Progressive Analysis Scheme for Web Document Classification

Li-Chun Sung, Chin-Hwa Kuo
Department of CS&IE
Tamkang University
g7190030@tkgis.tku.edu.tw

Meng Chang Chen
Institute of Information Science
Academia Sinica
mcc@iis.sinica.edu.tw

Yeali Sun
Department of IM
National Taiwan University
sunny@im.ntu.edu.tw

Abstract

In this paper, a web document classification scheme, Progressive Analysis Scheme (PAS) is proposed to efficiently and effectively classify HTML web documents. When an author writes a web document, HTML tags are used to visually emphasize the texts related to main concepts. The design of PAS is to catch the authoring convention in terms of the contributions of nested HTML tags to document classification. During the learning phase, PAS provides an enhanced tag sequence model to resolve the sample lacking problem in learning the classification contributions of HTML tag sequences. While in classification phase, PAS decomposes a web document into regions based on the DOM tag-tree, and analyzes the regions in the descending order of their classification contributions. PAS also provides a mechanism called emphasis degree adjustment to defer the processing of noisy region during classification. The simulation results shows that PAS has better performance than full-text (e.g. SVM) and sequential classifier.

1. Introduction

Based on the DOM tag-tree structure, the Web document content can be partitioned into a set of small textual regions (called tag-region in this paper) that every tag-region is a text node of a DOM tag-tree and its parent is HTML tag-pair (called bounding tag-pair). Besides the bounding tag-pair, a tag-region is often surrounded by other tag-pairs. For simplification, we represent the surrounding tag-pair sequence (or surrounding type or surrounding for short) as a comma-separated tag-pair list delimited with angle brackets. Let $\Theta$ denote the set of tag-pairs defined in HTML language. $\Theta = \{TP_1, TP_2, \ldots, TP_n\}$, where $TP_i$ represents the $i$-th tag-pair in $\Theta$ and $n = |\Theta|$. A surrounding $S$ is defined as $S = \{TP_n, TP_m, \ldots, TP_1\}$, where $n_i \in [1, |\Theta|]$.

By the HTML definition, each type of tag-pair presents its embedded text in a specific visual style. As a common practice, web document authors deliberately use the presentation characteristics of HTML tag-pairs for their web document designs. Thus, the contributive terms related to the main concept of the document are often surrounded by some intensely emphasizing tag-pairs. Consequently, the contribution degree of a tag-region and the emphasis degrees of tag-pairs in its surrounding are intermingled or even synergetic. In other words, the contribution degree of a tag-region to web document classification can be inferred from its surrounding. Preferably the tag-regions are sorted by their contribution degrees, and the classifier works on the tag-region sequence from high to low contribution degree. In addition, the classifier does not need to process all the tag-regions if there is sufficient support to the classification decision. In this way, not only can classification performance be improved by the analysis deferment of noisy tag-regions, but also computation time reduced.

1.1. Related works

There are proposals for identifying emphasized tag-regions based on their surrounding tag-pairs. For example, the Voting scheme [2] and the ACIRD system [3] consider that contributive terms should appear in the tag-regions bounded by one of a small set of specific tag-pairs. For both methods, the emphasis degrees of tag-pairs are assigned experientially. Furthermore, they only consider the bounding tag-pair. Some works on XML documents, e.g. [4], employ the entire surrounding of tag-region. However, for HTML surroundings, some of them may not exist in the training data set, or their training data is not sufficient for a sound training of emphasis degree.

Some researchers apply the structural information to obtain the contribution of a tag-region of a concept segment from other tag-regions of this segment. The locale of a concept segment can be identified according to raw DOM tag-tree [5,6] or visual layout [7] of the web
document. However, they evaluate roughly the contribution per segment, rather than per tag-region. Neither do they consider the analysis order.

