM Transactions on Information System*, vol. 19, no. 4, pp. 431-456, 2001.

[11] G. Chen and B. Choi, "Web Page Genre Classification," *The 23rd annual ACM Symposium on Applied Computing*, pp. 2353-2357, March 2008.

[12] K. Matsuda, Toshikazu, and Fukushima "Task-oriented world wide web retrieval by document type classification," in *Proc. the 8th International Conference on Information and Knowledge Management*, pp. 109-113, 1999.

[13] S. Haas and E. Grams, "Page and link classifications: connecting diverse resources," *ACM DL*, pp. 99-107, 1998.

[14] B. Choi, "Making Sense of Search Results by Automatic Web-page Classifications," *WebNet 2001 -- World Conference on the WWW and Internet*, pp. 184-186, 2001.

KankanaShukla is a Masters student in Computer Science and Biomedical Engineering at Louisiana Tech University. She did her Bachelors in Electronics and Instrumentation. Her research interest includes Data Mining, Web Mining, Big Data Analysis, Machine Learning, Bioinformatics, Biostatistics, Robotics and Artificial Intelligence. Her future work includes pursuing a Ph.D. degree in Data Mining and Biostatistics.

Ben Choi has a Ph.D. degree in Electrical and Computer Engineering and also has a Pilot certificate for flying airplanes and helicopters. He is an associate professor in Computer Science at Louisiana Tech University. He received his Ph.D., M.S., and B.S. degrees from The Ohio State University, studied Computer Science, Computer Engineering, and Electrical Engineering. His areas of research include Humanoid Robots, Artificial Intelligence, Machine Learning, Intelligent Agents, Semantic Web, Data Mining, Fuzzy Systems, and Parallel Computing. His future research includes developing advanced software and hardware methods for building intelligent machines and theorizing the Universe as a Computer.