Technische Universität Wien
Vienna University of Technology

GraphWrap: Interactive data extraction from PDF documents using graph matching techniques

Database and Artificial Intelligence Group

Have you ever needed to extract large quantities of data such as product specifications, measurements, contact details or prices from one or more PDF files? Then let us introduce you to *GraphWrap* – a new technique for user-guided data extraction, or “wrapping”, which uses graph matching techniques to locate data instances.

There are already a number of existing systems which offer similar wrapping solutions for web pages. These systems make use of the hierarchical structure inherent in the HTML format, which explicitly defines the individual data elements and how they are grouped together. In PDF, there is no such explicit structure, and data extraction from PDF is therefore a far more challenging task.

By developing a graph-based representation of a PDF and accompanying graph matching methods, we have achieved a solution to this problem. This prototype allows you to view and explore the structure of any PDF of your choice and interactively create and run wrapper programs on it.

How to create a wrapper in four simple steps:

1. Open the document

- Click on “*Open document*” and select the PDF file you wish to open. The document will be displayed on the screen.

- You will notice that the light blue rectangles correspond to each individual line of text. These are the nodes of the graph. For certain extraction tasks, a coarser granularity is more desirable.*

- To show the edges of the graph, click on the *“Edges”* check-box in the *“Layers”* list. This list allows you to show or hide the other individual layers in our representation. Try it out!

2. Define the wrapper

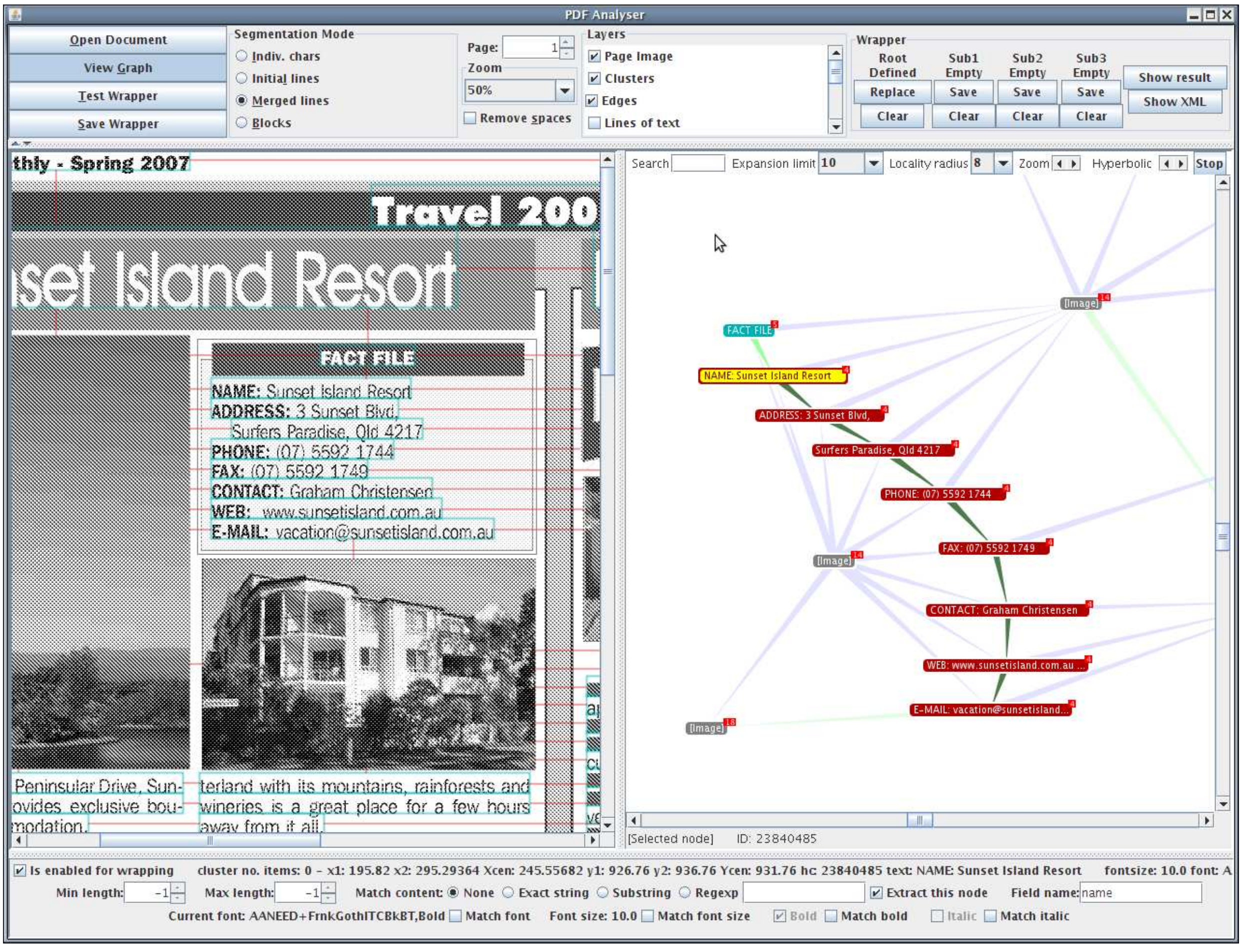
- Click on the *“View graph”* button. The screen will now be split into two halves: the page view will now occupy the left-hand side and the interactive graph will be displayed on the right-hand side.

