Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Extraction of Citation Data from Websites based on Visual Cues

Masterarbeit

zur Erlangung des akademischen Grades
Master of Science (M. Sc.)

eingereicht von:	Tim Repke
geboren am:	31. Dezember 1989
geboren in:	Berlin
Gutachter/innen:	Prof. Ulf Leser
	Prof. Niels Pinkwart
eingereicht am:	15. September 2016

Abstract

In this master's thesis a system for extracting meta-information, specifically citation data, from webpages is proposed. Machine Learning models like Artificial Neural Networks and Random Forests are trained to classify elements on a given webpage based on visual cues. Visual properties of elements are analysed in detail in order to derive meaningful numerical features for classification. After applying sensible post-processing filters, the system is able to recall up to 80% of the desired data at a precision of up to 90%. Relying purely on visual cues however has it's limitations for robust extraction of some of the citation data. Possible approaches to facilitate that are discussed at the end.

Contents

1. Introduction

In this master's thesis, a system capable of extracting meta-information, specifically *citation data*, from arbitrary webpages is developed. Accurately identifying the title, author, and date of an article in the web is a very challenging task due to the large diversity of designs and underlying source code. Since humans can easily locate these information, the system uses visual cues to describe elements on a webpage to be used for classification by Machine Learning models. In order to define meaningful numerical features, properties of elements from hundreds of webpages are analysed. They are used to train Artificial Neural Networks and Random Forests to distinguish citation data from other content on a given webpage. After applying sensible post-processing filters, the proposed system reaches a mean F-score of 75% for extracting desired information from previously unseen webpages. This thesis also shows limitations of purely relying on visual cues.

1.1. Motivation

Every scientific article, be it a student's essay or a paper in a journal, or even an entry on the Wikipedia, provides citations and a bibliography to support its arguments. Publishers and universities require their academics to format bibliographies according to specific style guidelines. Correctly organising and formatting those can be a tedious task. Fortunately though, there are software tools, so-called *Citation Generators*, to help with that. To simplify that process even further, those tools sometimes even acquire relevant citation data automatically, when given an identifier like the ISBN of a book, DOI of a journal article, or URL to a webpage. This thesis focuses on latter case.

The information needed to cite something, like the title, author, date of publication and more, is called *citation data*. Writing software that extracts these meta information from webpages using the same underlying template is trivial, however a Citation Generator has to deal with arbitrary webpages with lots of different layouts. The diversity of visual appearance and underlying markup structure is

overwhelming, so writing specific wrappers for each template is virtually impossible. General approaches that search for citation data within a webpage's markup are limited by syntactical identifiers, which are defined by developers and don't follow standards nor reflect the appearance. Content is intended to be rendered by a browser where humans can easily recognise relevant information, which suggests that using visual cues from the rendered representation provides valuable properties to describe and extract citation data elements from a webpage. Manually developing conditions based on those properties would quickly result in unmaintainable code, whereas Machine Learning algorithms can learn from training samples.

1.2. Description of Citation Data in Webpages

Figure 1.1.: Example webpage and its corresponding formatted citation.

In this section the notion of the term *citation data* will be defined, as it is critical to understand what exactly is meant by that in order to follow the concepts presented in this thesis. Citation data are meta information about a resource, which are required when referencing it in a bibliographic entry. As mentioned earlier, only the title, author and date are considered in this work and are emphasised in the example shown in Figure 1.1.

Online content commonly displays a title, author and date. For a proper citation nuances for each of those fields need to be differentiated. The webpage of an article for example contains multiple candidates that could be considered a title: The website's name, the descriptor of a section, the articles title and subsection headers within the article. This thesis only targets article titles. It is also important to note, that the content of the title tag of a webpage does not necessarily represent

the actual title for that page [54, 57]. Similarly, only the actual directly visible (human) authors of an article are looked at, not publishers or editors. There might also be different dates associated with the content, namely the date of submission, publication or later updates. In this context, the latest available date is used and from now on only referred to as "date" and should at least contain the year.

