Online Template Matching Over a Stream of Digitized Documents

Michael Stockerl, Christoph Ringlstetter
Gini GmbH
Eirini Ntoutsi, Matthias Schubert, Hans Peter Kriegel
University of Munich (LMU) CS Department
Personal Business Documents

Implement assistance systems for digitization of paper documents, document understanding, digital workflows

Digital Responses to Paper

Pay a bill.
Put an date into a calendar.
Archive a contract

Entry Point: smartphone as a scanner
Technological Environment
Technological Setting implies Complications

(1) Digitization: Image distortions (angle), Blur, Contrast

   Seeing the document as a set of localized strings
   - varying term positions
   - wrong or even missing terms

(2) Consumer Service:

   Stream of Arbitrary Documents (no closed set)
Idea of Document Templating

Preliminaries:

(1) Documents are issued using templates with certain Information bits at predefined locations and stable meta information.

(2) One User tends to receive often the documents from the same issuer. The crowd of users tend to receive a high portion of documents from few sources (amazon, telcos).

Conclusion: the template of a document can support information extraction, overcome Noise in a specific instance.
Motivation
Motivation

- **Understand** document using similar documents
Motivation

- Understand document with similar documents

document type: Invoice
Motivation

- Understand document with similar documents

Look for sender name at coordinates: x, y
Motivation

- Understand document with similar documents

Scope of the work!
Simple String Matching? Doesn’t work.

Same terms at different positions => high bag of words similarity e.g. tf-idf
Our working definition of a Template:
If two documents from different users share a template, documents are aequivalent except for user specific information

Words of header/footer with their normalized postitions
Our System dynamically creates templates and assigns incomming documents also represented as Header and Footer (Query) to them.
Methods to Match Documents with Templates

Must be robust against digitization errors: Strings missing, Strings different, Strings distorted

Methods for Query-Ranking Document2Template

(1) Word-Pos
(2) Bounding-Box
(3) Weight Decay
A grid is used to divide the document into regions. Each word is concatenated with its position of occurrence. Matching through tf-idf scoring of these artificial strings.

„Invoice-ID_101_88“

Wordpos

Evaluation

- Grid size defined at index time
- Strict borders for movements of Strings
- No ranking based on the layout
- Tf-idf scoring skewed by artificial Strings

Practical Experience: Precision too low for mobile images
Store words and positions into a spatial index

```json
{ word: 
  term: Mobilfunk-Rufnummer,
  location: {
    x: 530,
    y: 115
  }
}
parent: someld
```
Filter Index with bounding boxes of the query-document
Search for each word in its bounding box
Best template based on sum of tf-idf scores of terms
Evaluation:

+ flexibility search for a match around the query term.
- strongly dependent on a well selected size of the bounding box: too large => Precision Problem, too small => Recall Problem

⇒ Idea to compute the distances between the query term and a template term: the greater the distance, the smaller the weight.
Weight Decay

- Full-text search with query terms
- Score the distance between query term and matching term
- Aggregate scores with one hit per term (minimum distance)
Online Template Extraction from Stream

Template database learned from stream = open set of templates.

1. The photo/scan of the incoming paper document is OCRed => terms and their location
2. Footer/Header terms with normalized positions => query
3. Search for a fitting template in the template database (e.g. weight decay)
4. Match=> assign the document to the retrieved template and update the template representation
5. No Match: introduce new template

Figure 3: Overview of the overall workflow for processing a new document.
No Match
If \( d \neq t \in T \): new template created from the local terms of \( d \). normalized origin: upper and lower terms represent header/footer confidence value for terms \( \rho \) is initialized with one. (practically we work with a cut-off value to prevent matches)

Match
If \( d \approx t \in T \): all term positions of \( t \cap d \) are updated and for each term the confidence weight \( \rho \) is updated
- according to weight decay if match of the term
- discounted with penalty else

Elimination of terms:
If \( \rho \) goes below a threshold: term is discarded
The hope: personal data not template data will be discarded
Dataset small 771 labeled Documents

3 media types: scanned, mobile and rectified
20 classes and 10 invoices for each class (200 documents)
171 noise documents
- Weight decay works best
- Discard templates with small score improves quality
  = cut-off value
Performance is best on scanned documents.
Only *weight decay* maintains quality for the other media types.
→ *weight decay* robust to skewing of mobile images.
Large Dataset 12,000 scanned+mobile

Additional Experiments found:

(1) Evidence for reachable **Precision >95%** with weight decay
(2) Stream order has influence but even worst case scenario of documents inflowing per user leads to only **1-3% drop**
(3) Qualitative investigation of Template Refinement by term weight $\rho$:
   - **drop-outs: personal information**
   - high rated terms: header/footer template terms
(4) System **scaling**: processing times as expected **robust**
Conclusions

• Template representation by a set of localized terms of the header/footer
• Online System on stream with automatic template acquisition and update
• Weight decay method superior and in applicable terrain of Precision >95%
• System in place for fallback solution of imperfectly captured images and additional source for our information extraction cascade.

Recent Work:
• Further use for document anonymization, aggregation: inhouse dataset + BI
• Robust techniques for harvesting information from matched templates
Thank You!

Questions?!

Again thanks to LMU for the collaboration.