A WEBPAGE CLASSIFICATION ALGORITHM CONCERNING WEBPAGE DESIGN CHARACTERISTICS

Shih-Ting Yang*
Department of Information Management
Nanhua University
Chia-Yi (622), Taiwan

ABSTRACT

Owing to the booming growth of Internet technology, the number of web documents has significantly increased over the Internet. If the webpage can be effectively managed, the knowledge demanders (i.e., Internet users) can efficiently absorb and use the knowledge documents; it has become the core topic in this information explosion era. Webpage classification technology with high accuracy can improve the efficiency for Internet users to search required knowledge and to save lots of knowledge-searching time. Differing from previous researches, this paper explores webpage design characteristics for webpage classification. That is, concerning complexity of webpage structure, this paper analyzes the webpage design characteristics including tag attributes and tag-region layout to develop an algorithm for webpage classification. Therefore, based on webpage design characteristic analysis, the text contained in specific tag-regions can be identified. Also, the keywords extracted from each tag-region are weighted according tag attributes and tag-region locations; then, the categories of the target webpage can be determined. Furthermore, based on the hyperlink tag, the similar webpage with higher correlations can be collected to re-determine target webpage categories. In addition to the webpage classification algorithm, a web-based webpage classification system is developed to demonstrate feasibility of the proposed model. The attempt of this research is to analyze and use the characteristics of webpage design for webpage classification technology to improve the effectiveness of classification.

Keywords: Tag-region, Webpage Classification, Webpage Design, Keyword Extraction, Knowledge Management

1. INTRODUCTION

With the advancement of Internet technologies, the number of Internet users is increasing and the amount of information online has growth explosively. As browsing information or files on the Internet has become one of important channels for knowledge acquisition, how to effectively manage Internet information/files to assist the users in efficiently absorbing and utilizing required information has become an important issue. That is, if the webpage that contains these Internet information/files can be effectively classified, it would enhance user convenience and increase webpage browsing rate to derive their required information.

On the basis of this issue, many technologies for webpage classification have been developed. Since webpage contents contain texts, pictures or films, most researches analyze and classify these kinds of data for categorization. Also, some researches maintain domain keywords in database as a basis for determination of webpage categories. Furthermore, as the tags appearing in pairs (i.e., tag-region) contain words of certain segmentation (e.g., <title></title > and <h1></h1>) in webpage, some researches apply the standardized programming pattern of UML and HTML used by webpage creators/designers to analyze webpage tags for webpage classification. In addition, in case of insufficient data for webpage classification, other researches employ hyperlinks contained in webpage for such purpose. The AS-IS model of webpage classification is as shown in Figure 1.

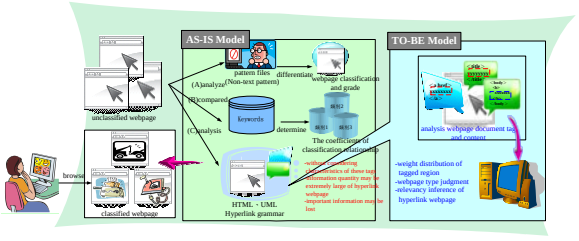

Figure 1: The As-Is and To-Be models of webpage classification

* **Corresponding author: stingyang@mail.nhu.edu.tw**

As shown in Figure 1, although HTML, UML or webpage tags may be used for webpage classification, most of these technologies remove webpage tags to capture texts for webpage classification (i.e., webpage text captures are considered to be equally important). In such case, the critical information may be ignored. Taking tag-region of subject words (i.e., webpage subject terms contained in tag-region <title> and </title>) as an example, most subjects are expressed in concise words, so few or no keywords can be extracted from tag-region of subject terms and leading to unideal classification result. Moreover, if hyperlink webpage are used without screening as classification information of the target webpage, webpage classification time and incorrect information are both increased. As for HTML or tag used for webpage classification currently, this paper summarizes the following main issues.

- The texts contained in different tag-regions are regarded as equal important without considering tag attributes in most current researches.
- Hyperlink webpage without screening mechanism used as a basis for webpage classification lead to reduce efficiency of classification.

Different from previous webpage classification methodologies, this paper concentrates on webpage design characteristics (including tag attributes and tag-region layout, etc.) for webpage classification. That is, this paper analyzes tag attributes (e.g., head tag <title> always contains more representative words with respect to the target webpage) and also considers tag-region layout (i.e., tags of the same type located in different tag-region layout contains words of different importance) to assign the corresponding weights for all considered tags. After that, this paper employs keyword extraction technology to extract all the keywords contained in different tag-regions and given the corresponding weights. Finally, to avoid problems of insufficient or excessive analysis information, this paper establishes a hyperlink webpage screening mechanism to collect the hyperlink webpage with higher correlations to slightly modify or adjust the categories of the target webpage. The TO-BE model of this paper is shown in Figure 1.

2. LITERATURE REVIEW

Concerning the webpage classification issue, most researches focus on analyzing the texts, tags, hyperlinks and images contained in webpage.

