GraphWrap: Interaktive Datenextraktion aus PDF-Dokumenten mittels Graph-Matching-Techniken

Haben Sie schon einmal größere Mengen von Daten wie Spezifikationen, Abmessungen, Kontaktdaten oder Preise aus einer oder mehreren PDF-Dateien extrahieren wollen? Dann lassen Sie uns Ihnen *GraphWrap* – ein neues Verfahren zur gezielten Datenextraktion, sogenanntes „Wrapping", anhand von Graph-Matching-Techniken – vorstellen.

Es existerien bereits mehrere Systeme, die ähnliche Lösungen für Webseiten bieten. Im Gegensatz zu HTML-Dateien stellt die Datenextraktion von PDF eine größere Herausforderung dar, denn die hierarchische Struktur, die jedes Datenelement explizit abgrenzt und zusammengehörige Datenelemente verbindet, ist nicht vorhanden.

Indem wir eine Repräsentation eines PDFs in einer Graphstruktur samt Matchingalgorithmus entwickelt haben, haben wir eine Lösung für dieses Problem geschaffen. Mit diesem Prototypen können Sie selber die Struktur eines beliebigen PDFs ansehen und Wrapper-Programme interaktiv erstellen und ausführen.

In 4 Schritten zum Wrapper

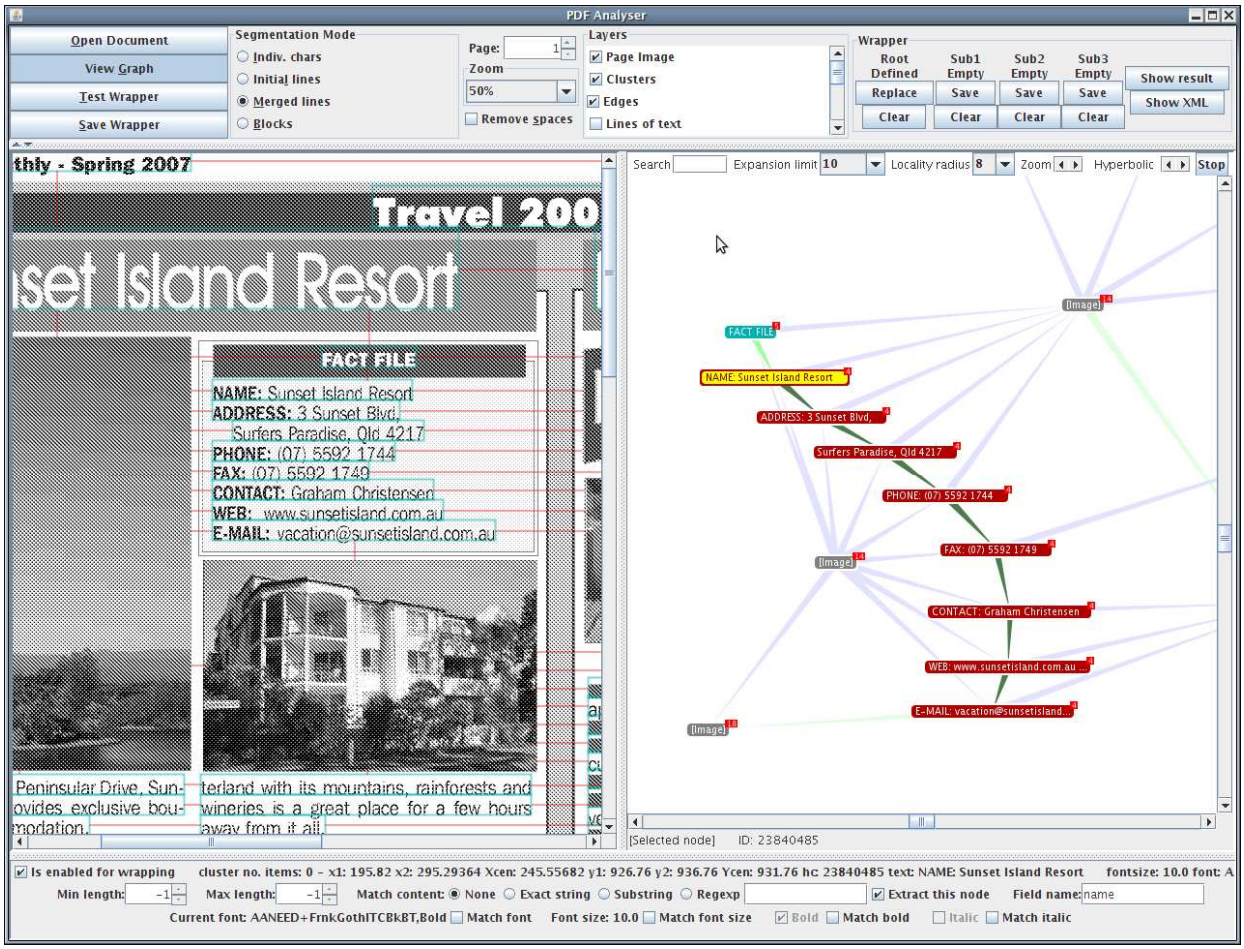
1. Dokument öffnen

- Klicken Sie auf *„Open document"* und wählen Sie die gewünschte PDF-Datei aus. Das Dokument wird angezeigt.
- Sie werden merken, dass die hellblauen Rechtecke jeder einzelnen Textzeile entsprechen. Diese sind die Knoten des Graphen. Für bestimmte Extraktionsaufgaben ist eine gröbere Repräsentation besser geeignet.*
- Um auch die Kanten einzublenden, klicken Sie auf das Häkchen *„Edges"* in der *„Layers"*-Liste. So können Sie auch die anderen Ebenen ein- und ausblenden. Probieren Sie es aus!

2. Wrapper verfassen

- Klicken Sie auf den Button *„View graph"*. Der Bildschirm wird nun in 2 Teile geteilt: Die Seitenansicht bleibt links; der interaktive Graph wird rechts gezeigt.
- Wählen Sie durch Click-and-Drag einen Teil des Dokuments aus. Die entsprechende Graphstruktur wird auf der rechten Bildschirmhälfte gezeigt. Die Kanten des Graphen entsprechen den Nachbarschaften (in den 4 Richtungen) zwischen den Knoten. Nachbarschaften von links nach rechts werden mit blauen Pfeilen gezeigt; von oben nach unten mit grünen Pfeilen.
