Measuring Web Page Similarity Based on Textual and Visual Properties

Vladimír Bartík*

Brno Universtiy of Technology, Faculty of Information Technology, IT4Innovations Centre of Excellence.
Božetěchova 2, 612 66 Brno, Czech Republic
bartik@fit.vutbr.cz

Abstract. Measuring web page similarity is a very important task in the area of web mining and information retrieval. This paper introduces a method for measuring web page similarity, which considers both textual and visual properties of pages. Textual properties of a page are described by means of modified weight vector space model. General visual properties are captured via segmentation of a page, which divides a page into visual blocks, properties of which are stored into a vector of visual properties. These both vectors are then used to compute the overall web page similarity. This method will be described in detail and results of several experiments are also introduced in this paper.

Keywords: Web Page Similarity, Clustering, Vector Space Model, Vector Distance, Term Weighting, Visual Blocks

1 Introduction

The amount of documents stored on the Web is still growing rapidly. There is a growing need for effective and reliable methods of information retrieval on the Web. The other important task is to organize and navigate the information. Therefore, we need to develop methods for web content mining.

This paper is focused on measuring similarity of web pages. The results can be used to search on the web, web page clustering or to reveal phishing web pages. If we are able to store web pages, which are frequently accessed by a user, it is possible to warn user, if a phishing page is visited by the same user.

There are several factors, which affect similarity of web pages. At first, we have to mention the hyperlink structure of web pages, which is the most frequently used factor to measure the similarity. There is also text content and visual structure of a document used for this purpose.

The main idea of the method proposed here is that the pages are similar only if their text content and visual properties are similar simultaneously. A new measure based on these two factors is introduced. To represent text contents of

* This work was supported by the research programme MSM 0021630528 and the IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070.

the pages, we use the vector space model with the modified TF-IDF weights for text terms. The visual properties of a web page are represented as another vector of properties. If we have these both vectors, we can compute the textual and visual similarity of web documents. The combination of both similarity measures forms the aggregate similarity of web pages.

In the first part of the paper, the way of getting necessary information from web pages and its representation is introduced. Then, the measures, which are used to compute the similarity from the information obtained by the previous step, are described. Finally, some results of experiments are proposed.

2 Related Work

The methods developed to count web page similarity differ in the information, which is used to compute it. In most of methods, hyperlink structure of a web document is used. In [1], to represent a page, keywords from all pages linking to a given page, are extracted and used for clustering algorithm. In [2], the hyperlink information is used to find related pages on the web, i.e. pages with similar topic.

Another possibility to represent web page content is to extract its textual information and use some of text mining methods. Typically, the Vector Space Model is used to represent text content. For example, TF-IDF can be used as a weighting method to express the term's significance [3]. A method, which uses only the proper names from the text, is proposed in [4]. Proper names are presented as more significant for the task of similarity search in this paper.

The visual layout of the page is also very important aspect of web page content. The basic idea is that similar pages would also have a similar visual layout and structure. In [5], the visual similarity is deduced from frequencies of individual HTML tags. Here, the information about general web page's structure and layout is acquired in a form of tag frequency. This brings some interesting results, but it does not affect the order in which tags appear in pages. In the visual representation based approach proposed in [6], a tree representation of the HTML structure is used as the input to count web page similarity.

Another possibility to extract visual information of web pages is to use page segmentation methods. Segmentation is a process, which divides a web page into visual blocks that are separated from each other. Segmentation algorithms usually work in hierarchical manner. This means that in the beginning, a page is divided into a few blocks, which are divided furthermore into smaller ones (top-down); or a small pieces of a page are joined together to form the greater ones (bottom-up). Probably the most popular segmentation algorithm working in top-down manner is VIPS (Vision-based Page Segmentation) [7]. Another segmentation method, which works in bottom-up manner, is presented in [8].

The tree of visual blocks obtained by segmentation can also be used to determine the web page similarity. Some results of this approach and comparison of different segmentation methods are presented in [9].