2.1 Webpage text information analysis

Previous researches extract keywords for webpage classification based on webpage text information [4,10,22]. The automatic webpage classification of Wolverhampton Web Library uses DDC (Dewey Decimal Classification) and manual definition of keywords to increase the accuracy of webpage classification [13]. Besides the above extraction of keywords, the semantic in webpage text is also analyzed [20]. The SRG (Semantic Relationship Graph) is constructed according to the spatial scale that could be searched under the guidance of combinative table and association list, and then the Naive Bayesian classifier is used to develop a semantic relationship graph based multi-relationship Naive Bayesian classifier [4]. Such classifier removes unnecessary characteristics and relationships according to the analytical results of semantic relationship graph to avoid generating uncorrelated associations [8,23]. In addition, the Syntactic Similarity of files is proposed to analyze webpage text for identification of the similarity among various webpage; then, each two highly related webpage are classified into the same category [3].

2.2 Webpage tag information analysis

Lim et al. [15] proposes UML and HTML grammars or tag characteristic in webpage documents as the analytic data of webpage classification. The extracted data are taken as analysis characteristics and data of webpage classification for further studies [2]. In addition, based on the DOM (Document Object Model) tag-tree structure, a webpage can be segmented into small tag-regions. Each tag-region can be displayed in the browser by visualized types corresponding to a specific nested combination of tag-pairs. The profitability of tag-regions for webpage classification is varying among visual types caused by the web authoring convention [19]. Furthermore, webpage design and advanced digest algorithms in LookSmart webpage directory are both used to improve the efficiency of webpage classification and reduce extra webpage messages [17].

2.3 Webpage hyperlink information analysis

Furnkranz [9] classifies webpage documents considering the hyperlink ensembles in webpage. The classification data can be obtained from the text of the target webpage and the hyperlink webpage and used to classify the target webpage effectively. The OEM (Object Exchange Model) is employed to identify webpage categories [14]. In this methodology, the number of hyperlink of the target webpage is calculated and the contents of hyperlink webpage are converted into Node Similarity, Edge Similarity and Structural Similarity to obtain the similarity degree to classify the similar webpage into the same category. In addition, Davison [6] uses Web Crawlers is used to extract webpage link nodes to construct the website tree structure; then, the extracted tag information is applied to calculate the similarity degree as a basis for category identification of the target webpage.

2.4 Webpage image information analysis

The webpage contents contain not only texts, but also pictures and films. The image-block analysis technology is proposed for webpage classification [7]. This technology uses the recognition degree in image-block and data compact block to identify the importance of the image in the webpage to increase the webpage classification accuracy. Based on graphic structure, the webpage comparison technology is developed [1]. The graphic elements in the webpage can be obtained by analyzing the webpage structure, and the relevant webpage can be searched within these range conditions. These relevant webpage are presented in tree structure, and then the edition distance between subtrees is used to measure the similarity degree among webpage. Therefore, the search area can be enlarged and the classification effectiveness can be improved. In order to distinguish most analysis text information, the content type of each separated area of the image file can be determined based on the area feature vector in the unit of 25 dimensions [21]. Furthermore, the CART (Classification and Regression Tree) classifier is constructed for out digital document classification by low-level perceptual features in images [16].

3. WEBPAGE CLASSIFICATION MODEL

The webpage classification model proposed in this paper is based on analysis of tag attributes and tag-regions to search for text contained in tag-regions and extract the corresponding keywords. Based on tag attributes and specific tag-region layout of webpage, the corresponding weight values are assigned to various tag-regions. Therefore, according to keywords extracted from various tag-regions and weights assigned for tag-regions, the categories of target webpage can be determined. Finally, hyperlink tag (i.e., <a href>) is used to search for hyperlink webpage with higher correlation to modify the categories of the target webpage. Therefore, this model can be divided into three kernel modules including "tag-region weight assignment (TWA) module" (as shown in Part 1 of Figure 2), "webpage category determination (WCD) module" (as shown in Part 2 of Figure 2) and "hyperlink webpage determination (HWD) module" (as shown in Part 3 of Figure 2).

3.1 Tag-region weight assignment (TWA) module

As designing the webpage, most designers use HTML to program and arrange article contents in webpage. Since HTML mainly employs tags in pairs to specify exhibition mode of webpage texts, webpage designers only need to utilize suitable webpage tags to produce the same reading mode as traditional articles. Just like writing mode of common articles, webpage designers may apply webpage tags to define title, abstract, keywords, chapter titles and other contents of webpage. The webpage designers may also emphasize uniqueness and importance of words by bold, italic or underline, etc. Besides these typical tag attributes, tags of the same type but located in different tag-region layout (i.e., external tag-regions may also contain internal tag-regions) may contain contents of different significance. To differentiate these tags, this module analyzes spatial planning of webpage (i.e., locations of tag-regions in webpage) to assign different weights to the corresponding tag-regions in different tag-region spatial layout.

In TWA module, this paper acquires the tags (i.e., tag extraction mechanism) correlated with webpage classification (i.e., which contain text data); then, the weight assignment of tag-region in different tag-region spatial layout (i.e., tag-region location analysis mechanism) are discussed.