In the scope of this work only directly visible data is used to extract citation data. For example entries in an online encyclopaedia usually don't display an author or date but it would be possible to find this information in a separate version history. Also, on personal websites it is implicitly clear who authored individual blog posts and thus its often not displayed next to each post.

1.3. Structure

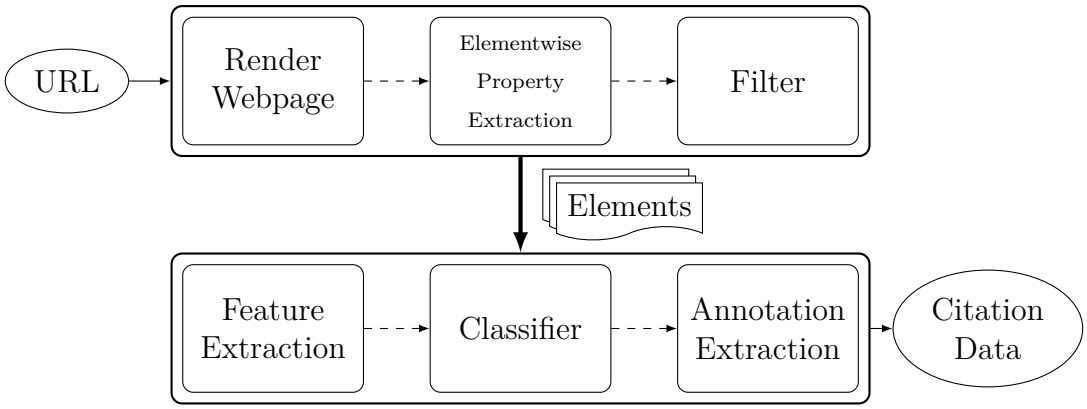

Figure 1.2.: Schematic of the proposed system

The structure of this work is based on the schematic representation of the proposed system as shown in Figure 1.2. Given the Unified Resource Locator (URL) to a webpage, its citation data will be returned. For reference, a set of 870 webpages with annotated citation data was created as described in the first half of Chapter 4.

The data processing from a set of URLs to the extracted citation data is comprised of two main components as indicated by the thick boxes in the mentioned graphic. The second half of Chapter 4 provides details on how the webpages are rendered, so that each elements' calculated styling and positional properties can be extracted from the Document Object Model (DOM) (Section 4.2). The Gold Standard for the extraction task is the set of elements labelled according to the acquired annotations.

After filtering outliers, the set is passed on to the second component. There, numerical features are constructed from previously extracted properties. An analysis, along with a comprehensive description of the feature extraction process, can be found in Chapter 5.

Each elements' features are used to classify it into one of four classes: `Title`, `Author`, `Issued` or `Unassigned`, whereas the latter indicates a node, which carries no citation data. The class assigned to an element is called *label*. Note that the classifier returns element-wise probabilities for all labels. Theoretical background for the machine learning algorithms used in this thesis is provided in Chapter 3, namely Artificial Neural Networks and Random Forests.

Chapter 6 describes in more detail how these classifiers are set up, and how their output is used to return the most likely annotations to the URLs in the gold standard. The citation data extraction is later evaluated on set of previously unseen examples in Chapter 7 followed by a discussion.

All analyses, experiments and evaluations were done on a Dell R920 shared compute server ("gruenau6") with four 4 Intel Xeon E7-4880 v2 (15 cores with 2.5GHz each) and 1024 GigaByte RAM running the SuSE 13.1 linux operation system, provided by the computer science department of the Humboldt-University Berlin.

Although URLs for the selection of webpages in the gold standard are provided by the online citation manager RefME[1], that also kindly supported the crowdsourced annotation, this work in no way biased or specifically developed for their business interest. Any additional data is acquired using publicly accessible data endpoints.

[1] `https://refme.com`