Database and Artificial Intelligence Group
Institute of Information Systems
Technische Universität Wien

Using Graph Matching Techniques to Wrap Data from PDF Documents

Tamir Hassan
hassan@dbai.tuwien.ac.at

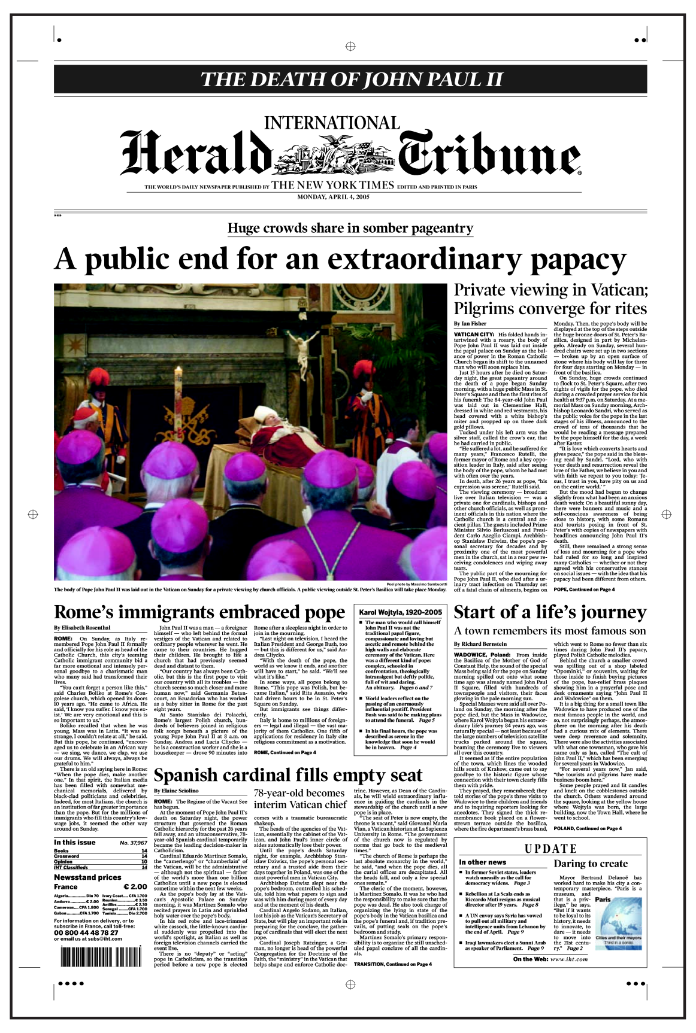
Robert Baumgartner
baumgart@dbai.tuwien.ac.at

Background

The **Lixto Visual Wrapper**, a product of research at DBAI, allows a user to interactively create a **wrapper program** to automatically extract data items from Web pages. The **hierarchical structure** of the HTML code is used by the wrapper to locate the data instances to be extracted.

Our goal is to extend the functionality of Lixto to PDF documents. As the PDF format is relatively unstructured, this is a challenging task. Documents can generally be thought of as possessing a **geometric structure** and a **logical structure**. In HTML documents, the structure of the code usually closely resembles the document's logical structure, and it is this structure that allows us to locate the data we require.

In PDF documents, the logical structure is usually not explicit in the code. In many cases, it is possible for a machine to automatically infer the logical structure from the geometric structure. This process is known as **document understanding**.



Left: A sample page of the *International Herald Tribune*, together with its corresponding segmentation

Right: Screenshot from our prototype, showing a table within a page and its respective graph

The Wrapping Process

First, we use the **PDFBox** library to parse the raw PDF file and return the text elements as a list of elementary objects, or **text fragments**, which represent one PostScript instruction.

The next step is to **cluster** these text fragments into blocks that can be said to be **atomic**, i.e. to correspond to one entity in the document's logical structure. In our case, the resultant elements are titles, headings, paragraphs, captions and individual data cells inside tables. A set of heuristics achieve this. An example of clustering is shown above.

We then use a number of document understanding methods to infer the logical structure of the document. This allows us to convert the PDF file into a well-structured HTML document, which can then be fed directly into the Lixto Visual Wrapper.

