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.

Installation

- Voraussetzungen für den Prototyp sind, dass eine neue Version der Java Runtime Environment (*java.sun.com*) und Ghostscript (*www.ghostscript.com*) auf Ihrem Rechner installiert sind. Damit der Prototyp reibungslos funktionieren kann, wird es empfohlen, JRE 6 und GNU Ghostscript 8.64 (erhältlich unter *http://mirror.cs.wisc.edu/pub/mirrors/ghost/GPL/current/*) installiert zu haben.

- Extrahieren Sie die *zip*-Datei *graphwrap-testversion-0.x.zip* in ein beliebiges Verzeichnis.

- Das Wurzelverzeichnis beinhaltet drei Shellskripten (für Unix) und Batchdateien (für Windows). Bitte überprüfen Sie, dass die Pfade zu Java und Ghostscript Ihrer Installation entsprechen und diese ggf. ändern. Dies ist vor allem nötig, wenn Sie den Windows-Betriebssystem benutzen und eine andere Version von Ghostscript als 8.64 oder Ghostscript in einem anderen Verzeichnis installiert haben.

- Der Prototyp ist nun einsatzbereit. Um das GUI zu öffnen, führen Sie die Stapeldatei *gui* bzw. *gui.bat* aus. Im nächsten Abschnitt erfahren Sie, wie man mit dem GUI interaktiv Wrapper-Programme verfassen kann.

In vier Schritten zum Wrapper

1. Dokument öffnen

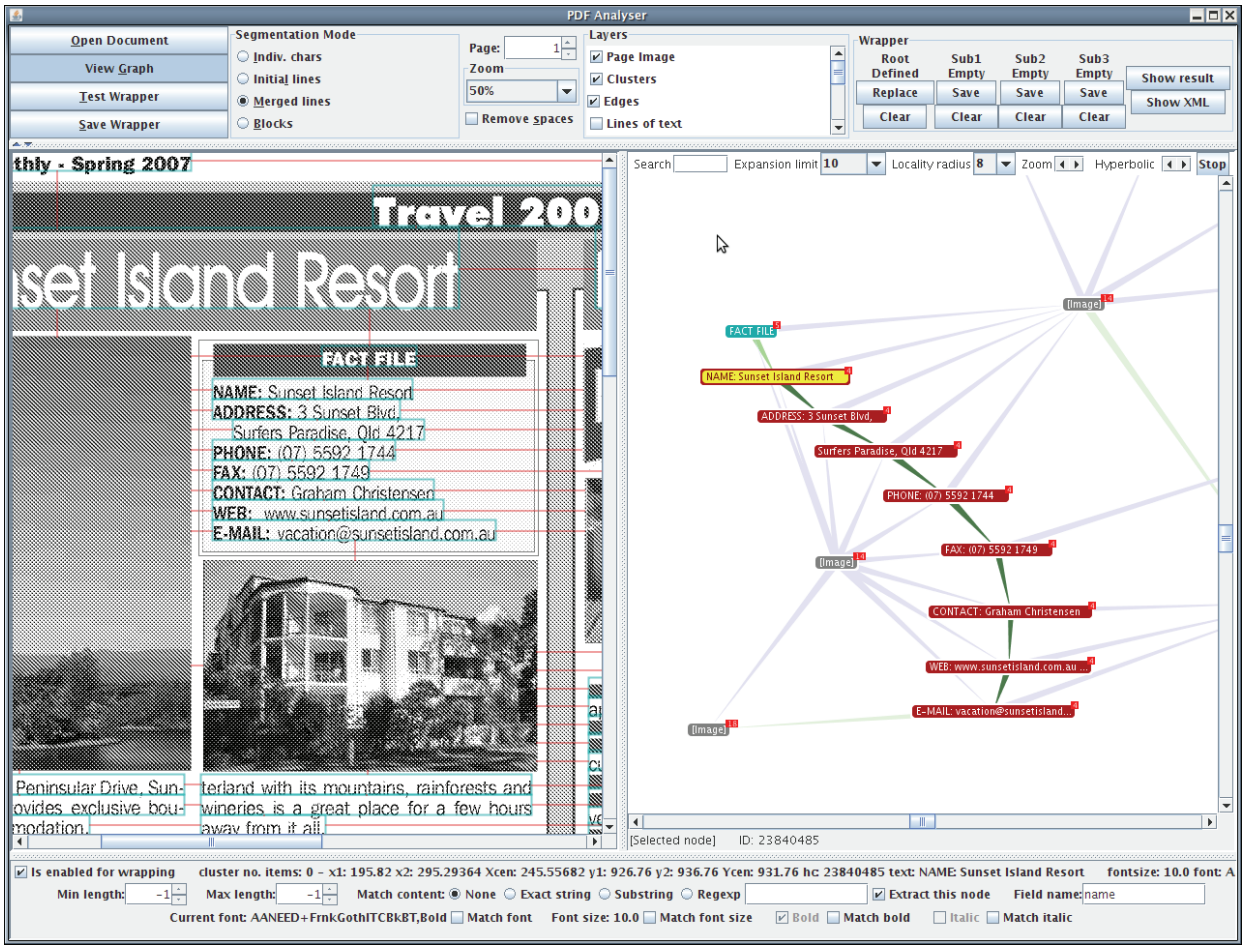
- Klicken Sie auf *„Open document"* und wählen Sie die gewünschte PDF-Datei aus. Das Dokument wird angezeigt. In dieser Anleitung wird das Beispieldokument *travel.pdf* verwendet, welches Sie im Unterverzeichnis *example* finden können.