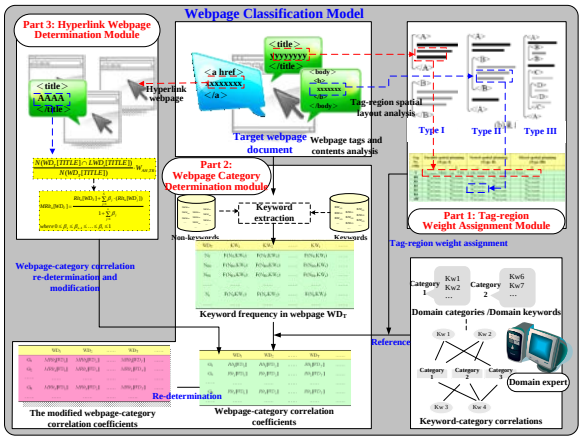

Figure 2: Architecture of webpage classification model

(A) Tag extraction mechanism

In HTML, all tags have their respective purposes, and contents contained in tag-regions often reflect the attributes of those tags. For example, head tags (including <head>, <title>, <bgsound>, <meta>, <style>, <script>, etc.) are mainly used to define segments of such setting values as format, form, name, Script language and pattern list of the target webpage. Among which, head tag <title> is containing the subject of the webpage, and designers often employ tags <h1> to <h6> to display and highlight subject of different sizes. Webpage body tag (i.e., <body>) contains all texts, pictures and other multi-media files to be displayed. Concerning display of webpage texts, webpage designers often employ bold/italic tags (including <b>, , <cite>, and etc.), tabulating tags (including <ol>, <li>, <ul>, <dl>, <dt>, <dd>, etc.) and quoted text tag (i.e., <blockquote>) to emphasize importance of terms displayed in webpage.

In a webpage, texts contained in different

tag-regions denote different significance and importance. Therefore, webpage classification model in this paper takes text data in webpage as analysis basis. The webpage head tag and webpage body tag are mainly utilized, and text tone strengthening tags are also considered to serve as the basis for webpage classification. Webpage tags to be employed are summarized in Table 1.

Table 1: List of tags as the classification basis

Tag types	Tag Names	Text highlight tags	
Head	<title>	Subject tags (T)	<h1>, <h2>, <h3>, <h4>, <h5>, <h6>…
Body	<body>	Bold tags (B_1)	<b>, …
		Italic tags (B_2)	<address>, <cite>, <dfn>, , <i>….
		Tabular tags (B_3)	, , , <dl>, <dt>, <dd>…
		Quotation tags (B_4)	<blockquote>…
Hyperlink	<a href>	Hyperlink tag (AH)	

(B) Tag-region location analysis mechanism

Following tag extraction as classification basis, the extracted tags of the same type but located in different positions may contain texts of different significance. In order to differentiate these tags, this module analyzes the spatial layout of tag-region (Hsu, 2000), and assigns weights to tag-regions located in different spatial layout.

As HTML has the function of spatial planning of webpage, this module further discusses the relationship between tag-regions and webpage spaces. That spatial planning of tag-regions can be divided into three types including (1) Parallel spatial planning, (2) Nested spatial planning and (3) Mixed spatial planning (as shown in Figure 3 to Figure 6) and the principle of tag-region weight assignment for spatial planning are described as follows.

(1) Parallel spatial planning: As tag-regions are all independent from each other, weight assignment of tag-regions is only referred to tag attributes.

(2) Nested spatial planning: As one tag-region (i.e., external tag-region) contains not only contents but also other tag-regions (i.e., internal tag-region), some contents of the webpage may be contained in one external tag-region and several internal tag-regions simultaneously. This module argues that webpage contents in these overlapping regions are related with tag attributes of innermost tag-region. However, as the innermost tag-region is contained in several tag-regions, its significance should be properly strengthened. As a result, the innermost tag-region should also be provided with weight values of external tag-regions to enhance text significance therein. Weight calculation of the inner tag-region is shown in Equation (1) and symbols used in this mechanism are defined as follows.

$W_{j,TR}$ The weight value of the j'th tag located in TR, $j \in \{T, B_1, B_2, ...\}$ (see Table 1) and $TR \in \{TypeIA, TypeIB, TypeIIA, TypeIIB, TypeIIIA, ...\}$ (see spatial planning in Figure 3 to Figure 6).

α_i The added weight value of the i'th external tag contained in the j'th tag (located in TR).

$$W^*_{j,TR} = \left[1 + \sum_{all\ i}(\alpha_i)\right] \cdot W_{j,TR} \quad \text{where} \quad 0 \le \alpha_i \le 1 \qquad (1)$$

(3) Mixed spatial planning: This spatial planning applies the principles of the above two spatial planning; so that, weight assignment is the same as that of nested spatial planning.

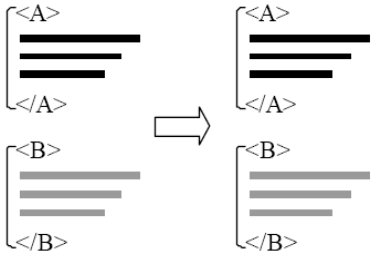

Figure 3: Parallel spatial planning (Type I)

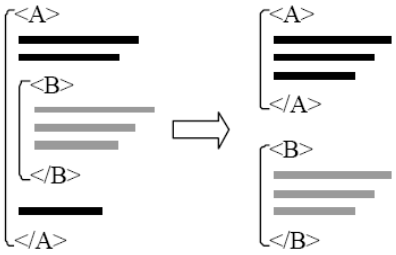

Figure 4: Nested spatial planning (Type II)

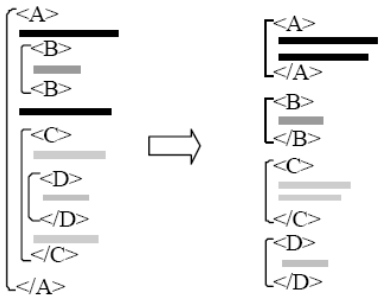

Figure 5: Mixed spatial planning (Type III)

To avoid tag-regions with the same form and name but different significance as the same ones (in such case, critical information for webpage classification may be lost), this module extracts tags first, along with spatial planning of tag-regions to differentiate those tag-regions which may have different importance, and then assign the corresponding weight values. The weight values of tag-regions can be summarized in Table 2.

