Detecting the Noise from Web Pages Using Entropy Measure

Swe Swe Nyein
University of Computer Studies, Mandalay
sweswenyein@gmail.com

Abstract

The rapid expansion of the Internet has made Web a popular place for disseminating and collecting information from the web. The noisy items in web pages are one of the major problems to extract the main contents. It is also important how to detect noises and distinguish valuable information from noisy data within a single Web page. In this paper, we propose a noise detection technique is based on the Document Object Model (DOM) tree. In DOM tree, weight of each node calculated by tf-idf scheme is added in entropy measure to get the respective value, which will be compared with a threshold value. Those less than threshold value are regarded as noise. Experimental results on a range of datasets using precision and recall measure show that our framework can improve noise detection accuracy.

1. Introduction

More and more advance technology research and development uses the Web as its data source. Almost all Web pages contain significant amount of unrelated information such as navigation panels, copyright notices, banner ads, decoration picture, etc. mixed with the relevant information. The irrelevant information, which is called noisy information, negatively affects users of small display devices such as PDAs and mobile phones but they are functionally useful human browsers and necessary for web site owners, Moreover, it is time consuming when collect and mine information due to the large amount of information on the web. Many researchers have been developed noise elimination techniques.

In this paper, before eliminate the noise we need to detect the noise correctly. While noise filtering algorithms are usually used to improve the accuracy of induce classification models, our aim is to detect noisy instances to be inspected by human expects in the phase of data understanding, data cleaning and outliers detection.

In web environment, there are two kinds of noise: Global noises and Local noises. Global noises mean duplicated web pages, old versioned Web pages to be deleted, etc. Local noises are noisy items within a Web page. We focus on local noise.

The proposed technique based on the DOM based analysis and actual contents of the web page. For detection noise, entropy based measure is apply which can measure the important of the content weight. To calculate the weight of each node tf-*idf* scheme is used. We choose a threshold value which was defined by the experience researchers. If the entropy value of each node is less than the threshold value, then we regard this node as noise.

This paper is organized in the following sequence. A briefly survey of many researchers is presented in section II. The proposed system is dealt with in section III. The section IV describes the evaluation results. This is followed by the

conclusions in section V, ongoing research work in section VI and the references.

2. Related Work

Detecting and eliminating noise from web pages is an important problem. Various techniques have been developed to deal with this problem.

L. Yi et al. have proposed a new Style tree to capture the actual contents and common layouts (or presentation styles) of the Web pages in a Web site. Their method can difficult to capture the common presentation style for many web pages from different web sites in [3].

Another approach mentioned in S. H. Lin and J. M .Ho [7] was InfoDiscoverer system to discover informative content blocks from web documents. It first partitions a web page into several content blocks according to HTML tag <TABLE>. They considered only <TABLE> tag for blocking.

The approach described in C. Li et al. [1] extracted informative block from a web page based on the analysis of both layouts and semantic information of the web pages. They needed to identify blocks occurring in a web collection based on the Vision-based Page Segmentation algorithm.

In [5], P. S. Hiremath et al. proposed an algorithm called VSAP (Visual Structure based Analysis of web Pages) to exact the data region based on the visual clue (location of data region / data records / data items / on the screen at which tag are rendered) information of web pages. In [2] D. Cai et al. proposed a Vision-based Page Segmentation (VIPS) algorithm that segments web pages using DOM tree with a combination of human visual cues, including tag cue, color cue, size cue, and others.

P. M. Joshi et al. proposed an approach of combination of HTML DOM analysis and Natural Language Processing (NLP) techniques for automated extractions of main article with associated images form web pages. Their approach did not require prior knowledge of website templates and also extracted not only the text but also associated images based on semantic similarity of image captions to the main text in [4].

In [9], Y. Li and J. Yang proposed a tree called content structure tree which captured the importance of the blocks. In [6], S. Gupta et al. proposed content extraction technique that could remove clutter without destroying webpage layout. It is not only extract information from large logical units but also manipulate smaller units such as specific links within the structure of the DOM tree. In [8], Y. Fu et al. proposed a technique to discover informative content block based on DOM tree. They removed clutters using XPath. They removed only the web pages with similar layout.