- 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.

- Sie können eine gröbere Granularität einstellen, indem Sie den Segmentierungsmodus *„Blocks"* auswählen und das Dokument erneut öffnen. Dann werden die Knoten gröberen Einheiten wie Absätzen und Tabellenzellen entsprechen. Bitte beachten Sie, dass Änderungen

zu den Einstellungen der Dokumentverarbeitung erst bei der Öffnung eines neuen Dokuments wirksam sind.

- 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“*.

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!

- Beachten Sie bitte, dass der Wrappingalgorithmus nach *allen möglichen Vorkommnissen* der Subgraphstruktur auf der Seite sucht. Standardmäßig werden alle Knoten mit allen anderen Knoten und alle Kanten in einer bestimmten Richtung mit allen anderen Kanten in dieser Richtung gepaart. Durch Einstellen von Bedingungen können sie dieses Matching beschränken, in dem z.B. nur jene Knoten, die einen bestimmten Text beinhalten bzw. nur jene Kanten, deren Länge zwischen bestimmten Grenzwerten liegt, gepaart werden.

4. Wrapper speichern

- Klicken Sie auf *„Save wrapper“* und geben Sie einen beliebigen Namen ein. Bitte vergessen Sie die *.xml*-Endung nicht.

Beispiel *„Travel Monthly“*

Im Unterverzeichnis *example* finden Sie *travel.pdf*, eine Beispielseite von mehreren hundert Seiten, die von *www.travelmonthly.com.au* herunterladen werden können. In den nächsten Schritten erfahren Sie, wie datenbankorientierte Daten, wie die vier *„fact files“* auf der Beispielseite, mit GraphWrap extrahiert werden können. Nachdem Sie einen Wrapper für diese Seite verfasst haben, können Sie diesen auch auf den anderen Seiten von der Travel-Monthly-Website ausprobieren.

Bitte öffnen Sie zuerst das Dokument *travel.pdf* mit dem Segmentierungsmodus *„Merged lines“*. Nachfolgend klicken Sie auf *„View graph“*, um die Graphansicht zu öffnen.

1. Verfassen eines einfachen Wrappers

- Wählen Sie durch Click-and-Drag die ganze *„fact file“* oben links auf der Seite aus. Bitte überprüfen Sie, dass auch die Überschrift *„FACT FILE“* innerhalb der Auswahl liegt. Der entsprechende Subgraph wird nun rechts gezeigt.
- Wie schon beschrieben, definiert dieser Subgraph schon ein Wrapper-Programm. Durch Klicken auf *„Test wrapper“* können Sie dieses ausprobieren und werden merken, dass es viele überlappende Ergebnisse liefert. Der Grund dafür ist, dass es nach allen möglichen Kombinationen von vertikal nebeneinander Textzeilen sucht, egal welcher Text sie beinhalten oder wieviel Platz zwischen den Zeilen liegt.
- Klicken Sie in der Graphansicht auf den Knoten mit dem Text *„FACT FILE“*. Die möglichen Bedingungen für diesen Knoten werden dann unten auf der Seite gezeigt. Klicken Sie auf *„Substring“* und geben Sie *„FACT FILE“* im Textfeld ein. Probieren Sie anschließend den Wrapper nochmal aus, indem Sie wieder auf *„Test wrapper“* klicken. Jetzt sollen nur vier Ergebnisse geliefert werden, da der oberste Knoten auf die Überschrift *„FACT FILE“* nun „verankert“ ist.
- Wenn Sie sich das Ergebnis genauer anschauen, werden Sie merken, dass nur zwei von den vier Fact files zur Gänze ausgewählt wurden. Das liegt daran, dass die anderen Datensätze mehrere Zeilen beinhalten als die Beispielinstanz. Um flexibelere Wrapper zu verfassen, die alle Fact files ganz auswählen würden, benötigen Sie *Multiple-match-Kanten*.

2. Multiple-match-Kanten

- Wählen Sie durch Click-and-Drag die ganze *„fact file“* oben links auf der Seite aus. Bitte

überprüfen Sie, dass auch die Überschrift „*FACT FILE*“ innerhalb der Auswahl liegt. Der entsprechende Subgraph wird nun rechts gezeigt.

- Wie schon beschrieben, definiert dieser Subgraph schon ein Wrapper-Programm. Durch Klicken auf „*Test wrapper*“ können Sie dieses ausprobieren und werden merken, dass es viele überlappende Ergebnisse liefert. Der Grund dafür ist, dass es nach allen möglichen Kombinationen von vertikal nebeneinander Textzeilen sucht, egal welcher Text sie beinhalten oder wieviel Platz zwischen den Zeilen liegt.

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)