Table 2: List of weight values of tag-regions

Tag No. (TR)	Parallel spatial planning (Type I)		Nested spatial planning (Type II)		Mixed spatial planning (Type III)			
	A	B	A	B	A	B	C	D
T	$W_{T,TypeIA}$, where head tag <title> is only existed in the A location of Type I							
B0	$W_{B0,TypeIA}$	$W_{B0,TypeIB}$	$W_{B0,TypeIIA}$	$W_{B0,TypeIIB}$	$W_{B0,TypeIIIA}$	$W_{B0,TypeIIIB}$	$W_{B0,TypeIIIC}$	$W_{B0,TypeIIID}$
B1	$W_{B1,TypeIA}$	$W_{B1,TypeIB}$	$W_{B1,TypeIIA}$	$W_{B1,TypeIIB}$	$W_{B1,TypeIIIA}$	$W_{B1,TypeIIIB}$	$W_{B1,TypeIIIC}$	$W_{B1,TypeIIID}$
B2	$W_{B2,TypeIA}$	$W_{B2,TypeIB}$	$W_{B2,TypeIIA}$	$W_{B2,TypeIIB}$	$W_{B2,TypeIIIA}$	$W_{B2,TypeIIIB}$	$W_{B2,TypeIIIC}$	$W_{B2,TypeIIID}$
B3	$W_{B3,TypeIA}$	$W_{B3,TypeIB}$	$W_{B3,TypeIIA}$	$W_{B3,TypeIIB}$	$W_{B3,TypeIIIA}$	$W_{B3,TypeIIIB}$	$W_{B3,TypeIIIC}$	$W_{B3,TypeIIID}$
B4	$W_{B4,TypeIA}$	$W_{B4,TypeIB}$	$W_{B4,TypeIIA}$	$W_{B4,TypeIIB}$	$W_{B4,TypeIIIA}$	$W_{B4,TypeIIIB}$	$W_{B4,TypeIIIC}$	$W_{B4,TypeIIID}$
AH	$W_{AH,TypeIA}$	$W_{AH,TypeIB}$	$W_{AH,TypeIIA}$	$W_{AH,TypeIIB}$	$W_{AH,TypeIIIA}$	$W_{AH,TypeIIIB}$	$W_{AH,TypeIIIC}$	$W_{AH,TypeIIID}$

3.2 Webpage category determination (WCD) module

Based on weight values of tag-regions in different spatial planning obtained TWA module, as well as the correlations between keywords and categories established by domain expert in advance, the categories of the target webpage can be determined in WCD module. The symbols used in this module are defined as follows.

D_i	The i'th training webpage
$F(N_j,KW_i)$	The frequency of the i'th keyword exist in the j'th tag-region
G_k	The k'th category
KW_i	The i'th keyword in keyword set after keyword mergence
N_j	The tag-region contained in the j'th tag, where $j \in \{T, B_0, B_1, B_2, ...\}$
$R(G_k,KW_i)$	The correlation coefficient between KW_i and G_k
$Rlt_k^{'}[WD_T]$	The relationship coefficient between WD_T and G_k
$Rlt_k[WD_T]$	The correlation coefficient between WD_T and G_k
WD_i	The i'th webpage
WD_T	The target webpage

Before classifying the target webpage, the keywords should be established by domain experts in advance. Then, the training webpage D_i (i.e., the webpage of known contents and categories) are used to calculate the frequencies $N(D_i,KW_i)$ of keywords existing in each training webpage (as shown in Table 3). After that, by using correlations between training webpage (such as news webpage) and domain categories, the correlation coefficient $R(G_i,KW_i)$ between keyword KW_i and category G_i can be established in Table 4 [11].

The WCD module in this paper is also based on these two established Tables. The details for webpage classification are introduced as follows.

Step (C1): Definition of webpage tag-regions

As discussed in TWA module, all the webpage considered in this paper can be segmented through tags. So that, the each webpage WD_j can be divided into several tag-regions (N_T,N_{B0},N_{B1},N_{B2},N_{B3},N_{B4}) (see Equation (2)).

$$WD_j = \{N_T, N_{B_0}, N_{B_1}, ..., N_{B_4}\} \quad (2)$$

Where N_T~N_{B4} represent webpage texts contained in webpage head tag <title>, as well as webpage texts contained in text tag B_0 or text highlight tags B_1 to B_4 in body tag <body>.

Step (C2): Calculation of frequencies of keywords in the target webpage

After webpage tag-regions are defined, keywords extraction technology [18] is adopted to extract keywords contained in tag-regions in the target webpage WD_T (as shown in Table 5).

