Pattern Recognition for Finding Similarity of Web pages

[1]N. Pughazendi , [2]G. Pattusamy Student
[1]Associate Professor, Department of Computer Science, Panimalar Engineering College
[2]PG scholar, Department of Computer Application, Panimalar Engineering College

Abstract- We proposed a functional technique for identifying similar Web pages that is based on measuring tree similarity. In this paper we introduce an experiment with two methods for evaluating the similarity of web pages. The results of these methods can be used in different ways for the reordering and clustering a web page set. Both of these methods belong to the field web content mining. The first method is purely focused on the similarity of web pages. This method segments web pages and compares their layouts based on the image processing and graph matching. The second is based on detecting of objects that result from the user point of view on the web page. The similarity of web page is measured as an object match on the analyzed web pages. The key idea behind the method is to transform each Web page into a compressed, normalized tree that effectively represents its visual structure.

Keywords: Tree structure, Clustering, Web comparison, Edit distance.

I. INTRODUCTION

The idea of information visualization is to gain insights from great amounts of abstract data. Comparing document sets that are found by searching the World Wide Web is a new challenge for visualization–related techniques. Typical scenarios include automatically identifying documents that are visually similar to ones previously viewed or fixed by the user. Also, considering visual similarities can sometimes improve Web page clustering quality [1].Apart from sheer volume, one particular vexing problem of information retrieval (IR) and related techniques such as filtering or categorization is that they mainly deal with text. However, the possibility to recognize heterogeneous depictions of pieces of information that have similar semantics is a common situation in the Web that should be taken into account. Contemporary research has begun to supplement basic IR approaches with techniques that collect indicators of "information value" and are able to assess semantic similarities [2]. Unfortunately, one factor that is hardly considered by state-of-the-art Web handling techniques is whether two different pieces of code can express the same visual sensation.

When we look at a Web page, we are not aware of the underlying HTML code and we are only able to distinguish the visual structure given by the groupings, columns, rows, and data page" as the apparent structure of the Web page that is perceived by a human independently of the source code that produces it. The ability to assess visual similarities in an accurate, automated, and scalable way can be a key determinant of the effectiveness of information handling and decision support software that deals with the WWW. For example, in [3], several methods are proposed to determine the structure of a given Web page. A ranked list of possible structures for the page is produced, which can be used for several important applications. For instance, it cans be used to define suitable wrappers1 for Web-based information integration systems. As reported in [3], by adding some visual requirements regarding to the Web page that we are interested in exploring, the fitness of the proposed top ranked structures is meaningfully improved with respect to an earlier ranking that lacks this extra information. As a second example, in the scenario of Web security, the visual similarity of Web pages can also be applied to phishing detection [4]. A novel application of the techniques for recognizing the visual information within Web pages is presented. There, Web pages are considered to comprehend not only the information but also social aspects where there is a hidden interaction among users, developers, and Web site owners. The paper then establishes that the visual perception is usually in dependent from the user's ability (e.g., it is expected that a sales page can be recognized by an Arab-speaking user, a Chinese-speaking user, or an English-speaking user).Thus, everybody should be able to use the functionalities of the Web page. Moreover, the visual information that can be extracted from the pages that users visit could be productively used to determine the social groups of the users themselves. In this manner, this information could be used to improve the accuracy of search engines. The above-mentioned applications have motivated us to investigate the problem of Web page visual similarity. To the best of our knowledge,

the recognition of the visual structural information is an area that is still unexplored by this aspect.