1.2. Three derivative classification issues

Ideally, only the high contribution degree tag-regions are analyzed for classification to improve classification performance and reduce computation time. But, it raises three issues:

1.2.1. The sufficient number of tag-regions for classification. The threshold to cut off noisy regions is still difficult to determine. A high threshold value (e.g. [2]) is that only a small portion of web document is included for classification, which may not be sufficient for correct classification. A lower threshold (e.g. [3]), results in the inclusion of noisy terms in the classification process.

1.2.2. The sufficient training data for sound emphasis degree training. While treating each individual surrounding differently achieves best result in classification, the emphasis degrees of some surroundings may be incorrect due to unsound training phrase; the training samples of some surroundings may not exist, or only appear very few times, in the training data set. As a result, a classifier may conclude an incorrect outcome due to unreliable emphasis degrees.

1.2.3. The impact of emphasis trap. Emphasis trap is the situation when noisy tag-regions are embedded in high emphasis degree surroundings and its true classification contribution is under expectation. For example, for the web documents of “image processing” category, we may expect that a <H1> surrounded tag-region should contain high contributive terms, like “JPEG”, to confirm the corresponding category. If a <H1> surrounded tag-region contains noisy terms, like “Overview”, the classifier is likely to fall into “emphasis trap”. In a web document, the emphasis traps often occur repeatedly for the tag-regions with the same surroundings or in the same concept segment due to the authoring practices. Following the above example, the same web document may contain a trap surrounded by <H1> embedding a “Conclusion” term.

In the paper, we propose a novel tag-region based web document classification method, named Progressive Analysis Scheme (PAS), aimed at tackling the three issues. The details of the PAS are in Sec. 2. The experimental results, shown in Sec. 3, verify the high classification performance and low computation cost of PAS. Finally, Section 4 is the conclusion of this study.

2. Models of progressive analysis scheme

2.1. Classification process of PAS

The idea for the first issue is to efficiently and effectively extract sufficient number of high contribution degree tag-regions for classification. During classification, a web document will be divided into small tag-regions based on its DOM structure. These tag-regions are analyzed in descending order of the emphasis degrees of their surroundings. After a tag-region instance is analyzed, the embedded features of this instance and the already analyzed instances are aggregated and represented by a feature vector in VSM for similarity calculation with each category. The classification analysis proceeds until its corresponding category is confirmed, or its instances are all analyzed.

2.2. Basic surrounding model

The basic model of surrounding includes the list definition of surrounding type, as defined Sec. 1, and the emphasis degree. As mentioned in Sec. 1, the classification contribution of a tag-region is positively related to the emphasis degree of its surrounding. PAS follows this property, and defines the contribution of a tag-region to be its category similarity over the average category similarity of all tag-regions of the document. Before presenting the formula, we introduce the notations. A tag-region \( \mathcal{R}_{iP} \) can be read as the \( i \)-th instance of document \( P \) with surrounding \( S \). \( \text{Sim}() \) is the similarity function. \( C_P \) is the corresponding category of \( P \). Then the contribution of a tag-region can be represented as below.

\[
\text{Cont}(\mathcal{R}_{iP}) = \frac{\text{Sim}(\mathcal{R}_{iP}, C_P)}{\text{avg}(\text{Sim}(\mathcal{R}_{jS}, C_P))}
\]

(1)

In PAS, the implementation of the similarity function is flexible that one can apply schemes like the cosine of angle scheme [1]. The emphasis degree of a surrounding \( ED(S) \) is defined by averaging the contributions of tag-regions with the same the surroundings in the training data set, defined as follows.

\[
ED(S) = \frac{\text{avg}(\text{Cont}(\mathcal{R}_{iP}))}{|\text{Training set}|}
\]

(2)

2.3. Enhanced surrounding model

In the proposed enhanced surrounding model, several approaches are taken to alleviate the problem arisen from the second issue. The first enhancement is surrounding merge aiming to merge close surroundings. The second enhancement is enhanced emphasis degree to borrow training samples from similar surroundings.
2.3.1. Surrounding merge. The scheme merges close surroundings into a new merged surrounding, and the training samples of surroundings to be merged together support the training of the new surrounding. As a tag-pair determines a particular presentation style, the same visual effect may be maintained by exchanging the sequence of tag-pairs of a surrounding. Therefore, the enhanced surrounding model ignores the order relation of a list and uses the set representation. The enhanced model simplifies surroundings with same tag-pairs, instead of surroundings with same bounding tag-pair, because the contributions of surroundings with same bounding tag-pair may be dramatically different.

