DOCUMENT COMPARISON BASED ON THE PAGE LAYOUT

Martin Milička
Doctoral Degree Programme (1), FIT BUT
E-mail: xmilic01@stud.fit.vutbr.cz

Supervised by: Radek Burget
E-mail: burgetr@fit.vutbr.cz

Abstract: The paper suggests the pre-processing method for the document comparison based on visual features. The method tries to extract the basic layouts of the web pages. Afterwards, it uses these layouts in the comparison based on the web page layout that is the pre-processing part of the complete comparison process. The idea of pre-prosessing phase is based on the knowledge where the layout comparison is faster than the complex document comparison with visual features. It makes sense in the case where two documents with different layouts are in the comparison process.

Keywords: document comparison, document structure, visual features, layout detection

1 INTRODUCTION

Internet became a part of our live. It connects people around the world. It breaks the borders and distances. The WWW (World Wide Web) is the most used service of the Internet because it allows to present information such as documents, images, etc. Every day the number of this information is growing into incredible sizes. Due to the growing number of documents (information) we are using search engines that try to register them into some databases. The search engines go over the WWW and index the important information. Such created indexes are used by users for the information retrieving. However it brings a problem of effective information mining and following retrieving. There must be solved the problem of document duplicities as well.

The mining based on visual features is an interesting topic of data mining from the web pages. This approach can help to identify the meanings of the document parts. If the block identification is done in the right way there can be reduced the redundant information such as advertisements, web page menus, etc.

Another reason of document block identification can be a visual comparison process. Just visual processing has a connection to the human perception. Where we try to find the model that describes the web page as good as the human can perceive it. This comparison is mainly done because of the phishing. There are many cases when users want to log in into the internet banking but they are deceived by a different web page with the similar appearance.

2 VISUAL SIMILARITY ASSESSMENTS

Visual similarity assessment is a complex process that tries to cover all visual features that can be perceived by human. It is useful in the case when we want to substitute a human by a computer in the visual comparison process. Liu [5] proposes the way of similarity assesment for web documents. He suggests three metrics of web page similarity evaluation. These metrics are *block level similarity*, *layout similarity* and *overall style similarity*.

Block-Level Similarity. This similarity is defined as the weighted average of the visual similarities of all matched block pairs between two Web pages. At first, each web page is processed separately. It searches the main content blocks and it tries to categorize them as either text or image where each type has its own characteristics. In the categorization process it also extracts the specific features from the original block. For instance, the text-blocks extract information like Block content, Block colour, Block boundary, Block font, Block text and Block navigation. Also the image-blocks have special feature types like Block content, Block colour, Block size, Block source and Block navigation. Two blocks' total similarity is defined as a weighted sum of all individual features similarities. The values of features can be enumerative (font family, etc.) or continuous (for example, font size, colour, etc.). Two blocks are considered to match if the similarity value is over the given threshold. When all similarity values are obtained for all block pairs then the matching schema must be found between the two pages' blocks.

Layout Similarity. This approach has two default parts. The first one tries to match the several blocks with identical contents. Afterwards, the second part uses *neighbourhood relationship model* to match other blocks according to the spatial relations of all existing blocks. Two blocks are marked as matched if they have high visual similarity and the same position in the web page (layout relationship). Finally, layout similarity is calculated as the ratio between the weighted number of matched blocks and the number of total blocks in the genuine web page.

Overall Style Similarity. The researchers discovered that users ignore details (graphical and textual). If the styles of two pages are too similar, most people have problem to decide which page is genuine and which is the phishing one. The overall style similarity is an important feature because it gives a complex view of web page with a human perception. For example, the visual styles of Web pages are represented by following types of features - Page content, Page colour, Block boundary, Page font and Page text. Each feature type contains a group of features (For instance: Block boundary contains features as border style and border weight, etc.). Finally, the overall style similarity between two pages is defined as a correlation coefficient (range of [0,1]) of the pages' histograms with the style feature values.

3 WEB PAGE LAYOUT

Web page is represented by document object model (DOM). This representation has a tree structure. Each element may have zero or more child elements and it creates the hierarchy. There is a connection between the specific element from the source code and the element that is showed in the rendered web page.

In order to simplify the following description we are defining *Block*. It is a substitution of default element that has the specific features.

3.1 BLOCK DEFINITION

Block is defined as a default element of the web page layout. All blocks of layout are defined at the same level of document hierarchy. It implies that all blocks are children of one root element. We can see different levels of block resolution in figures 1 and 2. Figure 1 shows the most abstract layout of web page. The block detection process found three blocks where two of them are from two columns layout and the last one is a top part of web page. In the second figure 2, we can see the similarity with the first figure. However, the second method displayed in figure 2 detected smaller blocks. Of course, the second method can bring more appropriate approach in some cases because it recognizes more blocks. Each block is represented its width, height and position. This method does not use any

Figure 1: two columns layout

Figure 2: two columns layout with smaller blocks

visual features because this method is used as a pre-prosessing part of the proper comparison based on visual features.

In the figures, you can see that the method does not want to find the elementar elements. It searches the logic groups of these elements that are packed into blocks (areas). These blocks create the layout. The number of basic elements that are merged into a logic block depends on the level of element recognition. We can see that the level of recognition in figure 1 is smaller than the recognition level in figure 2.