We developed a preliminary technique for Webpage comparison that takes into account their visual structure. The key idea behind the method is to transform each Web page into a compressed, normalized tree that effectively represents its visual structure. The transformation is based on a classification of the set of html–tags that is guided by the visual effect of each tag in the entire structure of the page. Then, a metric to compute the distance between two Web pages that is based on the classical tree edit distance algorithm was formalized. This algorithm defines the dissimilarity of two trees by determining the minimal cost of the edit operations (node insertion, node deletion, and label change) that are needed to transform one tree into the other. A preliminary implementation that is written in Maude [10] was also presented in [8]. Unfortunately, the vast (fix point) computation that is involved in the Web comparison algorithm of [8] leads to unsatisfactory performance, so the prototypical tool can only efficiently process very small structure even though their respective HTML encodings are plainly different. The tool computes a similarity measure of 92% between the two Web sites in a very short time. The second case study is a Web document clustering. The aim was to identify a set of Web page templates within a large collection of Web pages.

II. SYSTEM ANALYSIS

Existing System

Social networking can be both overwhelming and addictive at the same time. You'll have a lot of noise to filter out if you want to find something specific. And you may find yourself checking for updates several times throughout the day when you really should be doing something else. Given a query for retrieving a piece of information from the web, the search for this information typically involves three aspects: textual information within the Web page, page structure layout, and the patterns of the query. However, an extra factor that is hardly considered by current tools is whether two different pieces of code can express the same visual sensation. When they look at a Web page, we are not aware of the underlying HTML code, but are only able to distinguish the visual structure given by the groupings, columns, rows and data. This suggests us the idea to define as “visual structure of a Web page” the apparent structure of a Web page that is perceived by a human independently of the source code that produces it. The horizontal compression packs together those sub terms which represent repetitive structures. The vertical compression shrinks those chains of tags that does not influence visually the perceived result.

Disadvantages

- We are not aware of source code of the web pages.
- One particular vexing problem of information retrieval (IR) and related techniques such as filtering the web content.

Proposed System

We developed a preliminary technique for Web page comparison that takes into account their visual structure. The key idea behind the method is to transform each Web page into a compressed, normalized tree that effectively represents its visual structure. The transformation is based on a classification of the set of html–tags that is guided by the visual effect of each tag in the entire structure of the page. Then, a metric to compute the distance between two Web pages that is based on the classical tree edit distance algorithm was formalized. This algorithm defines the dissimilarity of two trees by determining the minimal cost of the edit operations (node insertion, node deletion, and label change) that are needed to transform one tree into the other. A preliminary implementation that is written in Maude was also presented in page. Unfortunately, the vast (fix point) computation that is involved in the Web comparison algorithm of leads to unsatisfactory performance, so the prototypical tool can only efficiently process.

Advantages

- It analyzing the source code of web pages and filtering the web content.
- This project used while designing and developing a webpage.
- It visualizes the set of HTML code for the web pages.

III. SYSTEM ARCHITECURE

The main purpose of this project is to identify the similarity of two different web pages.

This architecture are represents the user after entering URL to the Google search engine then filtering the web content and analyzing source code. This system used to designing and developing web pages. The Matching descriptors yield a similarity degree for a suspect page and an authentic page. Finally, we use the similarity degree between the two pages.

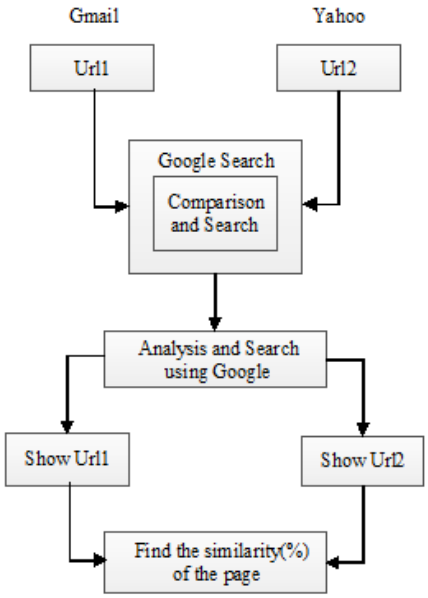

Figure 3.1 System Architecture

IV SYSTEM IMPLEMENTATION

The modules in this process are

- Translation
- Canonical representation
- Web page compression
  - Horizontal compression
  - Vertical compression
- A similarity measure between terms

Translation

ur translation technique is based on a classification of the set of html–tags, which is guided by the visual effect of each tag on the whole structure of the page, that allows us to infer the visual structure of the page from the HTML tags in it.