As the results of above mentioned methods are not suitable, several proposed methods focus on the combination of the above mentioned approaches to

represent the web page content. Such approaches typically try to combine the information about hyperlink structure and textual contents of web pages. One of these methods [10] creates a vector with information about text content and links with different weights. This is used as an input to cluster pages.

This paper presents a new method, which uses a representation of web page content, which is combination of text content and visual structure extraction. An approach regarding this combination has been described in [11]. The structural information is taken also from HTML tags. Tags are divided into several classes. A weight is set for each class, according to its importance.

In our paper, we use information obtained by web page segmentation instead of HTML tags, because of the fact that the formation of various web pages is different and the same information can be represented in different ways in HTML. To represent textual contents of a page in a form of weight vectors, we use the modified TF-IDF weights, which affect the visual blocks, in which the text terms appear. This technique has been used for web page classification and it has brought some improvement of classification accuracy [12].

3 Description of the Page Similarity Method

The proposed method uses two vectors to measure similarity. The first one is the vector of modified TF-IDF weights, which reflects the textual content of a page. The second one is the vector of visual properties obtained by the segmentation.

3.1 Method Overview

At first, it is necessary to perform web page rendering, which takes the page source code and returns its visual layout. Then, this visual layout is an input for web page segmentation. The output of the segmentation process is a set of visual blocks, each of which has a set of properties assigned. This includes its position on the page, some other visual properties and text contained in the block.

This is then used to perform visual block classification, which is used to assign a class for each visual block. Each class has different significance for the web page representation. The results are then used to compute the modified TF-IDF weights for text of all visual blocks from the whole web page.

If we are able to create these vectors for each page, it is possible to count the textual similarity of pages. Some visual properties taken from the results of web page segmentation process are used to count the visual similarity of pages. According to some experiments, a set of suitable properties has been chosen. These properties are stored in a form of the second vector called visual vector.

Finally, we can combine these two similarities into the aggregate web page similarity, which is the main result of our method.

3.2 Visual Block Clasification

The first step of the process is using of a web page rendering machine, which interprets an input document and shows its final form. This allows analyzing

the visual properties of component blocks of a page and obtaining necessary information about them.

Information about visual blocks is obtained by a segmentation method that produces hierarchical structure of visual blocks. A visual block represents a rectangular region in the page that is visually separated from other parts of the page. Our segmentation algorithm presented in [8] works in a bottom-up manner.

The resultant tree of visual blocks is created recursively: if a block is visually separated, a new corresponding block is added to a tree. Then, the same is applied to the child boxes. After that, the blocks are clustered - we find all visual blocks that are placed in the adjoining cells and are not visually separated from each other. Such areas are joined into a single area. This step corresponds to the detection of content blocks (for example text paragraphs) that consist of several boxes. Finally, we look for areas that are not separated but they are delimited with the visually separated areas around. These blocks are also clustered.

For each detected visual block, we can determine its visual properties, which are then used for classification of the blocks as input attributes. These attributes include: text font size, dominating text weight and style, count of text/number characters, count of lower/upper case/space characters, average luminosity, background color, contrast of the text and position of visual block.

If we have this set of attributes obtained, we can use arbitrary classification method to perform visual blocks classification. Our experiments presented in [14] showed that the classification accuracy is above 90%. This can be a good foundation for further processing of this data. The best results have been achieved by decision tree based methods. Here is the list of classes assigned to visual blocks:

- Heading - The main heading and subheadings of a page.
- Main Text - The main text content.
- Links - Links to other related web pages, some of them may be irrelevant.
- Navigation - Link to other sections of the web site.
- Date/Authors - Information about date and authors of a page.
- Others - Other unimportant parts, such as advertisement, caption etc.

Textual Similarity Measuring It is obvious that different classes of visual blocks have different significance for representation of page content. This is reflected by our weighting method.