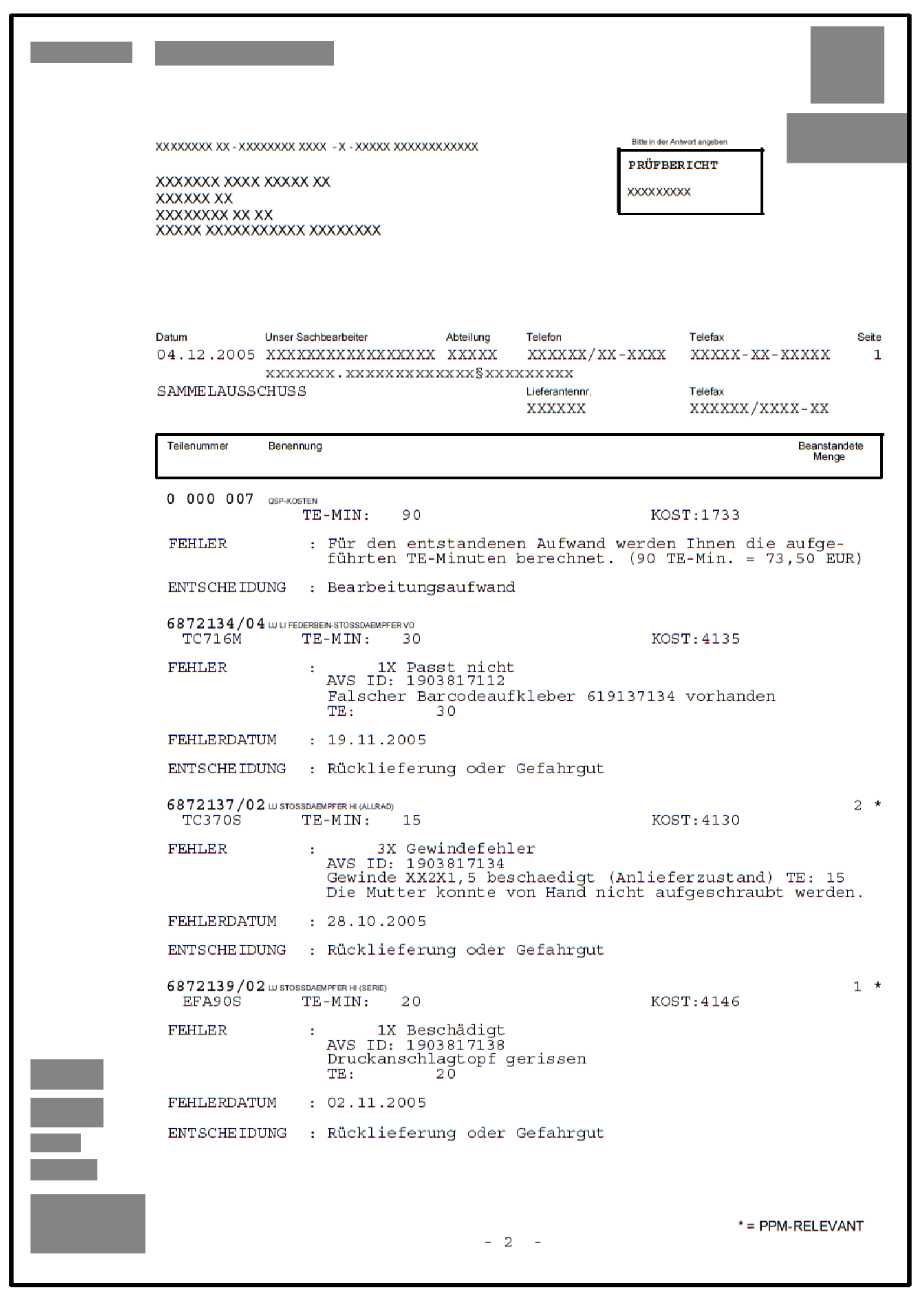
However, this approach has its limitations. Document understanding is inherently a very imprecise task, and there are many complex documents that cannot be sufficiently well understood without additional input from the user.

Our Graph Representation

We can also represent the atomic clusters as nodes in an **attributed relational graph**. Edges between nodes can be used to represent both geometric and logical relations. When the user interactively selects a wrapping instance, its corresponding sub-graph can easily be found.

We are currently investigating the use of error-tolerant graph matching techniques to locate other **logically similar** instances in the document.

This approach allows us to make use of the document's geometric structure, as well as the inferred logical structure, and any other metadata resulting from the document understanding process. As this representation is closer to the original than the HTML conversion, the burden of the document understanding process is partially shifted to the user, resulting in a more robust system.