The signature of abstract(visual) HTML tags, where grp, col, row, and text can be seen as the "abstraction" of a number of different concrete HTML tags according to the abstraction classification.

It is straight forward to translate the Web pages into ordinary terms. This translation brings to light the repetitive structures that occur on the page and gets rid of those chains of tags that do not influence the visual aspect of the page.

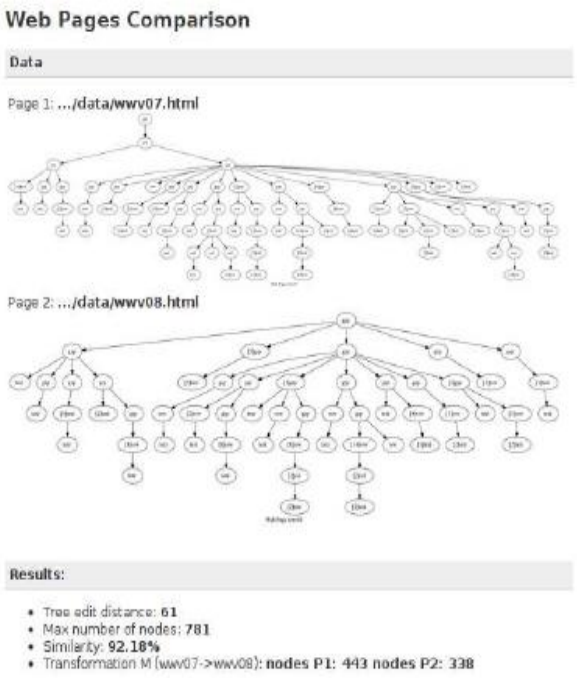

Figure 4.1 Translation

Canonical representation

Canonical representational order to avoid losing information after compressing Web page, we need to record the approximate number of nodes before applying the compression. Let us introduce the notion of marked term [N]t, which allows us to record the number N of "similar" sub terms that appear in a given term. For example, consider the terming Figure (a). The corresponding marked term is shown in Figure (b). The sub term "[2]row([1]text)" represents that the term row([1]text) appears twice. Note that this representation is not commutative, thus (in Figure (a)) the first sub term row(text) cannot be packed together with the last two. When no confusion can arise, we simply write grp([2]row([1]text)) = grp([2]row(text)).

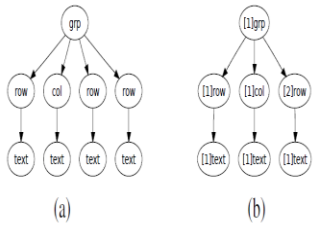

Figure 4.2 Canonical Representation

Web page compression

Reduce the complexity of tree comparison, two compression techniques that dramatically reduce the size of the terms.

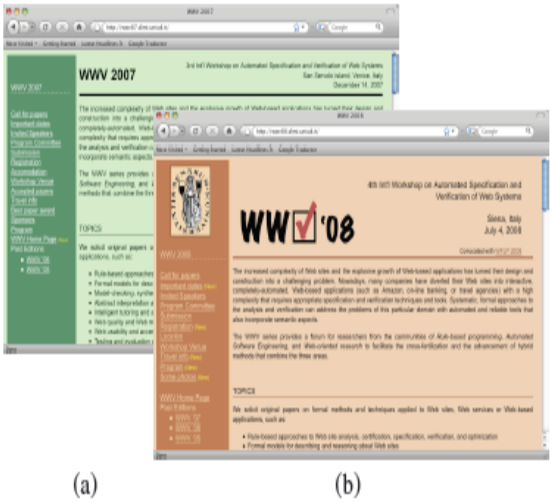

Figure 4. Web pages of workshops WWV'07 and WWV'08

Horizontal compression: Horizontal compression packs together those sub terms that represent repetitive structures

Vertical compression: Vertical compression shrinks those chains of tags that do not visually influence the result.