In the beginning, the text information is stored for each visual block. After the visual block classification, a class for each block is known. Then we have to perform two standard preprocessing procedures for the text contents of all visual blocks - stop words removal and stemming. If we set a significance coefficient for each class, it could be reflected in the text term weighting.

This is ensured by modified TF-IDF weighting, description of which follows. Let us denote an input set of web documents as $D = \{d_1, ..., d_n\}$ and a set of terms $T = \{t_1, ..., t_m\}$, which appear in documents from the set D.

It is possible to divide each document into several visual blocks, each of which can be classified. Let us denote a vector of all class labels as $C = (c_1, ..., c_k)$. Each class label is evaluated by a coefficient according to its importance for

representation of page contents. We can represent it as a vector of coefficients $V = (v_1, ..., v_k)$, where v_j is the coefficient of a class label c_j. The modified document frequency of a term $t \epsilon T$ in a document $d \epsilon D$ is defined as:

$$MTF(t, d) = \sum_{i=1}^{k} v_i * F(t, d, c_i) \quad (1)$$

where $F(t, d, c_i)$ is a frequency of term t in all blocks with class label c_i in a document d. The MTF weight is obtained as a summarization of all weights for visual blocks, in which the term is present. This weight should be normalized, as it is usually done for the TF weight. Modified IDF weight is obtained as:

$$MIDF(t) = 1 + \log(\frac{n}{k_V}) \quad (2)$$

where t is a term from the set of terms T, n is the count of all documents and k_V is the count of documents, in which content visual blocks at least once contain the term t. The resultant modified TF-IDF weight is obtained as a multiplication of modified TF and modified IDF weight.

This allows us to omit the non-content parts of a web page, if we set zero significance coefficients for respective class of visual blocks. We can mention "Navigation" as an example of a non-content block. On the other hand, "links" are considered as content block class with lower significance coefficient. A suitable setting of significance coefficients has been suggested in [12].

The textual similarity of two web documents represented as modified TF-IDF weight vectors can be computed by means of some similarity measure. We use the Cosine similarity, which is typically recommended for this purpose.

Visual Similarity Measuring To represent visual layout and structure of a page, we use the results of the visual block classification. These results include information about position, visual properties and class of each individual visual block. This is used to create a vector of visual properties containing attributes that characterize overall layout of a page. Here is the list of properties used:

- Font size, font name and color of the main heading - the main heading is discovered as a block of class "Heading" with the highest font size.
- Font size, font name and color of the main text - dominating values among all blocks of the class "Main text".
- Top-left position of the main heading and the main text.
- Amount of main text and the heading text, both expressed as the count of all characters inside these blocks.
- Length of the main heading and the count of all headings placed on the page.
- Dominating color of the main text background.
- Co-ordinates of the bottom-right corner of the main text, obtained as maximum value of co-ordinates between all "Main text" blocks.
- Minimum and maximum positions of the "Links" blocks - co-ordinates of top-left and bottom-right links.

- Amount of links expressed as the count of characters, which occur in all blocks of the "Link" class.
- Amount of text (number of characters) on the whole page (including the non-content visual blocks) and number of digit characters on the whole page.
- Number of different background colors and text colors on the whole page, including all content and non-content blocks.
- The "height" of the page, expressed as the maximum horizontal co-ordinate of any visual block of the page.

Then, the vectors consisting of all these properties can be used to measure the visual similarity of web pages. We can use arbitrary distance metric for this purpose, for example the Cosine similarity mentioned above.

3.3 Aggregate Web Page Similarity Computation

In the next phase, we have to aggregate both similarities into one similarity measure reflecting both visual and textual properties of pages. If these two similarity values are comparable, it is straightforward - we have to count up those two values to obtain the aggregate similarity measure. But during experiments, it turned out that the differences between values of textual similarity are much smaller than between values of visual similarity. That's why the aggregate similarity was more reflected by the visual similarity value.

