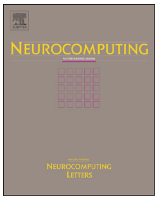
Neurocomputing 135 (2014) 61–68

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



A class of neural-network-based transducers for web information extraction

Hassan A. Sleiman*, Rafael Corchuelo

University of Sevilla, ETSI Informática, 41012 Sevilla, Spain.

ARTICLE INFO

Article history:
Received 31 October 2012
Received in revised form
8 March 2013
Accepted 30 May 2013
Available online 18 January 2014

Keywords:
web wrappers
web information extraction
neural networks
finite automata
machine learning
supervised method

ABSTRACT

The Web is a huge and still growing information repository that has attracted the attention of many companies. Many such companies rely on information extractors to integrate information that is buried into semi-structured web documents into automatic business processes. Many information extractors build on extraction rules, which can be handcrafted or learned using supervised or unsupervised techniques. The literature provides a variety of techniques to learn information extraction rules that build on ad hoc machine learning techniques. In this paper, we propose a hybrid approach that explores the use of standard machine-learning techniques to extract web information. We have specifically explored using neural networks; our results show that our proposal outperforms three state-of-the-art techniques in the literature, which opens up quite a new approach to information extraction.

1. Introduction

The growth of the Web has raised many companies′ interest in using the information that it provides to feed their business processes. Unfortunately, extracting information from web documents is not easy at all, which has motivated many researchers to work on the so-called information extractors. The common theme is to help transform web documents into structured information; that is, information for which there is an explicit model, so that it can be easily consumed by automatic business processes.

The information in the Web ranges from free-text to semi-structured information. Free-text web documents provide information that is buried into text that is written in natural language, with a few HTML tags that endow it with a little structure, e.g., headings, sections, or sidebars. Unfortunately, this little structure is not enough to characterise the information to be extracted; thus, natural language processing techniques are required to extract relevant information from these documents [22,31]. On the contrary, the information in a semi-structured web document is formatted using regular patterns, e.g., grids and lists. In these documents, HTML tags provide far more structure than in free-text documents since the pieces of information to extract are usually enclosed within formatting tags. Such pieces of information are usually referred to as slots in this context [4,27].

Our work focusses on information extraction from semi-structured web documents. The proposals in this field can be classified into two categories, namely: rule-based and heuristic-based. Rule-based information extractors rely on the so-called extraction rules, which range from regular expressions to context-free grammars, horn clauses, tree templates, or transducers, to mention a few. Heuristic-based proposals do not rely on extraction rules, but are based on a number of heuristics that have proven to work well on a large number of web sources [1,8,9,24,28]. Extraction rules can be handcrafted [6,21], but the costs involved motivated many researchers to work on proposals to learn them automatically using supervised and unsupervised learning techniques. Supervised learning techniques require the user to provide samples of the information to be extracted, aka annotations [10,14,19,29,30]. Unsupervised learning techniques learn extraction rules starting from one or more web documents; these rules extract as much prospective information as they can and the user then gathers the relevant information from the results [2,3,7,11,17,26]. Since typical web documents are growing in complexity, some authors are also working on techniques whose goal is to identify the region within a web document where the relevant information is most likely to reside [27]. Other authors have studied the problem of verifying the information extracted [12,13,15,18]; their focus was on determining when an information extractor stops performing well due to changes in the structure of the documents from which they have to extract information. Other authors have focussed on how to maintain information extractors automatically, that is, on how to adapt them to changes with minimum user intervention [16,20].

* Corresponding author.
E-mail addresses: hassansleiman@us.es (H.A. Sleiman), corchu@us.es (R. Corchuelo).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.05.057

Since automatic techniques build on machine learning techniques or heuristics [27], their precision and recall may not generally expected to be 100% in every case. This makes this quite an active research area. A user might also handcraft ad hoc information extractors that might have perfect precision and recall for a given web site, but the effort this task requires the high precision and recall that automatic techniques may achieve, render handcrafting rules not very appealing in general. We have found out that, without an exception, the automatic proposals in the literature build on ad hoc machine-learning techniques. As far as we know, the idea of using standard machine-learning techniques in this field remains largely unexplored. In many conversations with other researches in conferences world wide, we have found out that the usual problem that motivates researchers to work on ad hoc machine-learning techniques is that it is not easy to map the problem of information extraction onto the dataset tables that are required by standard machine-learning techniques. This has motivated us to work on this topic: we wished to explore it in an attempt to find out if ad hoc machine-learning techniques can be outperformed by means of standard machine-learning techniques.