A similarity measure between terms:

The problem of comparing Web pages essentially boils down to comparing trees. In the literature, the most widely used approach for comparing trees consists in computing the "edit distance" between the two rooted ordered trees.

Algorithm analysis

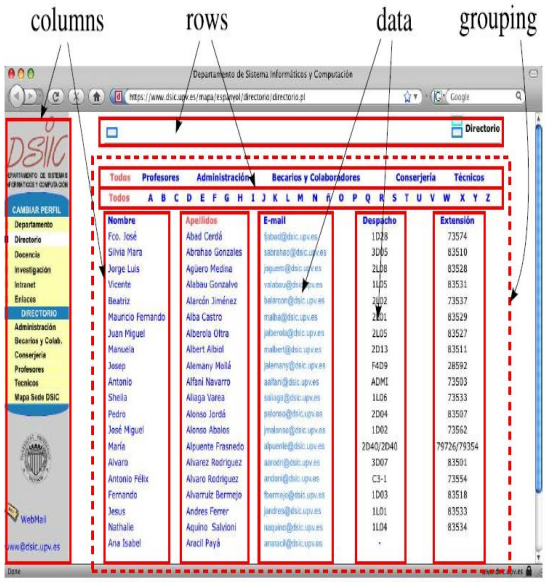

Figure 5.1 Analysis for column, row and text

A description of the few existing techniques and a comparison with our approach can be founded. We developed a preliminary technique for Webpage comparison that takes into account their visual structure. The key idea behind the method is to transform each Web page into a compressed, normalized tree that effectively represents its visual structure. The transformation is based on a classification of the set of html–tags that is guided by the visual effect of each tag in the entire structure of the page. Then, a metric to compute the distance between two Web pages that is based on the classical tree edit distance algorithm was formalized. This algorithm defines the dissimilarity of two trees by determining the minimal cost of the edit operations that are needed to transform one tree into the other.

A preliminary implementation that is written in Maude was also presented. Unfortunately, the vast(fix point) computation that is involved in the Web comparison algorithm of leads to unsatisfactory performance, so the prototypical tool can only efficiently process very small examples. In order to reduce the time and space complexity of the tree edit distance algorithm, different optimizations have been developed. These are based on computing only a subset of a large dynamic programming table by using memorization. In this paper, we present a powerful optimization of our previous Web comparison algorithm that is based on similar memorization techniques and allows us to compare documents that are extracted from real document

collections and databases. We have also implemented the former Web comparison tool in Maude in order to improve both the functionality and the performance of the previous version. Finally, we report two experiments in realistic scenarios.