We have to adjust the influence of both similarities to the aggregate similarity. This is enforced by the different powers applied to these two measures. Therefore, the aggregate similarity of two different web pages is defined as:

$$Agg\_Sim = Text\_Sim^M * Vis\_Sim^N \tag{3}$$

where Text_Sim and Vis_Sim are the textual and visual similarity values, M and N are the powers to be applied on them. The higher is the value of power; the lower is the influence of that value to the aggregate similarity value.

4 Results of Experiments

Experimental phase of our project consists of two parts. In the first one, we use a dataset of web pages coming from news web servers (CNN.com, usatoday.com, reuters.com, nytimes.com). This dataset contains approximately 500 web pages of various topics.

The second dataset is a small dataset of phishing pages taken from a PhishTank database. This experiment is only used to verify that our approach is applicable to discover this type of fraudulent web pages.

4.1 Experiments with News Web Pages

As mentioned above, this dataset comes from several news web sites and covers several topics. It is clear that this dataset is suitable to perform experiments regarding both textual and visual similarity.

Experiments of Textual Similarity There are many topics of web pages in our dataset, but there is a subset of pages about specific topics, which are expected to form clusters. There is a set of pages about war in Libya, earthquake in Japan and NHL playoffs. In our experiment, we will test if these three groups will form clusters. Other pages should be significantly different from these ones.

Our modified weighting approach will be compared to a simple text extraction with classical TF-IDF weighting. Because some text on the web page is non-content, it is expected that our approach will bring some improvement.

This experiment will be evaluated as follows: as queries, we will specify documents from those three clusters mentioned above. The ideal result of this experiment would be a response of exactly those (approximately 20) documents from the same cluster. In Table 1, the values of precision and recall values are compared for each of three clusters separately.

Table 1. Textual Similarity Experiment Results

	$MTF-IDF$		$TF-IDF$	
Cluster	*Precision*	*Recall*	*Precision*	*Recall*
Cluster 1 - Earthquake in Japan	91.2%	90.0%	77.5%	68.0%
Cluster 2 - Libya war	88.2%	86.0%	73.5%	66.8%
Cluster 3 - NHL playoffs	84.8%	89.9%	62.2%	60.0%

As we can see from the table, modified weighting of text terms brings significantly better results in measuring textual similarity. This is primarily caused by ignoring text from the non-content blocks of pages, even though the visual block classification could produce a small number of mistakes. Slightly worse results for the third cluster are caused by some web pages about other kinds of sport, which have been determined as similar to those pages from cluster 3.

It is evident from these results that this representation without computing visual similarity could be used to measure web page similarity because of higher precision and recall values.

Experiments of Visual Similarity In this experiment, we assume that pages from the same web site have also high visual similarity. This assumption could not be valid for some attributes reflecting amounts of text mentioned in Section 3.4. These attributes will be omitted in this experiment, although they are important in the general process of measuring the similarity of web pages.

Here we assume three clusters - cluster of pages from CNN.com, pages from usatoday.com and reuters.com. The results are shown in Table 2. From the table we can see that this computation cannot be used independently to detect similarity of web pages. But it is applicable as a supporting mean in combination with textual similarity. In addition, worse values of precision and recall are par-

Table 2. Visual Similarity Experiment Results

Cluster	*Precision*	*Recall*
Cluster 1 - CNN.com	84.4%	71.2%
Cluster 2 - usatoday.com	71.0%	69.4%
Cluster 3 - reuters.com	73.5%	70.4%

tially caused by much wider clusters of pages than in the first experiment. This experiment has been evaluated in the same way as the previous experiment.

Experiments with Aggregate Similarity In our manually created dataset, there are also some clusters which have the same topic, i.e. documents have high textual similarity and concurrently they are from the same web site.

The results presented in Table 3 show the results for three clusters having these properties. The way of evaluation is the same as in the previous two subsections. We can see that there are slightly better results than that achieved by the textual similarity measure. These results are measured, if the coefficients M and N from the formula (4) both have the value 1.

