You Need Only One Clue for Effective Record Segmentation[*]

Cheng Wang, Tim Furche, Georg Gottlob, Giovanni Grasso, Giorgio Orsi, Christian Schallhart
Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD
firstname.lastname@comlab.ox.ac.uk

ABSTRACT

Record segmentation is a core problem in data extraction. Previous approaches have focused on more and more sophisticated heuristics without knowledge of the concrete domain. In this work, we demonstrate that with only a single clue about mandatory attributes in a given domain, straightforward rules for record segmentation suffice to achieve 100% precise record extraction from the vast majority of web sites in that domain. These results are first outcomes of the just launched ERC project DIADEM on domain-specific intelligent automated data extraction.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: On-line Information Services—*Web-based services*

General Terms

Languages, Experimentation

Keywords

record segmentation, data extraction, deep web

1. INTRODUCTION

You want to rent a newly refurbished, two bedroom flat in Oxford which welcomes your pet and it is located close to your office and your children's schools. All this information is readily available on some webpage, but manually extracting, aggregating, and ranking that data is tedious and often unmanageable due to the size of the relevant data. Automatic data extraction is the last chance to make all such data accessible to further processing. With such automation, users remain overwhelmed by the huge amount of information the web can deliver.

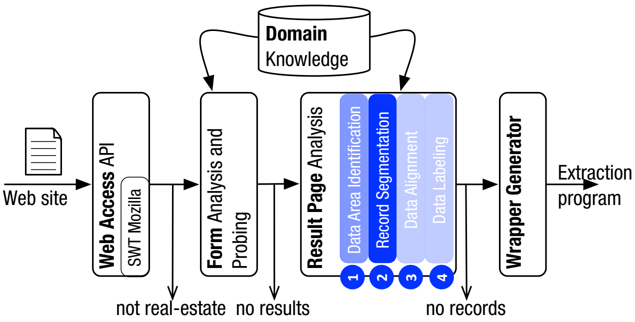
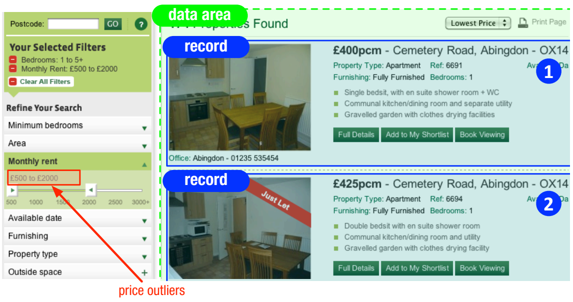


Figure 1: Record segmentation in DIADEM

The architecture of web data extraction systems follows the pipeline shown in Figure 1. We focus on result page analysis, i.e., the extraction of records such as property listings from result pages provided through search forms. Result page analysis can be divided into four phases: **(1)** *Data area identification:* what is the area within a page containing all relevant data; **(2)** *Record segmentation:* how to divide the data area into records; **(3)** *Data alignment:* which data items in the records belong to the same attribute; **(4)** *Data labeling:* what are the labels for these aligned attributes.

Previous unsupervised web extraction approaches mostly detect repeated patterns within a page employing a sophisticated combination of heuristics and similarity functions. Most of these tools are domain-independent, and at best use domain knowledge (e.g., ontologies) to refine their results. In [1], ontologies are used as one among six heuristics for record segmentation. However, the assumptions on the structure of HTML pages no longer hold for modern web pages. ODE [2] uses ontologies for data area identification, alignment, and labeling, but not for record segmentation.

In this paper, we show a novel technique for record segmentation that uses mandatory, domain-specific information


Your Selected Filters
Bedrooms: 1 to 5+
Monthly Rent: £500 to £2000

Monthly rent: £500 to £2000

£400pcm - Cemetery Road, Abingdon - OX14
Property Type: Apartment Ref: 6691
Furnishing: Fully Furnished Bedrooms: 1
- Single bedsit, with en suite shower room + WC
- Communal kitchen/dining room and separate utility
- Gravelled garden with clothes drying facilities

Office: Abingdon - 01235 535454

£425pcm - Cemetery Road, Abingdon - OX14
Property Type: Apartment Ref: 6694
Furnishing: Fully Furnished Bedrooms: 1
- Double bedsit with en suite shower room
- Communal kitchen/dining room and utility
- Gravelled garden with clothes drying facility

Figure 2: Price outlier result page (`finders.co.uk`)

[*] The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement no. 246858.

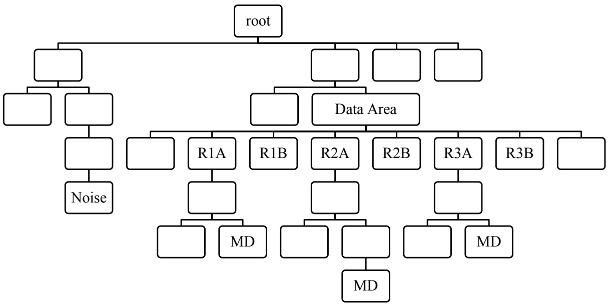


Figure 3: Result page pattern structure

(such that each property has a price) to segment records. We demonstrate that a single such clue suffices to achieve 100% precision for record segmentation on 96,7% of pages in a representative corpus of 60 real estate result pages.