Table 3: Frequency of keyword in training webpage

	KW_1	KW_2	.....	KW_i	.....
D_1	$N(D_1,KW_1)$	$N(D_1,KW_2)$	.....	$N(D_1,KW_i)$	.....
D_2	$N(D_2,KW_1)$	$N(D_2,KW_2)$	.....	$N(D_k,KW_i)$	.....
.....	.....	.....	.....	.....	.....
D_j	$N(D_j,KW_1)$	$N(D_j,KW_2)$	.....	$N(D_k,KW_i)$	.....
.....	.....	.....	.....	.....	.....

Table 4: The keyword/category correlation coefficients

	KW_1	KW_2	.....	KW_i	.....
G_1	$R(G_1,KW_1)$	$R(G_1,KW_2)$	.....	$R(G_1,KW_i)$	.....
G_2	$R(G_2,KW_1)$	$R(G_2,KW_2)$	.....	$R(G_i,KW_i)$	.....
.....	.....	.....	.....	.....	.....
G_j	$R(G_j,KW_1)$	$R(G_j,KW_2)$	.....	$R(G_i,KW_i)$	.....
.....	.....	.....	.....	.....	.....

Table 5: Frequencies of keywords contained in tag-regions of the target webpage

WD_T	KW_1	KW_2	.....	KW_i	.....
N_T	$F(N_T,KW_1)$	$F(N_T,KW_2)$	.....	$F(N_T,KW_i)$	.....
N_{B0}	$F(N_{B0},KW_1)$	$F(N_{B0},KW_2)$	.....	$F(N_{B0},KW_i)$	.....
N_{B1}	$F(N_{B1},KW_1)$	$F(N_{B1},KW_2)$	.....	$F(N_{B1},KW_i)$	.....
.....	.....	.....	.....	.....	.....
N_j	$F(N_j,KW_1)$	$F(N_j,KW_2)$	.....	$F(N_j,KW_i)$	.....
.....	.....	.....	.....	.....	.....

Step (C3): Calculation of relationship coefficient between target webpage and categories

Based on the derived keyword frequencies in the target webpage (see Table 5), keyword-category correlation coefficients (see Table 4) and tag-region weight assignment (see Table 2), the relationship coefficient $Rlt_k^{'}[WD_T]$ between the target webpage WD_T and category G_k can be obtained via Equation (3) to preliminarily determine the categories of the target webpage. The relationships between each webpage and categories are summarized in Table 6.

$$Rlt_k^{'}[WD_T] = \frac{\sum_{all\ i}\sum_{all\ j}\sum_{all\ TR} R(G_k, KW_i)\cdot F(D_j, KW_i)\cdot W_{j,TR}}{\sum_{all\ i}\sum_{all\ j}\sum_{all\ TR} F(D_j, KW_i)\cdot W_{j,TR}} \quad (3)$$

where $j \in \{T, B_0, B_1, B_2, ...\}$

and $TR \in \{TypeIA, TypeIB, TypeIIA, TypeIIB, TypeIIIA, ...\}$

Table 6: Relationship coefficients between webpage and categories

	WD_1	WD_2	.....	WD_T	.....
G_1	$Rlt_1^{'}[WD_1]$	$Rlt_1^{'}[WD_2]$	.....	$Rlt_1^{'}[WD_T]$	.....
G_2	$Rlt_2^{'}[WD_1]$	$Rlt_2^{'}[WD_2]$	.....	$Rlt_2^{'}[WD_T]$	.....
.....	.....	.....	.....	.....	.....
G_k	$Rlt_k^{'}[WD_1]$	$Rlt_k^{'}[WD_2]$	.....	$Rlt_k^{'}[WD_T]$	.....
.....	.....	.....	.....	.....	.....

Step (C4): Calculation of correlation coefficients between target webpage and categories

As the sum of relationship coefficients of target webpage is not equal to 1 (i.e. $\sum_{all\ k} Rlt_k^{'}[WD_T] \neq 1$). In Step (C4), the relationship coefficient $Rlt_k^{'}[WD_T]$ between target webpage WD_T and category G_k should be standardized (as shown in Equation (4)) to obtain the webpage-category correlation coefficient $Rlt_k[WD_T]$ (as shown in Table 7). If the webpage-category coefficient is greater, the target webpage approaches the corresponding category. On the other hand, if the value is equal to zero, the target webpage WD_T have no relation to category G_k.

$$Rlt_k[WD_T] = \frac{Rlt_k^{'}[WD_T]}{\sum_{all\ k} Rlt_k^{'}[WD_T]} \quad (4)$$

Table 7: The correlation coefficients of webpage and categories

	WD_1	WD_2	.....	WD_T	.....
G_1	$Rlt_1[WD_1]$	$Rlt_1[WD_2]$	.....	$Rlt_1[WD_T]$	.....
G_2	$Rlt_2[WD_1]$	$Rlt_2[WD_2]$	.....	$Rlt_2[WD_T]$	.....
.....	.....	.....	.....	.....	.....
G_k	$Rlt_k[WD_1]$	$Rlt_k[WD_2]$	.....	$Rlt_k[WD_T]$	.....
.....	.....	.....	.....	.....	.....

3.3 Hyperlink webpage determination (HWD) module

As single webpage cannot cover all knowledge to be described, webpage designers often use hyperlink tags (i.e., <a href>) in target webpage to build webpage hyperlink. Based on the hyperlink, Internet users can link from the target webpage to another webpage for acquisition of more relevant knowledge. Therefore, the relationship existed between target webpage and linked webpage should be discussed.

