

Chapter 79

HTML Segmentation for Different Types of Web Pages

Evelin Carvalho Freire de Amorim

Departamento de Ciência da Computação (UFMG), Brazil

ABSTRACT

Search engines manage several types of challenges daily. One of those challenges is locating relevant content in a Web page. However, the concept of relevance in information retrieval depends on the problem to be solved. For instance, the menu of a website does not impact the results of an algorithm to detect duplicate Web pages. An HTML segmentation algorithm partitions a Web page visually in such a way that parts from a same partition are semantically related. This chapter presents two strategies to segment different types of Web pages.

INTRODUCTION

Search engines manage redundant and non-structured content daily. However, redundant and non-structured data generate problems that affect the performance of search engines. For example, redundant data are not useful for a query; nevertheless redundant data can be exhibit in results if they are not removed from the dataset. Partitioning a web page into cohesive visual pieces and selecting the most relevant piece can improve algorithms for detection of redundant data. The task of partitioning a web page into cohesive visual pieces is called HTML segmentation.

Web browsing in mobile devices is also enhanced by HTML segmentation (Yin & Lee, 2004). The web browser of a mobile device can partition a web page and exhibit the most relevant part of the web page in the center of the screen. This feature improves the user's experience in the mobile device.

Another task to be solved by HTML segmentation is the ranking quality of standard web pages searching schemes (Fernandes, Moura, da Silva, Ribeiro-Neto, & Braga, 2011). Ranking of web pages is an important task in Information Retrieval and search engines are concerned about the best ranking of web pages.

DOI: 10.4018/978-1-5225-1759-7.ch079

HTML Segmentation for Different Types of Web Pages

There are two main types of HTML segmentation techniques: general or topical. The latter technique segments only specific types of web pages, for instance blogs or news. Although topical techniques achieve robust results they are inflexible for particular Information Retrieval tasks. General techniques face the challenge of finding a model that conciliates features from different web pages like personal web pages and e-commerce web pages.

Considering that general techniques for HTML segmentation are uncommon and still constitute a challenge for the data mining area, because web pages displays relevant content in different ways. For instance, describing news web pages and an e-commerce web page in one model is not an intuitive task.

This chapter has the following goals:

1. Describing general techniques for HTML segmentation;
2. Comparing two general HTML segmentation techniques. The first strategy is called ETL HTML segmentation and the second strategy is called TPS segmentation.

The remaining of this chapter also reviews some topical techniques, the main results of HTML segmentation algorithms and issues to solve in HTML segmentation.

BACKGROUND

HTML segmentation covers concepts from information retrieval and data structures. The following subsection defines data structures concepts employed in HTML segmentation algorithms. The next subsection describes how HTML segmentation improves some tasks of the information retrieval area.

Data Structures Concepts

A rendering web page is purely an HTML document in its visual form. However, a more suitable representation of HTML is required in order to automatically extract information from web pages. Therefore algorithms that process web pages use a data structure called DOM (Document Object Model) Tree, which defines a logical structure of documents and the way a document is accessed and manipulated (Le Hégaret, P. Wood, L., & Robie, J., 2000). The process of building a DOM Tree transforms each HTML tag into a DOM node, which also involves assigning attributes of a tag to the corresponding DOM node. Figure 1 shows an HTML code on the left side and its corresponding DOM Tree on the right side.

Web browsing requires a visual representation of HTML though. Besides that, a rendering web page allows the user to locate continuous visual parts of his or hers concern, for instance, menu, product pictures, and many others parts. The concept of HTML segment is based on these visual parts of a rendering web page. For instance, an e-commerce web page usually shows the following visual parts: describing product, pictures of product, reviews of products, and so on.

Chakrabarti, Kumar, and Punera (2008) formally defined HTML segment as a visual continuous and cohesive piece of a web page. HTML segmentation tasks aims to find a set of web page segments in the same way a human divides a web pages in different semantic parts. By using this kind of segmentation, it is possible to improve the ranking quality of standard web pages searching schemes (Fernandes et al., 2011) (Song et al., 2004). Also, the duplicate detection of web pages is enhanced by HTML segmentation algorithms (Chakrabarti et al., 2008).

22 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/html-segmentation-for-different-types-of-web-pages/173409

Related Content

Detection of Cardiovascular Disease Using Ensemble Feature Engineering With Decision Tree

Debasmita GhoshRoy, P. A. Alviand João Manuel R. S. Tavares (2022). *International Journal of Ambient Computing and Intelligence* (pp. 1-16).

www.irma-international.org/article/detection-of-cardiovascular-disease-using-ensemble-feature-engineering-with-decision-tree/300795

Intelligent Industrial Process Control Based on Fuzzy Logic and Machine Learning

Hanane Zermaneand Rached Kasmi (2020). *International Journal of Fuzzy System Applications* (pp. 92-111).

www.irma-international.org/article/intelligent-industrial-process-control-based-on-fuzzy-logic-and-machine-learning/245272

Artificial Intelligence in Education as a Means to Personalize Learning

Ana María Lacárcel (2022). *Handbook of Research on Artificial Intelligence in Government Practices and Processes* (pp. 285-308).

www.irma-international.org/chapter/artificial-intelligence-in-education-as-a-means-to-personalize-learning/298911

Self-Service Technology Banking Preferences: Comparing Libyans' Behaviour in Developing and Developed Countries

Fouad Omran Elgahwashand Mark Bruce Freeman (2013). *International Journal of Intelligent Information Technologies* (pp. 7-20).

www.irma-international.org/article/self-service-technology-banking-preferences/77871

Ambient Media Culture: What Needs to be Discussed When Defining Ambient Media from a Media Cultural Viewpoint?

Artur Lugmayr (2012). *International Journal of Ambient Computing and Intelligence* (pp. 58-64).

www.irma-international.org/article/ambient-media-culture/74370