In most cases, two or more identical tag-pairs give same effect to the surrounding tag-region as one tag-pair, e.g. <B> tag-pair. However, there exist some exceptions, e.g. the <UL> tag-pair. The tag-pair set is expanded to include these nested tag-pairs, defined as follows. The notation \(<UL^i>\) indicates an i-th level of nested \(<UL>\).

\[ \Theta^i = \Theta \cup \{ TP^j \mid j \in \mathbb{N}^* \text{ and } j \geq 2 \} \]

The merged surrounding \(MS\) of a surrounding \(S\) is represented as a set of its tag-pairs, defined as follows.

\[ MS = \{ TP \mid TP \in \Theta^i \text{ and } \exists i, TP = S(i) \} \quad (3) \]

The emphasis degree \(ED\) of a merged surrounding \(MS\) is represented as follows.

\[ ED(MS) = \frac{\text{avg}_{Pc \text{ Training samples}} \left( \text{avg}_{MS \rightarrow \text{Merged}(S,i)} \left( \text{Contrib}(\mathbb{R}^S_{TP}) \right) \right)}{\text{avg}_{MS \rightarrow \text{Merged}(S,i)} \left( \text{Contrib}(\mathbb{R}^S_{TP}) \right)} \quad (4) \]

2.3.2. Enhanced emphasis degree. While the merged surrounding scheme is more relaxed than previous research [4], there might be still some merged surroundings that lack of samples for a sound training. To solve this problem, PAS proposes the enhanced emphasis degree for a sample-lacking merged surrounding to use supplementary training samples from similar surroundings.

For each merged surrounding, a vector representation in VSM is defined as \(V^{MS} = (v_1, v_2, \ldots, v_n)\), \(n = |\Theta|\), and each element \(v_i\) is defined as:

\[ v_i = \begin{cases} \frac{1}{w_i} \sum_{S \in \text{Training samples}} \left( ED(S) \right)_{TP} & \text{if } TP \in MS \\ 0 & \text{Otherwise} \end{cases} \quad (5) \]

The weight \(w_i\) of tag-pair \(TP\) is the averaged emphasis degree of surroundings with it. The similarity between two merged surroundings is defined as the cosine value of their feature vectors.

If the number of training samples of a merged surrounding is less than the threshold \(T_{\text{sound}}\), it borrows the training samples from similar surroundings in the descending order of similarities. The borrowing continues until the threshold constraint is met. The emphasis degree for a sample-lacking \(MS\) is obtained by averaging the weighted emphasis degrees of its supportive surroundings. And the weighting multiplier is the surrounding similarity between the \(MS\) and its supporter.

2.4. Emphasis degree adjustment

During the classification process, the process will be prolonged if the classifier falls into an emphasis trap (i.e. noisy tag-region being analyzed). To avoid falling into the same type of emphasis trap repeatedly, the possibly noisy tag-regions associated with this type of trap are deferred for analysis by reducing their emphasis degrees. In this way, other unanalyzed tag-regions may move ahead for analysis.

The emphasis degree of a surrounding and its standard deviation indicates their content homogeneity. A high deviation means high diversity of tag-regions and high probability of noisy tag-regions. Fig. 1 shows the deviation degrees of several surroundings from experiment data in Sec. 3 that the unit of x-axis is a standard deviation and y-axis is the percentage of tag-regions. For example, the purpose of surrounding \{BODY, CENTER, H1, HTML\} is to highlight the document topic that is very consistent in authoring conventions; therefore, it has the lowest deviation degree. On the other hand, for the surrounding \{A, BODY, HTML, P\} that is used to highlight anchor texts, the contents of the tag-regions vary substantially.