- Select a rectangular section of the document by clicking and dragging on the page view. The corresponding graph structure will be displayed on the right-hand side of the screen. The edges of the graph represent neighbourhoods (in the four directions) between adjacent nodes. Neighbourhoods from left to right are shown with blue arrows; neighbourhoods from above to below are shown with green arrows.

- This sub-graph is the current wrapper definition. For each node and edge, you can set a wide variety of conditions by clicking on it and setting the appropriate values at the bottom of the screen.

- Other nodes in the document, which are not part of the current wrapper definition, are shown in a pale colour. These can be added to the wrapper definition at any time by right-clicking on the node and toggling the check-box *“Remove from instance”*. Current nodes can be removed from the wrapper definition in the same way.

* If you select the *“Blocks”* segmentation mode and re-open the document, the nodes will be grouped to represent coarser page elements such as paragraphs and table cells.



[Selected node] ID: 23840485

☑ Is enabled for wrapping cluster no. items: 0 - x1: 195.82 x2: 295.29364 Xcen: 245.55682 y1: 926.76 y2: 936.76 Ycen: 931.76 hc 23840485 text: NAME: Sunset Island Resort fontsize: 10.0 font: A

Min length: -1 Max length: -1 Match content: ● None ○ Exact string ○ Substring ○ Regexp ☑ Extract this node Field name: name

Current font: AANEED+FrnkGothITCBkBT,Bold □ Match font Font size: 10.0 □ Match font size ☑ Bold □ Match bold □ Italic □ Match italic

3. Try out the wrapper

- Click on *"Test wrapper"*. The results will be shaded in purple on the page view. If you did not set any conditions in Step 2, you could receive a large number of overlapping results at this stage!

4. Save and execute the wrapper

- This prototype allows you to define one root wrapper and up to three sub-wrappers. For each root wrapper result, the sub-wrappers will be run on the extracted nodes within that particular result. In this way, you can extract a complete record in the root wrapper and the individual data items in the sub-wrappers.

- Click on *"Show result"* to view the results of the root and sub-wrappers. The root wrapper result will be shown in purple and the sub-wrapper results in green on the page view. To see the corresponding XML output file, click on *"Show XML"*.

Contact details

Tamir Hassan
Database and Artificial Intelligence Group 184/2
Information Systems Institute
Vienna University of Technology
Favoritenstraße 9-11
A-1040 Wien
Austria

*This work is funded by the
Austrian Federal Ministry for
Transport, Innovation and
Technology
(Project no: 815128/9306)*

Email: *hassan@dbai.tuwien.ac.at* Web: *http://www.tamirhassan.com*