Statistical Web Object Extraction

Jun Zhu
Tsinghua University, China

Zaiqing Nie
Web Search and Mining Group Microsoft Research Asia, China

Bo Zhang
Tsinghua University, China

INTRODUCTION

The World Wide Web is a vast and rapidly growing repository of information. There are various kinds of objects, such as products, people, conferences, and so on, embedded in both statically and dynamically generated Web pages. Extracting the information about real-world objects is a key technique for Web mining systems. For example, the object-level search engines, such as *Libra* (http://libra.msra.cn) and *Rexa* (http://rexa.info), which help researchers find academic information like papers, conferences and researcher's personal information, completely rely on structured Web object information.

However, how to extract the object information from diverse Web pages is a challenging problem. Traditional methods are mainly template-dependent and thus not scalable to the huge number of Web pages. Furthermore, many methods are based on heuristic rules. So they are not robust enough. Recent developments in statistical machine learning make it possible to develop advanced statistical Web object extraction models. One key difference of Web object extraction from traditional information extraction from natural language text documents is that Web pages have plenty of structure information, such as two-dimensional spatial layouts and hierarchical vision tree representation. Statistical Web object extraction models can effectively leverage this information with properly designed statistical models.

Another challenge of Web object extraction is that many text contents on Web pages are not regular natural language sentences. They have some structures but are lack of natural language grammars. Thus, existing natural language processing (NLP) techniques are not directly applicable. Fortunately, statistical Web object extraction models can easily merge with statistical NLP methods which have been the theme in the field of natural language processing during the last decades. Thus, the structure information on Web pages can be leveraged to help process text contents, and traditional NLP methods can be used to extract more features.

Finally, the Web object extraction from diverse and large-scale Web pages provides a valuable and challenging problem for machine learning researchers. To nicely solve the problem, new learning methodology and new models (Zhu et al., 2007b) have to be developed.

BACKGROUND

Web object extraction is a task of identifying interested object information from Web pages. A lot of methods have been proposed in the literature. The wrapper learning approaches like (Muslea et al., 2001; Kushmerick, 2000) take in some manually labeled Web pages and learn some extraction rules (wrappers). Since the learned wrappers can only be used to extract data from similar pages, maintaining the wrappers as Web sites change will require substantial efforts. Furthermore, in wrapper learning a user must provide explicit information about each template. So it will be expensive to train a system that extracts data from many Web sites. The methods (Zhao et al., 2005; Embley et al., 1999; Buttler et al., 2001; Chang et al., 2001; Crescenzi et al., 2001; Arasu, & Garcia-Molina, 2003) do not need labeled training samples and they automatically produce wrappers from a collection of similar Web pages.

Two general extraction methods are proposed in (Zhai & Liu, 2005; Lerman et al., 2004) and they do not explicitly rely on the templates of Web sites. The method in (Lerman et al., 2004) segments data on list

pages using the information contained in their detail pages, and the method in (Zhai & Liu, 2005) mines data records by string matching and also incorporates some visual features to achieve better performance. However, the data extracted by (Zhai & Liu, 2005; Lerman et al., 2004) have no semantic labels.

One method that treats Web data extraction as a classification problem is proposed in (Fin & Kushmerick, 2004). Specifically, Fin & Kushmerick (Fin & Kushmerick 2004) use a support vector machine to identify the start and end tags for a single attribute. For the task of extracting multiple attributes, this method loses the dependencies between different attributes. Instead, the statistical models, which are the theme of this article, can effectively incorporate the statistical dependencies among multiple related attributes, such as a product's name, image, price and description, and achieve globally consistent extraction results.

MAIN FOCUS

Statistical Web object extraction models focus on exploring structure information to help identify interested object information from Web pages. Zhu et al. (Zhu et al, 2005; Zhu et al, 2006; Zhu et al, 2007a; Zhu et al, 2007b) have developed a complete statistical framework for Web object extraction and text content processing on Web pages. The key issues in statistical Web object extraction are selecting appropriate data representation formats, building good graphical models to capture the statistical dependencies, and exploring the structure information of Web pages to process text contents.

Data Representation Formats

Most existing Web mining methods take the HTML source codes or the HTML tag trees as their data representation format. However, due to the low-level representation these methods often suffer from many problems, such as scalability and robustness, when being applied to large-scale diverse Web pages. Instead, the higher level visual information, such as font, position, and size, is more robust and expressive, which will be the features and data representation formats used in statistical models.

Existing statistical models are built based on two different views and data representation formats of Web pages.

1. **2D spatial layout:** When a Web page is displayed to readers, it is actually a two-dimensional image. The HTML elements have their spatial information, such as position (i.e., coordinates in the 2D plane) and size (i.e., height, width, and area). With this spatial information, the elements are well-laid in the 2D plane for easy reading. Thus, the first representation format of Web data is a two-dimensional grid, each of whose nodes represents an HTML element. The edges between the nodes represent the spatial neighborhoods of the HTML elements.
2. **Hierarchical organization:** 2D spatial layout is a flat representation of a Web page. But if we look at the HTML tag trees, which are natural representations of Web pages, we see that the HTML elements are actually pended on a hierarchy. The hierarchical structure in some sense reveals a specific type of organization of the elements. This hierarchical organization information can be helpful in identifying the boundaries of data records or even in identifying the target attributes.

Statistical Web Object Extraction Models

According to the two different data representation formats—2D spatial layout and hierarchical organization, two types of statistical models have been studied.

1. **2D local model:** The two-dimensional Conditional Random Fields (Zhu et al, 2005) are introduced to model a data record, which consists of a set of HTML elements. The 2D model put the elements on a grid according to their spatial information and neighborhood relationships. Each node on the grid is associated with a random variable, which takes values from a set of class labels, such as name, image, price, and description in product information extraction. The model defines a joint distribution of all the variables and we can do some statistical inference to get the most probably labeling results of all the random variables. From the labeling results, we know which element is