3.2 BLOCK TYPES

This subsection brings a definition of different types of blocks. There is not any formal definition of these blocks. However, it helps for a future description. In general, blocks have a rectangular shape in a document description.

Horizontal block. This block is characterized its shape. Its width is longer than height. In figure 1, we can find this block at the top of web page.

Vertical block. Vertical block is defined as a block that has its height size longer than width size. Two vertical blocks can be seen under horizontal block in figure 1.

4 LAYOUT STRUCTURE EXTRACTION

Every web page is described its document object model that has its appropriate visual appearance. Our approach tries to find the main blocks of document that describe a document layout. The extraction of blocks is based on document source code. There was introduced a similar approach by Hu et al. in [4]. They are using a graphical method to extract the main blocks on the page. That research is concerned to segmentation and classification of paper documents.

As was mentioned before, our approach is much simple because it uses the source code. Also there is an advantage because the source code corresponds to visual appearance.

Basically, we are searching the elements that create the logic groups with similar content. For instance, we are searching a head part of page, columns in the document, etc. Usually, it is not hard to go through document object model and define the blocks at first level as the retrieving blocks. The problem starts when we want to define a different level of recognition. In this case, we have to walk deeply in DOM to find the smaller parts.

The layout extraction expects a position definition and size definition. The position of blocks has to be stored in the specific format. The information about block sizes and their positions can be reached by CSSBox [2]. We are storing the block positions in a multi array. The inspiration comes from Burget [1] who introduced a grid for a document description. Our approach uses the same methodology. Detail information can be seen in figure 3. The idea comes from the grid's cell occupation where each block occupies the specific cells.

Finally, the document structure is described by the following parameters:

- Document width and height
- Cell height and width
- List of blocks with their occupations in the grid

For instance, the list of blocks from the figure 3 is:

1.block { {1,1}, {1,2}, {1,3} }
2.block { {2,1}, {3,1}, {4,1} }
3.block { {2,2}, {3,2} }
4.block { {4,2} }
5.block { {2,3}, {3,3}, {4,3} }

5 LAYOUT COMPARISON

The comparison process expects the segmented web page in rectangular blocks. Basically, the blocks from the segmented document are partitioned into a *m* by *n* grid. Details of block occupation can be seen in figure 3.

The inputs of layout comparison are structures that were introduced in section 4. In general, these structures may have different parameters. Therefore, the documents have to be normalized into the similar width and height in the beginning. All parameters have to follow the defined ratio as well.

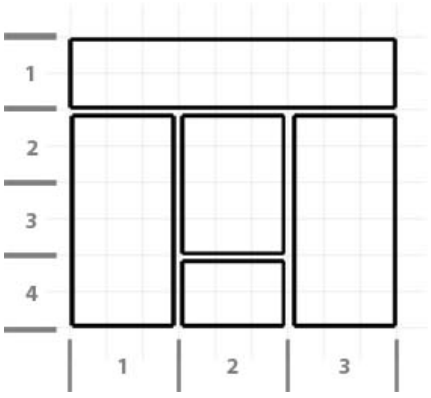

Figure 3: Layout grid

If the document structures are normalized they have similar width and height. Afterwards, there can be done the comparison. The solution is really straightforward because we are searching the overlapping between two layouts. In fact, there is searched the number of similar sub arrays (positions in the grid) between the lists of blocks. Of course, we have to keep in mind that this approach cannot

be used as standalone method for a document comparison based on visual features. This method is designed as a preprocessing part of total comparison method based on visual features.

6 FURTHER RESEARCH

The proposed approach will be a part of complete document comparison based on visual features. Due to a complexity of visual properties that have to be solved in the comparison process we proposed the layout structure comparison. It should speed up a decision of document similarity in the cases where the documents are completely different.

In the future approach, the comparison method will contain more information about the visual appearance. There can be information about colours, fonts, borders, etc. In fact, there will be satisfied the features that were introduced in section 3.

The layout comparison can be used for the information mining as well. It is based on the knowledge of block meaning where we try to get just information from the specific blocks.

7 CONCLUSION

This paper suggested a method of document comparison based on the document layout structure. In the beginning, we got some motivation for a document comparison and then we introduced the metrics of document comparison based on visual features. Due to that complexity we introduced the comparison method based on a document layout structure. Just this comparison may speed up a process of comparison if the documents have completely different structures.

The standalone layout comparison method is not force enough for a document comparison based on visual features because there are missing the important visual features in a comparison process. However, this method can be used in the information mining where we want to extract the information from the specific blocks in the web page.

ACKNOWLEDGEMENT

This work was partially supported by the research plan MSM0021630528, the specific research grant FIT-S-11-2 and the IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070.

REFERENCES

[1] Burget, R. Vizuálně orientované modelování dokumentů na WWW. Datakon, 2006, s. 159–168.

[2] Burget, R. CSSBox rendering engine. Available at: http://cssbox.sourceforge.net/ [online] (February 2012).

[3] Fu, A. Y., Wenyin, L.: Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD). IEEE Transactions on Dependable and Secure Computing, 2006, s. 301–311

[4] Hu, J., Kashi, R., Wilfong, G. Document Image Layout Comparison and Classification. In Proc. of the Conf. on Document Analysis and Recognition, 1999, s. 285–288

[5] Liu, W., Deng, X., Huang, G., Fu, A.Y. An antiphishing strategy based on visual similarity assessment. Internet Computing, IEEE, 2006, vol. 10, iss. 2, s. 58–65