This module uses the hyperlink tags in target webpage to derive the hyperlink webpage with higher correlation for redetermination and modification of the categories of the target webpage. Firstly, referring to the weight assignment of tag-regions (see Table 2), the weight values of all hyperlink tags and subject words of the corresponding hyperlink webpage can be acquired. Secondly, the relationship between each hyperlink webpage and the target webpage can be calculated and ranked to select the hyperlink webpage within predefined selection degree. After that, the categories of target webpage can be re-determined accordingly. The symbols used in this module are defined and the details are introduced as follows.

β_j	The modification weight value of the j'th ranked hyperlink webpage
LWD_t	The t'th hyperlink webpage of target webpage
$M[LWD_t]$	The correlation value between WD_T and LWD_t
$MRlt_k[WD_T]$	The modified correlation coefficient between the WD_T and G_k
$N(WD_T[TITLE])$	Number of subject words contained in head tag-region <title> of the target webpage WD_T
$Rlt_k[WD_j^{'}]$	The classification coefficient between the j'th ranked hyperlink webpage $WD_j^{'}$ and category G_k
$WD_j^{'}$	The j'th ranked hyperlink webpage according their correlation values $M[LWD_t]$ (in descent order)

Step (D1): Calculation of correlation value of hyperlink webpage with respect to the target webpage

Firstly, the weight values of all hyperlink tags in the target webpage WD_T can be obtained from TWA module. Secondly, the repetition proportion of subject words between each hyperlink webpage LWD_t and WD_T can also be obtained. After that, the correlation value of LWD_t with respect to WD_T can be derived via Equation (5).

$$M[LWD_t] = \frac{N(WD_T[TITLE] \cap LWD_t[TITLE])}{N(WD_T[TITLE])} \cdot W_{AH,TR} \quad (5)$$

Where $N(WD_T[TITLE] \cap LWD_t[TITLE])$ denotes repetition number of subject words in target webpage and the t'th hyperlink webpage.

Step (D2): Setting of selection degree of similar hyperlink webpage

All the hyperlink webpage in WD_T are ranked according to their correlation values $M[LWD_t]$ (in descent order). Also, the selection degree S should be defined in advance. The top S ranking hyperlink webpage ($WD_j^{'}$, where $j \leq S$) are selected and the modification weight values β_j (j=1,…,S) of these selected webpage are assigned to re-determine the categories of the target webpage.

Step (D3): Re-determination of correlation coefficients of target webpage and categories

After top S ranking hyperlink webpage are selected, the hyperlink webpage with higher correlation values are given corresponding weight values (i.e., β_1, β_2,…, β_s, the weight value of the preceding one is greater than or equal to that of the following one). Based on these modification weight values, this module re-determines the correlation coefficients of target webpage and categories to

obtain the modified correlation coefficient $MRlt_k[WD_T]$ (as shown in Equation (6)). The correlations between each webpage and categories revised from Table 7 are summarized in Table 8.

$$MRlt_k[WD_T] = \frac{Rlt_k[WD_T] + \sum_{j=1}^{s} \beta_j \cdot (Rlt_k[WD_j'])}{1 + \sum_{j=1}^{s} \beta_j} \quad (6)$$

$$where\ 0 \le \beta_s \le \beta_{s-1} \le \ldots \le \beta_1 \le 1$$

Where $Rlt_k[WD_j']$ is the modified correlation coefficient between the ranked j'th hyperlink webpage WD_j' and the k'th category.

Table 8: The modified correlation coefficients of webpage and categories

	WD_1	WD_2	.....	WD_T	.....
G_1	$MRlt_1[WD_1]$	$MRlt_1[WD_2]$	.....	$MRlt_1[WD_T]$	.....
.....	.....	.....	.....	.....	.....
G_k	$MRlt_k[WD_1]$	$MRlt_k[WD_2]$	.....	$MRlt_k[WD_T]$	.....
.....	.....	.....	.....	.....	.....

3.4 Summary

As a whole, based on the TWA, WCD and HWD modules, the categories of unclassified webpage can be determined. As existing classification methods for webpage classification fail to consider two problems: (1) Texts contained in tag-regions may have different importance; (2) Tags of the same type but located in different spatial layout may contain texts of different significance. Therefore, this paper conceders tag attributes, analyzes tag-region layout, employs keywords extraction technology and utilizes hyperlink webpage to establish the model for webpage classification.

4. WEBPAGE CLASSIFICATION SYSTEM

In order to demonstrate feasibility of the proposed algorithm for tag-region layout analysis and webpage classification, a web-based portal, namely webpage classification system, is developed for webpage classification over Internet. Under the webpage classification system, the webpage documents could be maintained and the user authorities are properly managed so that the webpage classification results can be accurately provided to this developed staff.

Based on the user login information, the webpage classification system recognizes the user category (e.g., system administrator and common user) and provides the corresponding functions to the user. Under the system, the system administrator establishes the domain keywords with respect to the specified categories to database via keyword maintenance module as the foundation of system training (Figure 4). Also, the system administrator can upload the webpage documents with given categories to the database via training webpage document upload function (Figure 5). After that, the keyword-category correlations and webpage-category correlations can be established in system database. After uploading these training webpage documents, the webpage-category correlation coefficients of unclassified webpage documents uploaded by common users can be determined by system administrator through webpage classification function (Figure 6 and Figure 7). Furthermore, if all the correlations of the target webpage are not greater than predefined threshold, the system automatically recommends administrator to re-determine the webpage-category correlations via hyperlink webpage analysis function (Figure 8). In addition, the system administrator can set weight values of tags according tag attributes, category determination threshold and hyperlink webpage selection threshold, etc. through the system parameter maintenance module (Figure 9). Finally, the system administrator can maintain users' profiles and control users' authorities via user profile maintenance module.