[redacted] [redacted] [redacted]

XXXXXXXX XX - XXXXXXXX XXXX - X - XXXXX XXXXXXXXXXXX

XXXXXXX XXXX XXXXX XX
XXXXXX XX
XXXXXXXX XX XX
XXXXX XXXXXXXXXXX XXXXXXXX

Bitte in der Antwort angeben
PRÜFBERICHT
XXXXXXXXX

Datum	Unser Sachbearbeiter	Abteilung	Telefon	Telefax	Seite
04.12.2005	XXXXXXXXXXXXXXXXX	XXXXX	XXXXXX/XX-XXXX	XXXXX-XX-XXXXX	1
	xxxxxxx.xxxxxxxxxxxxx§xxxxxxxxx				
SAMMELAUSSCHUSS			Lieferantennr. XXXXXX	Telefax XXXXXX/XXXX-XX	

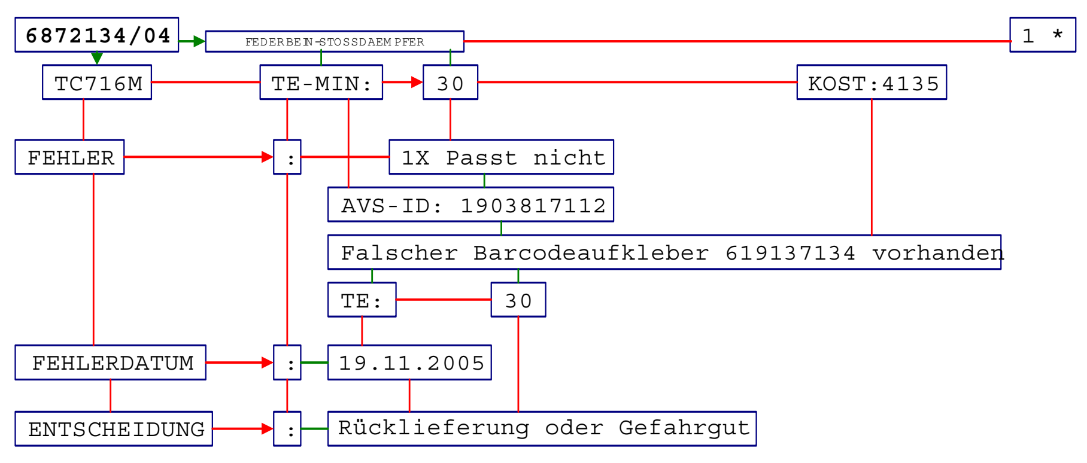
Teilenummer	Benennung	Beanstandete Menge

```
0 000 007 QSP-KOSTEN
           TE-MIN:   90                        KOST:1733
FEHLER       : Für den entstandenen Aufwand werden Ihnen die aufge-
               führten TE-Minuten berechnet. (90 TE-Min. = 73,50 EUR)
ENTSCHEIDUNG : Bearbeitungsaufwand

6872134/04 LU LI FEDERBEIN-STOSSDAEMPFER VO
  TC716M   TE-MIN:   30                        KOST:4135
FEHLER       :     1X Passt nicht
               AVS ID: 1903817112
               Falscher Barcodeaufkleber 619137134 vorhanden
               TE:        30
FEHLERDATUM  : 19.11.2005
ENTSCHEIDUNG : Rücklieferung oder Gefahrgut

6872137/02 LU STOSSDAEMPFER HI (ALLRAD)                        2 *
  TC370S   TE-MIN:   15                        KOST:4130
FEHLER       :     3X Gewindefehler
               AVS ID: 1903817134
               Gewinde XX2X1,5 beschaedigt (Anlieferzustand) TE: 15
               Die Mutter konnte von Hand nicht aufgeschraubt werden.
FEHLERDATUM  : 28.10.2005
ENTSCHEIDUNG : Rücklieferung oder Gefahrgut

6872139/02 LU STOSSDAEMPFER HI (SERIE)                         1 *
  EFA90S   TE-MIN:   20                        KOST:4146
FEHLER       :     1X Beschädigt
               AVS ID: 1903817138
               Druckanschlagtopf gerissen
               TE:        20
FEHLERDATUM  : 02.11.2005
ENTSCHEIDUNG : Rücklieferung oder Gefahrgut
```

[redacted]