Figure 2 shows a heavily scripted result page for the UK real estate agency Finders Keepers. Using only the fact that each property record has a price, our approach identifies all records on such a page as well as the data area. It is not fooled by prices that occur outside of the data area, as it exploits the structural properties of the records as well.

2. ALGORITHM DESCRIPTION

Our segmentation algorithm assumes that **(1)** each page has a single data area which **(2)** contains all records.

Data Area Identification. A data area in a result page is identified by leveraging *mandatory elements.* A mandatory element is a domain concept that appears in all the records of a given data area, e.g., the location in a real-estate website. Since in this phase the records are yet to be discovered, the mandatory elements are identified by matching the content of the text nodes with a domain ontology. The matching nodes are then labeled as MD-nodes as shown in Figure 3.

Since the matching process is intrinsically imperfect, it is possible to introduce false-positives during the identification of mandatory nodes. To reduce such false-positives, we group the MD-nodes with same (or similar) *depth* in the DOM and similar *relative position* among their siblings. We then consider only the MD-nodes belonging to the largest group as nodes of the data-area and we discard the other nodes. The least common ancestor in the DOM tree of all the identified MD-nodes is considered the data area root.

Record Segmentation. The records within a data area are identified as sub-trees rooted at children of the data area root. The segmentation process uses *record separators*, i.e., sub-trees interleaved with data records. In general, the root node of a record separator is a child of the data area root and does not contain any text or URL.

Since each record contains only one instance of an MD-node, the segmentation process heuristically identifies the areas belonging to a single record. As a first step, each subtree in the DOM containing a single MD-node and rooted at a direct child of the data area root is considered a *candidate record.* A subsequent step tries to "expand" the candidate record to adjacent subtrees in the DOM. We therefore consider the siblings of the candidate record (i.e., other direct-children of the data area root). If they are record separators, we consider each candidate record as a proper record; otherwise we apply the following steps: **(1)** Compute the *distance* l between two candidate records as the number of siblings between their root nodes. **(2)** Consider all the $2 \times (l - 1)$ expansions of a record to left and right adjacent subtrees. We apply the same expansion to all candidate records and compute the similarity between the identified expansions. The one with the highest similarity is considered optimal. **(3)** Whenever several expansions have the same similarity, we choose the one with the highest structural similarity among records. To break ties, we pick the one delimited by the largest number of record separators.

Experiments. We evaluate our prototype on 60 representative UK real estate websites against manually created ground truth annotations. We use a single mandatory attribute, the price of a property. With only this clue, we achieve we achieve 100% precision on 58 (96,7%) of the web sites both for identifying the data area and for record segmentation. The remaining 2 contain no price information, but using, e.g., the fact that also postcode is a mandatory attribute, we could also identify the records on these pages.

3. FUTURE WORK

Our approach improves with the number of relevant clues for prices and *other attributes,* such as postcodes or size. We can use this information to pursue two goals: the *tight integration* of the algorithm into the overall DIADEM pipeline and the *refinement* of the obtained results.

DIADEM Integration. We widen the scope of our approach to cover *all steps of the result page analysis,* from data area identification to data labeling. We have implemented the first two steps, and already obtain hints for data alignment and labeling, as we label mandatory fields whose instances ought to be aligned together. Once clues for further attributes are available, we immediately obtain a partial alignment which we expand structurally into a full record alignment. To this end, we also consider information on the query which generated the result pages at hand.

Algorithmic Refinement. **(1)** We improve the quality of the algorithm's results by a more *refined outlier detection and elimination.* We already identify outliers occurring outside the data area (see Figure 2), but those occurring in the records may remain undetected. For example, some real estate records contain additional prices, e.g., for an optional parking lot. In some cases our current heuristics are unable to distinguish these prices, simple structural properties to discriminate these instances. **(2)** We run our core algorithm with clues for further, non-mandatory attributes and *choose among the generated segmentations the best-fitting one*—measuring e.g. the size and location of the data area, the number of the records, or their average size. **(3)** Alternatively, we develop an integrated approach, where we use *different attributes simultaneously* to determine a common data area, and where we segment the records in order to distribute the attributes evenly on the resulting records.

4. REFERENCES

[1] D. W. Embley, D. M. Campbell, Y. S. Jiang, S. W. Liddle, Y.-K. Ng, D. Quass, and R. D. Smith. Conceptual-model-based data extraction from multiple-record web pages. *DKE*, 1999.

[2] W. Su, J. Wang, and F. H. Lochovsky. ODE: Ontology-Assisted Data Extraction. *TODS*, 2009.