Figure 4: Maintain keywords and keyword-category correlations

Figure 5: Upload training webpage documents to the system

Figure 6: Select webpage documents for classification

Figure 7: Results of webpage classification (correlation coefficients <predefined threshold)

Figure 8: Re-determine webpage-category correlations by hyperlink analysis function

Figure 9: Set hyperlink webpage selection threshold

Under the platform, common user also can upload the webpage documents via webpage upload function, so that the webpage documents can be efficiently managed and shared. After uploading webpage documents, system automatically extracts keywords and analyzes webpage tag-region layout; then, these extracted and analyzed information are maintained in system database. Also, common user can review or download all kinds of webpage documents provided by system administrator or other common users in the database via webpage search function (Figure 10). After the webpage categories are determined, common users also can inquire webpage-category correlations of target webpage documents via webpage search function (Figure 11).

Figure 10: Results of webpage inquiry (1)

Figure 11: Results of webpage inquiry (2)

In addition to demonstrating feasibility of the proposed webpage classification algorithm and system, the theoretical contributions and practical applications of this proposed algorithm and system are summarized as follows.

- ✓ **Theoretical contributions:** Since the traditional webpage classification technologies based on whole content-based webpage make it difficult to considerate of webpage layout designed from webpage designer for determination of webpage categories, this paper analyzes the webpage design characteristics including tag attributes and tag-region layout designed in webpage to improve the effectiveness of webpage classification.
- ✓ **Practical applications**: The proposed algorithm and developed modules can be integrated into the CRM and KM system for enterprise knowledge classification and knowledge sharing (e.g., the webpage training materials). So that, the classified knowledge documents (e.g., the classified training webpage documents) can be provided and enable the knowledge receivers to efficiently derive the original or critical expressions, concepts and information of domain experts or trainers.

As a result, this algorithm not only proposes an additional concern and feature (i.e., webpage design characteristics) for webpage classification technologies but can also be applied in real-world systems of enterprises.

5. CONCLUSION

Different form technologies for webpage classification, this paper analyzes tag attributes and tag-region layout in webpage to develop an algorithm for webpage classification including tag-region weight assignment (TWA) module, webpage category determination (WCD) module and hyperlink webpage determination (HWD) module. In TWA module, tag attributes and tag-region layout designed in webpage are analyzed to assign weight values to the corresponding tag-regions. In WCD module, the keyword extraction technology is employed to extract keyword contained in each tag-region and the corresponding weights are given to determine the webpage-category correlations. In HWD module, the hyperlink webpage with higher correlations with

respect to the target webpage are used to re-determine and modify the webpage categories. The attempt of this research is to improve the accuracy and efficiency of webpage classification by concerning the characteristics of webpage design. Also, the proposed webpage classification algorithm can assist the information demanders to efficiently and effectively search the required information over the Internet; so that, lots of researching energy and time can be reduced.