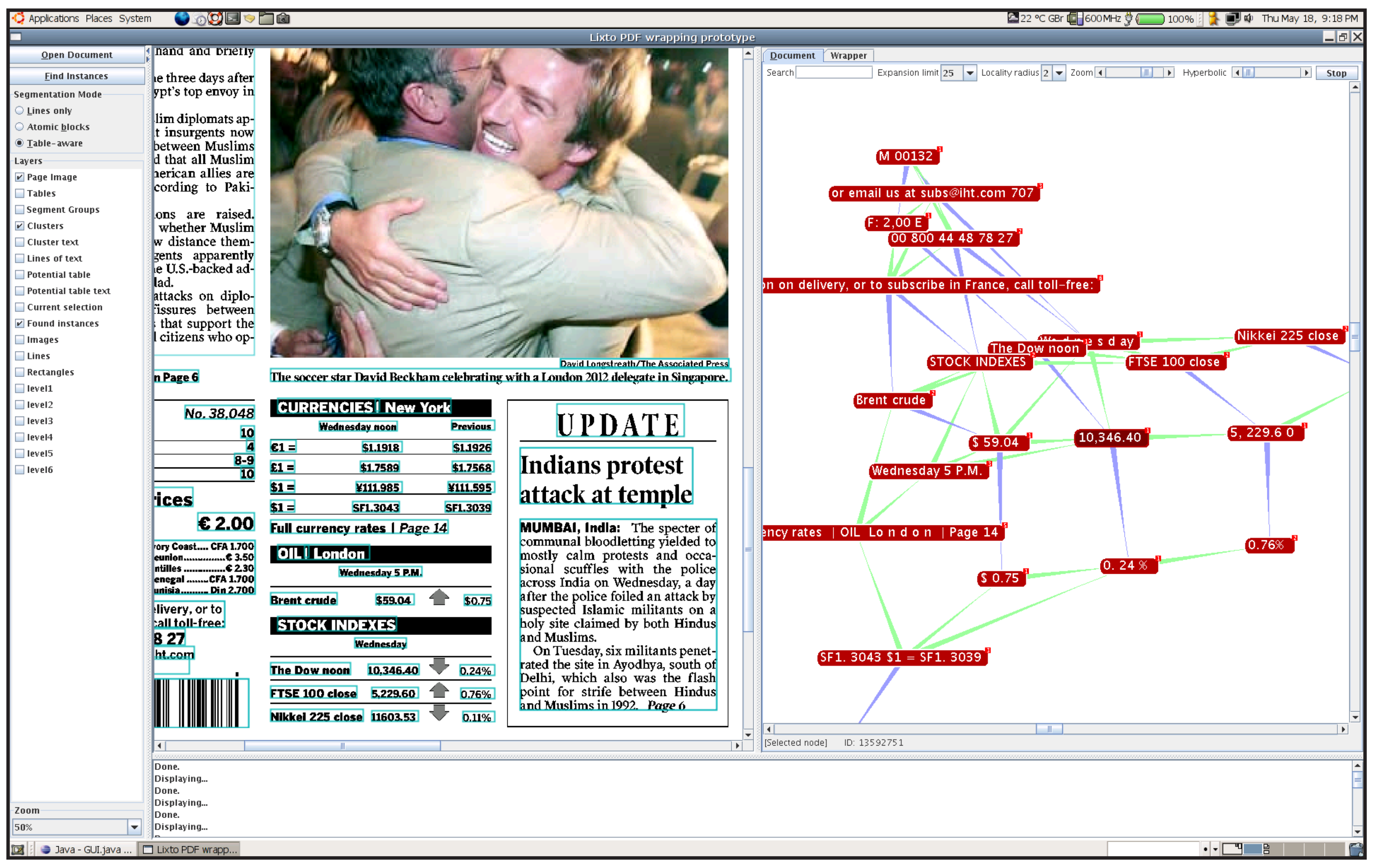
* = PPM-RELEVANT

- 2 -



Left: A real-life example of data to be wrapped. Brackets indicate the individual data instances. (*For confidentiality reasons, parts have been obliterated.*)

Above: Sub-graph for one record (wrapping instance) from this page. Edges with arrows represent **superior**-to-**inferior** relationships



Similarity Measures

The familiar notion of **edit cost** can be used to define the similarity of two sub-graphs. Allowed edits would include not just additions and deletions of single nodes or edges, but also complete rows of elements. For example, in a desired instance, a certain paragraph may be one line longer or a certain table may have one extra row added. Yet, the general **shape** of the instance will remain the same, as well as the logical structure (location of headings, etc.) Thus we are finding wrapping instances using both **logical** and **visual** similarity.

This measure could be extended to discriminate between headings and data. Any differences that affect heading elements would correspond to a profound change in the logical structure, and would therefore carry a higher associated edit cost.

We could also use the **weights** of each edge to represent the confidence or likelihood that the inferred relation is correct. An **error-tolerant** graph matching algorithm would use these weightings to select the best, or most likely instances.

This work is funded in part by the Austrian Federal Ministry for Transport, Innovation and Technology under the FIT-IT programme line as a part of the NEXTWRAP project, and by the REWERSE network of excellence.