Our proposal builds on transducers, which are regular automata that produce an output as they move from state to state; the outputs of our transducers are expected to be information that is extracted from input web documents or fragments. Transducers have proven to be very useful for information extraction in the past [5,10] since they have the ability to deal with web documents in which there is optional information, information that is displayed with different orderings, multi-valued information, and multiple formats for the same information. What is innovative in our proposal is that we have explored the idea of using neural networks to control how transitions take place. This resulted in a hybrid approach that can easily map the problem of web information extraction onto the dataset tables that are common to standard machine-learning techniques. The decision on why to focus on neural networks was because they are quite a powerful machine-learning technique that is characterised by its ability to find patterns in complex datasets.

The rest of the paper is organised as follows: Section 2 reports on the details of our proposal; Section 3 reports on our experimental evaluation, which proves that our proposal outperforms other state-of-the art techniques that build on ad hoc machine-learning techniques; Section 4 concludes our work and highlights a few future research directions.

2. Description of our proposal

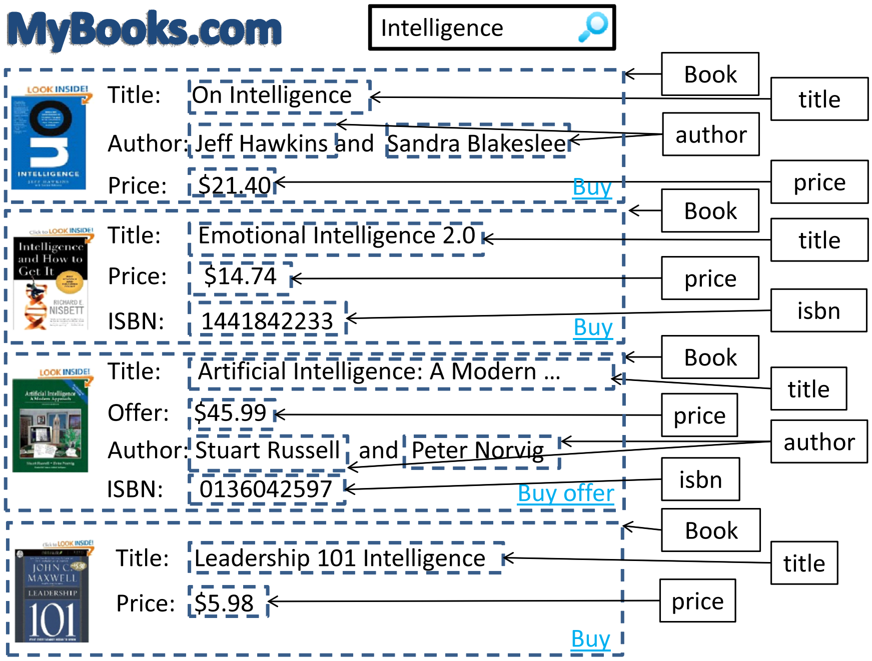
Our proposal builds on using transducers, so we first provide a formal definition of this concept:

Definition 1. A transducer is a tuple of the form (S, i, f, T), where S is a set of states, i denotes a state in S that is referred to as the initial state, f denotes a state in S that is referred to as the final state, and T is a set of transitions of the form (p, n, q), where p and q denote states in S and n is a neural network that encodes the conditions required to move from state p to state q.

In the following subsections, we delve into the details regarding how to learn a transducer from a training set and how to run it on an input web document. The details are illustrated using the sample web document in Fig. 1. Note that this sample web document is intentionally simple and that it is fully annotated since it is intended for illustration purposes only. Our experience proves that it generally suffices to provide a few annotations of every possible formatting, like in other state-of-the-art supervised proposals.

2.1. *Learning a transducer*

Our proposal relies on learning a transducer from a training set that consists of annotated web documents. By annotation we refer to a slot to which the user has assigned an explicit label; by slot we refer to either a record or an attribute inside a record. We do not preclude the possibility that an attribute can also be a record since a transducer can be applied to a whole web document or to a fragment in that document that was extracted previously by means of another transducer. For instance, the sample web document in Fig. 1 has four record annotations (`Book`) and several attribute annotations (`title`, `author`, `price`, and `isbn`).



Title: On Intelligence
Author: Jeff Hawkins and Sandra Blakeslee
Price: $21.40
Buy

Title: Emotional Intelligence 2.0
Price: $14.74
ISBN: 1441842233
Buy

Title: Artificial Intelligence: A Modern ...
Offer: $45.99
Author: Stuart Russell and Peter Norvig
ISBN: 0136042597
Buy offer

Title: Leadership 101 Intelligence
Price: $5.98
Buy

Fig. 1. A sample web document with annotations.