REFERENCES

1. Alpuente, M. and Romero, D., 2009, "A visual technique for web pages comparison," *Electronic Notes in Theoretical Computer Science*, Vol. 235, No. 1, pp. 3-18.
2. Artail, H. and Kassem, F., 2008, "A fast HTML web page change detection approach based on hashing and reducing the number of similarity computations," *Data & Knowledge Engineering*, Vol. 66, No. 2, pp. 326-337.
3. Broder, A. Z., Glassman, S. C., Manasse, M. S. and Zweig, G., 1997 "Syntactic clustering of the Web," *In Proceedings of the Sixth International World Wide Web Conference*, pp. 391-404.
4. Chen, C. M., Lee, H. M. and Chang, Y. J., 2009, "Two novel feature selection approaches for web page classification," *Expert Systems with Applications*, Vol. 36, No. 1, pp. 206-272.
5. Chen, H., Liu, H., Han, J., Yin, X. and He, J., 2009, "Exploring optimization of semantic relationship graph for multi-relational Bayesian classification," *Decision Support Systems*, Vol. 48, No. 1, pp. 112-121.
6. Davison, B. D., 2000, "Topical Locality in the Web," *In the Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pp. 272-279.
7. Fersini, E., Messina, E. and Archetti, F., 2008, "Enhancing web page classification through image-block importance analysis," *Information Processing & Management*, Vol. 44, No. 4, pp. 1431-1447.
8. Fujino, A., Ueda, N. and Saito K., 2007, "A hybrid generative/discriminative approach to text classification with additional information," *Information Processing and Management*, Vol. 43, pp. 379-392.
9. Furnkranz, J., 2002, "Hyperlink ensembles: A case study in hypertext classification," *Information Fusion*, Vol. 3, No. 4, pp. 299-312.
10. Horng, J. T. and Yeh, C. C., 2000, "Applying genetic algorithms to query optimization in document retrieval," *Information Processing and Management*, Vol. 36, pp. 737-759.
11. Hou, J. L. and Lin, F. H., 2004, "A document and user matching model via document keyword analysis," *Journal of Computer Information Systems*, Vol. 44, No. 4, pp. 1-15.
12. Hsu, H. C., 2000, "The web page classifier based on progressive tagged-region analysis," *Department of Information Engineering, Tamkang University, Master Thesis.*
13. Jenkins, C., Jackson, M., Burden, P. and Wallis, J., 1998, "Automatic classification of Web resources using Java and Dewey Decimal Classification," *Computer Networks and ISDN Systems*, Vol. 30, No. 1-7, pp. 646-648.
14. Kuo, Y. H. and Wong, M. H., 2000, "Web document classification based on hyperlinks and document semantics," *PRICAI 2000 Workshop on Text and Web Mining*, pp. 44-51.
15. Lim, C. S., Lee, K. J. and Kim, G. C., 2005, "Multiple sets of features for automatic genre classification of web documents," *Information Processing & Management*, Vol. 41, No. 5, pp. 1263-1276.
16. Schettini, R., Brambilla, C., Ciocca, G., Valsasna, A. and Ponti, M. D., 2002, "A hierarchical classification strategy for digital documents," *Pattern Recognition*, Vol. 35, No. 8, pp. 1759-1769.
17. Shen, D., Yang, Q. and Chen, Z., 2007, "Noise reduction through summarization for web-page classification," *Information Processing & Management*, Vol. 43, No. 6, pp. 1735-1747.
18. Sun, M. T. and Hou, J. L., 2003, "The architecture and models for security reasoning in an EDMS," *Journal of the Chinese Society of Industrial Engineers*, Vol. 20, No. 4, pp. 305-316.
19. Sung, L. C., 2006, "Progressive analysis scheme for web document classification," *Department of Information Engineering, Tamkang University, PhD Dissertation.*
20. Tan, S. and Zhang, J. 2008, "An empirical study of sentiment analysis for Chinese documents," *Expert Systems with Applications*, Vol. 34, pp. 2622–2629.
21. Wang, Y., Phillips, I. T., and Haralick, R. M., 2006, "Document zone content classification and its performance evaluation," *Pattern Recognition*, Vol. 39, No. 1, pp. 57-73.
22. Yang, C. C., Yen, J. and Chen, H., 2000, "Intelligent internet searching agent based on hybrid simulated annealing," *ELSEVIER Journal on Decision Support System*, pp. 269-277.
23. Youn, E. and Jeong, M. K., 2009, "Class dependent feature scaling method using naive Bayes classifier for text datamining," *Pattern Recognition Letters*, Vol. 30, No. 5, pp. 477-485.

ACKNOWLEDGMENT

This research is partially supported by the National Science Council under project No. NSC 99-2221-E-343 -004

ABOUT THE AUTHOR

Shih-Ting Yang is an assistant professor in the Department of Information Management at Nanhua University. Dr. Yang received his Ph.D. in Industrial Engineering and Engineering Management at National Tsing-Hua University and his research interests are knowledge management and mobile commerce.

(Received September 2010, revised October 2010, accepted November 2010)

考量網頁設計特徵之網頁文件分類技術

楊士霆*
南華大學資訊管理學系
嘉義縣大林鎮中坑里南華路一段 55 號

摘要

隨著網際網路相關技術之盛行，網路使用者亦日趨增加，網路環境資訊量已呈爆炸性成長，因此瀏覽網路上文件或資訊已成為現代人吸取知識的重要管道之一。故如何有效管理此些網路文件/資訊，讓使用者得以掌握，以協助使用者快速吸收並運用此些網路資訊，乃成為在現今資訊爆炸時代中之重要課題。目前網頁分類大多以關鍵字擷取或以HTML語法標籤內的文字區域為依據，作為關鍵資訊分析基礎並進行網頁分類。此些分類技術係將網頁標籤去除，以擷取當中文字型態資訊，進行網頁分類（亦即將所擷取之網頁文字視為同等重要性），但此種情況下，可能有多項關鍵資訊被忽略（如可能遺失網頁標題資訊）。有鑑於此，本研究提出一套以網頁標籤區域（Tagged-Region）為基礎之網頁文件分類模式；於模式中，首先本研究乃考量網頁標籤屬性，發展一套「標籤區域權重分配」模組，以尋找影響網頁文件分類之標籤，並解析各網頁標籤於不同網頁空間規劃下之重要性；之後以具分類代表性標籤區域為基礎，擷取當中關鍵字詞，發展一套「網頁文件類別判定」模組，以推論目標網頁文件之隸屬類別；最後再以鏈結網頁為基礎，發展一套「鏈結網頁關聯程度推導」模組，將關鍵性鏈結網頁之隸屬類別，修訂目標網頁文件之隸屬類別，以完成網頁文件之隸屬類別判定任務。本研究最後乃建立一套網頁文件自動分類系統，以呈現此模式與技術之可行性。綜合言之，本研究之目標乃為提昇網頁文件分類技術之正確率與效率性，因此，對於資訊需求者而言，本研究期望能協助資訊需求者於龐大之網路資訊/文件中，迅速且便捷地尋得其所需要之網路文件資料，以節省資訊需求者花費於資訊過濾與篩選之大量時間。

關鍵詞：標籤區域、網頁文件分類、關鍵字擷取、知識管理
（*聯絡人：stingyang@mail.nhu.edu.tw）