Retrieving OCR Text: A Survey of Current Approaches

Information Retrieval Lab
Illinois Institute of Technology

S. Beitzel
E. Jensen
D. Grossman
{steve, ej, grossman}@ir.iit.edu
Overview

• Models for OCR Text
• Processing OCR Text for Categorization
• Auto-correction of OCR Errors
Models for OCR Text

- Incorporate probabilities of typical OCR errors
- Harding, Croft, Weir (1997)
  - Addition of character-based n-grams to the model.
  - Ex: Environment
    - _en env nvi vir iro onm nme men ent – 3-grams
Auto-Correction of OCR Errors

• Liu (1991)
  – Classify each type of error
  – Use dictionary lookup to identify candidate terms

• Taghva, Borsack and Condit (1994)
  – Clustering to group mis-spellings in with their correctly mis-spelled terms
OCR Text for Categorization

• Hoch (1994)
  – Use of categorizer on OCR text, showed degraded performance with OCR data.

• Junker and Hoch (1997)
  – N-grams were used to show some improvement as well in [Junk97].
Summary

• Models exist for OCR retrieval
• N-grams have been shown to have some success
• No large standard test collection of OCR data, small collections exist with some early TREC data.