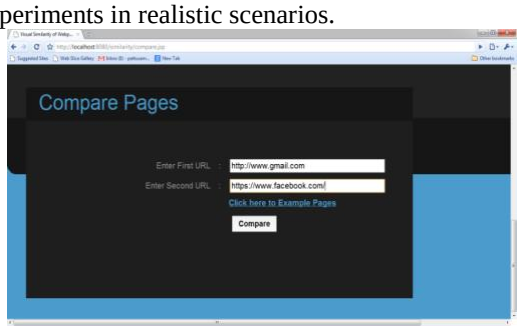

Figure 5.2 User Comparing URL

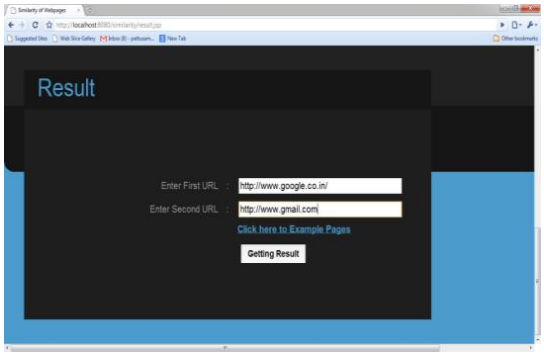

Figure 5.2 Result page

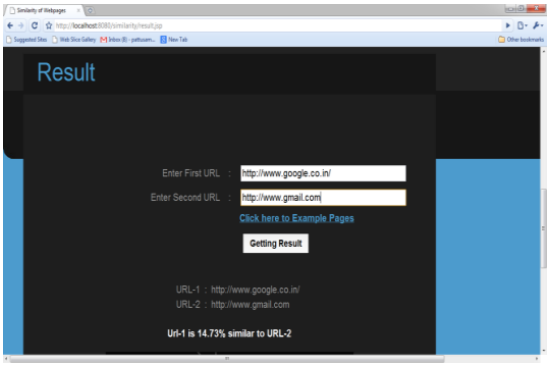

Figure 5.3 User getting similarity

V. CONCLUSION AND FUTURE ENHANCEMENT

Web page comparison is currently an open problem who's Importance extends from search engines to Web categorization. We define the "visual structure" of Web pages the structure perceived by a human, and then develop translation of Web pages to a canonical representative that effectively represents the visual structure of the page. A functional method for computing the similarity between two Web pages is also proposed which is based on generalization of the tree edit distance algorithm. We have implemented in Maude our technique for recognizing and comparing the visual structural information of Web pages by using memorization. Our experiments demonstrate considerable improvements in both time and space over the former version of the tool, which makes the technique applicable in real scenarios. As future work, we plan to extend our analysis by also considering style sheets and the visual effect of including Images on the web pages. This extension can be easily achieved by dividing a Web page into a collection of HTML code blocks that are determined by the CSS attributes of the Web page and the dimensions of the images. Then the tree structure of the Web page can be easily assembled by considering the coordinates given by the property "position "of its elements. This simple extension allows us to use our visual comparison methodology for collections of Web pages that include style sheets.

REFERENCES

[1] P. Lakkaraju, S. Gauch, and M. Speretta, "Document similarity based on concept tree distance," in Proc. of the 19th ACM Conf. on Hypertext and hypermedia. New York, NY, USA: ACM, 2008, pp. 127–132.

[2] A. Paepcke, H. Garcia-Molina, G. Rodriguez Mula, and J. Cho, "Beyond document similarity: understanding valuebased search and browsing technologies," SIGMOD Rec., vol. 29, no. 1, pp. 80–92, 2000.

[3] W. W. Cohen, "Recognizing structure in Web pages using similarity queries," in Proc. of the 16th Nat. Conf. on Artificial Intelligence and the 11th Innovative App. of Artificial Intelligence, Menlo Park, CA, USA, 1999, pp. 59–66.

[4] A. Y. Fu, L. Wenyin, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (emd)," IEEE Trans. Dependable Secur.

[5] J. Cao, B. Mao, and J. Luo, \A segmentation method for web page analysis using shrinking and dividing," JPEDS, vol. 25, 2010.

[6] A.Y. Fu, L. Wenyin, and X. Deng, \Detecting phishing web pages with visual similarity assessment based on earth mover's distance (emd)," TDSC, vol. 3, 2006.

[7] N. Thome, D. Merad, and S. Miguet, \Learning articulated appearance models for tracking humans: A spectral graph matching approach," Signal Processing: Image Communication, vol. 23, no. 10, 2008.

[8] S. Avila, N. Thome, M. Cord, E. Valle, and A. Araujo, \Bossa: Extended bow formalism for image classification," in ICIP 2011.

[9] K. Zhang and D. Shasha, "Simple fast algorithms for the editing distance betweentrees and related problems" SIAM Journal of Computing, Vol 18-6, (1989), p. 1245-1262.

[10] K. Zhang and D. Shasha, "Simple fast algorithms for the editing distance between trees and related problems" SIAM Journal of Computing, Vol 18-6, (1989), p. 1245-1262.