Although most of the existing approaches are based on entropy measure for detecting noise, they used either <table> tag only or, tag in their system. In this paper both tags will be used. It is intended to detect noise more accurately.

3. The Proposed System

The scope of the work considered in this paper is to detect the noise from the web page. It is based on the analysis of actual contents of the Web pages in a given Web site. A web page usually contains main content blocks and noisy content blocks. Only the main content blocks represent the informative part that is really we want to know. To detect the noise from the web page, common evaluation measures are used in this paper. We considered the methodology for the noise detection under the following five steps:

- preprocessing task

- creating DOM tree
- calculating weight
- using entropy measure
- detecting noise

3.1. Preprocessing Task

Most of the HTML web pages are not well-formed. So we need to check HTML documents using html tidy tool. And then stop words and stemming words are removed. We also need to remove the unnecessary nodes such as script, style, or other custom nodes and so on.

3.2. DOM tree structure

DOM stands for document object model. It builds the entire HTML (and XML) document representation in memory. DOM trees are highly transformable and can be easily used to reconstruct a complete web page. DOM tree is a well defined HTML document model. The HTML Parser will be used to build a DOM tree from a HTML Web page. HTML documents contain HTML tags and plain text. HTML lets format text, add graphics, create link, input forms, frames and tables, etc. In a DOM tree, tags are internal nodes and the detailed texts, images or hyperlinks are the leaf nodes. Figure 1 shows some html segments and its corresponding DOM tree. In the DOM tree, we need to tidy some unnecessary nodes, such as script, style or other customized nodes. Since the DOM tree follows the HTML code exactly, it is common that two visually similar HTML document have completely different DOM tree structures. HTML DOM is a standard object model. It defines the objects and properties of all HTML elements. According to the DOM, everything in an HTML document is a node. HTML Web pages begin from the BODY tag since all the viewable parts are within the scope of BODY.

```
<html>
<head><title> sample </title> </head>
<body>
   <div id = “wrapper”>
      <a href=”#”></a>
      <img src=”img”/ >
      <span>text</span>
  </div>
<img src=”img”/>
  <a href=”#”></a>
<table>
…..
</table>
<span>text</span>
</body>
</html>
```

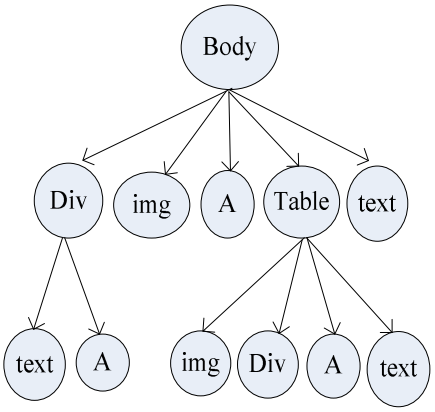

Figure 1. Segment of HTML code and its DOM tree

3.3. Calculating weight

This section describes how to calculate the weight of each node on the DOM tree using tf-idf scheme. The simplest approach is to assign the weight to be equal to the number of time that appears the term *t* in document *d*. This weighting scheme is referred to as term frequency denoted by tf_{ij}. Weight determined by the *tf* weighting function above or indeed any weighting function

that maps the number of occurrences of *t* in *d* to a positive real value. The *tf* scheme suffers a problem where a term appears in many documents of the collection. Instead, the document frequency df_i is use to define the number of documents in the collection that contain a term *t*. Denoting the total number of documents in a collection by *N*, the inverse document frequency (idf) inverse document frequency of a term t is given by Eq. (1)

$$idf_i = \log\left(\frac{N}{df_i}\right) \quad (1)$$

The *tf-idf* weighting scheme assigns to term *t* a weight in document d given by Eq. (2)

$$w_{ij} = tf_{t,d} * idf_i \quad (2)$$

From above the weight equation, we can measure the weight of each term in each element node on the DOM tree.
Each element node with their weighted term

	term 1..	term n
node 1	W_1	w_n
node 2	:	:
:	:	:
node 2	W_1	w_n

3.4. Entropy measure of Content Node

We calculate the entropy value of each element node (such as text, image, link) appearing in a block using their weight. The following is Shannon's famous general formula for uncertainty:

$$0 \le H = -\sum_{i=1}^{n} p_i \log_2 p_i \le \log_2 n,$$

where p_i is the probability of event.