- Dieser Subgraph ist nun der aktuelle Wrapper. Für jeden Knoten und jede Kante können Sie verschiedene Bedingungen setzen, indem Sie auf den Knoten oder die Kante klicken und die entsprechenden Werte unten am Bildschirm einstellen.
- Andere Knoten im Dokument, die nicht zum Wrapper gehören, werden in einer blasseren Farbe gezeigt. Sie können diese zum Wrapper jedoch hinzufügen bzw. andere Knoten wegnehmen. Um dies zu tun, klicken Sie mit der rechten Maustaste auf den Knoten und dann auf das Häkchen *„Remove from instance"*.

* Wenn Sie den Segmentierungsmodus *„Blocks"* auswählen und das Dokument erneut öffnen, dann werden die Knoten gröberen Einheiten wie Absätzen und Tabellenzellen entsprechen.



3. Wrapper ausprobieren

- Klicken Sie auf *„Test wrapper“*. Die Extraktionsergebnisse werden in der Seitenansicht in lila gezeigt. Wenn Sie gar keine Bedingungen in Schritt 2 eingestellt haben, kann es dazu führen, dass Sie mehrere überlappende Extraktionsergebnisse bekommen!

4. Wrapper speichern und ausführen

- Dieser Prototyp gibt Ihnen die Möglichkeit, einen Wurzelwrapper *(„root wrapper“)* und bis zu drei Subwrapper einzustellen. Für jedes Extraktionsergebnis des Wurzelwrappers werden dann alle Subwrapper ausgeführt, beschränkt auf die Knoten des Wurzelwrapperergebnisses. Somit können Sie z.B. im Wurzelwrapper einen ganzen Datensatz und in den Subwrappern die einzelnen Datenelemente extrahieren.

- Um die Wurzel- und Subwrapperergebnisse anzusehen, klicken Sie auf *„Show result“*. Das Wurzelwrapperergebnis wird in lila und die Subwrapperergebnisse werden in grün gezeigt. Um die entsprechende XML-Ausgangsdatei zu öffnen, klicken Sie auf *„Show XML“*.

Kontakt

Tamir Hassan
Abt. für Datenbanken und Artifizielle Intelligenz 184/2
Institut für Informationssysteme
Technische Universität Wien
Favoritenstraße 9-11
A-1040 Wien
Österreich

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

Diese Arbeit wird durch das österreichische Bundesministerium für Verkehr, Innovation und Technologie gefördert (Projekt-Nr. 815128/9306)