Table 3. Aggregate Similarity Experiment Results

Cluster	*Precision*	*Recall*
Cluster 1 - Japan, CNN	92.5%	90.6%
Cluster 2 - Libya, Reuters	90.5%	84.3%
Cluster 3 - NHL, usatoday	89.6%	93.4%

4.2 Experiment with Phishing Web Pages

We have made only a small experiment with phishing pages. It is clear that text and visual layout of phishing page and original page should be very similar. The only objective of this experiment is to prove that our similarity measure will reach significantly higher values for pairs of phishing and original pages than for other pairs of pages used to login into some internet banking applications. The values of both textual and visual similarity between the original and phishing login page have been significantly higher (close to 1.0) than both similarities for two different pages.

5 Conclusion and Future Works

In this paper, we have presented a new way to measure similarity of web pages. The main idea is to compute textual and visual similarity of pages separately

and then to join them into the aggregate similarity reflecting both similarities. To compute textual similarity, we use the modified text term weighting, which reflects the blocks, in which the text appears. Visual similarity is obtained by means of vectors of visual attributes. These attributes are obtained with use of rendering machine and web page segmentation algorithm.

In our future works, we are going to use web page representations proposed here to perform other web mining tasks. Another possibility is to extend our representation with hyperlink information.

We are also going to find some optimal settings for our method. This includes settings of coefficients M and N, which should be different for various types of pages and selection of suitable similarity measure for both similarities mentioned.

Finally, it may be convenient to find some better representation of visual layout, because the representation proposed here can be used only as a supporting mean, but not as an independent measure of similarity.

References

1. Halkidi, M., Nguyen, B. Varlamis, I. and Vazirigiannis, M. 2003. Thesus: Organizing web document collections based on link semantics. VLDB Journal, 12(4): 320-332.
2. Dean, J. and Henzinger, M. 1999. Finding related pages in the World Wide Web. In Proceedings of the 8th WWW Conference, Toronto, Canada, 1467-1479.
3. Salton, G. and Buckley, C. 1998: Term weighting approaches in automatic text retrieval. Information Processing and Management, Vol. 24, 1998, 513-523
4. Sannella, M. J. 1994. Constraint Satisfaction and Debugging for Interactive User Interfaces. PhD. Thesis. UMI Order No. GAX95-09398., University of Washington.
5. Cruz, I.F., Borisov, S., Marks, M.A. and Webb, T.R. 1998. Measuring structural similarity among web documents: preliminary results. In Proceedings of the 7th International Conference on Electronic Publishing, ICCC Press, Washington D.C., USA, 513-524.
6. Joshi, S., Agrawal, N., Krishnapuram, R. and Negi, S. 2003. A bag of paths model for measuring structural similarity in web documents. In Proceedings of the 9th ACM SIGKDD Conference, ACM, Washington D.C., USA, 577-582.
7. Cai, D., Yu, S., Wen, J.-R., and Ma, W.-Y. 2004: VIPS: a Vision-based Page Segmentation Algorithm. Technical Report MSR-TR-2003-79, Microsoft.
8. Burget, R. 2007: Automatic document structure detection for data integration. In Proceedings of Business Information Systems (BIS 2007), Lecture Notes in Computer Science, Vol. 4439, Poznan, Poland, 391-397.
9. Cai, D., Yu, S., Wen, J.-R., and Ma, W.-Y. 2004: Block-based Web Search. In The 27th Annual International ACM SIGIR Conference on Information Retrieval, ACM, Sheffield, UK, 440-447.
10. Modha, D.S. and Spangler, W.S. 2000. Clustering hypertext with applications to web searching. In Proceedings of the 11th ACM Conference on Hypertext and Hypermedia, ACM, San Antonio, USA, 143-152.
11. Cutler, M., Deng, H., Maniccam, S.S. and Meng W. 1999. A new study on using html structures to improve retrieval. In Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, IEEE, Chicago, USA, 406-409.
12. Bartik, V., 2010. Text-Based Web Page Classification with Use of Visual Information. In International Symposium on Open Source Intelligence & Web Mining, IEEE, Odense, Denmark.