By normalizing the weight of the node, the node entropy is:

$$H(node1) = -\sum_{i-1}^{n} w_{ij} \log_2 w_{ij},$$

where w_{ij} is the weight of each term in element node.

After getting the entropy value of each node, we can calculate the entropy value of each content block. The content block' entropy is the summation of its element nodes entropies as shown in the following equation.

$$H(ContentBlock1) = \sum_{i=1}^{k} H(node1),$$

where, *k* is the number of node in a content block

A content block contains different number of element node such as text, image, and link so the equation is normalized as:

$$H(ContentBlock1) = \frac{\sum_{i=1}^{k} H(node1)}{k}$$

3.5. Detecting Noise

Base on the entropy value, the content node can be divided into noise node or not. If the entropy value of a content node is less than threshold value, then we can define this node is noise node. The reasonable threshold value (0.5) is applied according to the experience researchers.

4. Evaluation

Most of the web site such as commercial and news sites contain significant amount of noise. We downloaded web pages from the commercial sites and news sites for detecting the noise. To evaluate the measurement of noise, precision and recall were calculated. Precision means number of correctly detected noise is divided by total number of retrieved noise. Recall means number of correctly detected noise is divided by total number of noise. The propose system correctly detect noise in web pages according to the threshold value. Firstly we created a sample web page. In this page, we could detect the noise 100 % in both precision and recall. Then, when we tested some Web pages from CBSNews and Commercial web sites by manually, the precision of detecting the noise is from over 70 % to around 90 % and recall is from 80 % to 90 %, respectively.

5. Conclusion

Detecting and eliminating noises from web pages is important to extract the useful information. We have proposed an approach for detecting noises. It is based on the DOM based analysis and actual contents of web page. We have measured the value of content nodes on DOM tree with entropy method. To correctly detect noise, each entropy value is compared with a threshold value which was tested by the experience researchers. We regarded nodes which are less than threshold value as noise nodes.

5. Ongoing research work

In real-world, noisy instances from web pages can lead to decrease performance of the web mining tasks. We intend to eliminate these noises and also extract the useful information by using the proposed algorithm.

References

[1] C. Li, J. Dong, and J. Chen, “Extraction of Informative Blocks from Web Pages Based on VIPS”, Journal of Computational Infromation Systems6:1(2010) 271-277.

[2] D. Cai, S. Yu, J. R. Wen, and W. Y. Ma, “VIPS: a Vision- based Page Segmentation Algorithm”, Technical Report, MSR-TR, Nov. 1, 2003.

[3] L. Yi, B. Liu, and X. Li, “Eliminating Noisy Information in Web Pages for Data Mining”, in Proc. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2003).

[4] P. M. Joshi, and S. Liu, “ Web Document Text and Images Extraction using DOM Analysis and Natural Language Processing”, ACM, DocEng, 2009.

[5] P. S. Hiremath, S. S. Benchalli, S. P. Algur, and R. V. Udapudi, “Mining Data Regions from Web Pages”, International Conference on Management of Data COMAD, India, December 2005.

[6] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm,”DOM-based Content Extraction of HTML Documents”, *Pro.* 12 th International Conference on WWW, ISBN: 1-58113-680-3, 2003.

[7] S. H. Lin and J. M .Ho, “Discovering Informative Content Blocks from Web Documents”, in Pro. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp.588-593, July 2002.

[8] Y. Fu, D. Yang, and S. Tang,”Using XPath to Discover Informative Content Blocks of Web Pages”, IEEE. DOI 10.1109/SKG, 2007.

[9] Y. Li and J. Yang, “A Novel Method to Extract Informative Blocks from Web Pages”, IEEE. DOI 10. 1109/ *JCAI*, 2009.

[10] http://www.w3.org/DOM

[11] http://www.tidy.sourceforge.net