![Figure 1. Deviation degree of contributions.](image)

Based on the standard deviation of training samples, PAS defines a revision factor to adjust the emphasis degree that the new emphasis degree of a tag-region after adjustment is the current emphasis degree multiplied by the revision factor. The revision factor of a surrounding \(MS\) is represented as follows.

\[ F(MS) = \frac{ED(MS) - \text{std}_{\text{Merged}(S)}}{ED(MS)} \quad (6) \]

where \(\text{std}(\cdot)\) is the function of standard deviation. For sample-lacking surrounding, the revision factor \(F\), likes the enhanced emphasis degree \(ED\), is defined as the weighted average of revision factors of its supporting surroundings.

In addition, PAS also adjusts the emphasis degrees of tag-regions in the same concept segment of an emphasis trap that the revision factor is the same as shown in Eq. 6. The construction of concept segments is simply based on some tag-pairs that can form visual content separators, similar to [5, 6].
3. Performance evaluations

There are five specific categories that each category has 300 web documents that all the experiments in this paper are based on the categories and web documents. The first three categories, which are Image Compression (labeled \( C_1 \)), ATM (Asynchronous Transfer Mode) Network (labeled \( C_2 \)), and Wireless Network (labeled \( C_3 \)), that their web documents are acquired from the Google search engine. The other two categories, which are Course (labeled \( C_4 \)), and Project (labeled \( C_5 \)), that their web documents are acquired from the WebKB database [8]. Cross validation is used in the experiments that one sixth of the web documents are used as training data and the remainders as testing data.

In the experiments, we adopt five contestants, including the PAS classifier, the PAS classifier without emphasis degree adjustment (labeled as “PAS#”), the sequential classifier (which analyzes tag-regions sequentially from the beginning of a HTML document), the traditional full text classifier and the popular SVM classifier which consists of multiple binary SVM classifiers with RBF kernel are tested on the web document sets. Note that the similarity function of the first four classifiers employs the VSM scheme with the cosine similarity measurement. The results of classification performances are listed in Table 1. The performances include the macro-averaged precision ratio, macro-averaged recall ratio, and F1 score.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Performance</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAS</td>
<td></td>
<td>0.90</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td>PAS#</td>
<td></td>
<td>0.88</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>Sequential</td>
<td></td>
<td>0.88</td>
<td>0.84</td>
<td>0.86</td>
</tr>
<tr>
<td>Full text</td>
<td></td>
<td>0.87</td>
<td>0.79</td>
<td>0.83</td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td>0.91</td>
<td>0.88</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Among the first three classifiers, the first two PAS classifiers reap the benefit of tag-regions analysis sequence sorted in their emphasis degrees and have better performance than the sequential classifier. Furthermore, the fully fledged PAS classifier achieves the best performance in F1 score by adopting emphasis degree adjustment, while the SVM classifier have the best performance in precision ratio but suffer from low recall for being interfered by the noisy term. The traditional full text classifier has the worst classification performance in both precision and recall as expected.

The PAS is supposed to have high classification efficiency as it only analyzes sufficient number of tag-regions. In Fig. 2, we show the classification efficiency of the five classifiers in the relationships between the classification rate and the percentage of document content needed for category confirmation. It is formed by summarizing the results of the first three categories. For PAS, 55% of the testing web documents are classified by analyzing only 10% of their contents and 92% are classified by analyzing 50% of their contents. In addition, the PAS outperforms the other classifiers also.

![Figure 2. Classification rate vs. the required percentage of document content.](image)

4. Conclusions

In this paper, we propose a progressive analysis scheme (PAS) for web document classification. The goal is to improve classification performance and efficiency by only analyzing sufficient number of contributive tag-regions. PAS analyzes tag-regions in descending order of their emphasized degree until the document category is confirmed or the classification process fails. PAS scheme can only use 50% or less of the document content to successfully classify most of the web documents while keeping a reasonable precision rate.

5. References