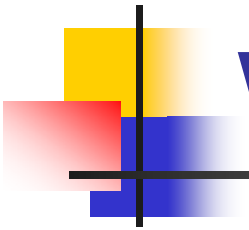
Testing appraisal models with digital corpora

Patricia Galloway
School of Information
University of Texas at Austin



Why appraise?

- Not enough room [Moore's Law?]
- Not enough time for description [Google?]
- Nobody will care about most of the material anyway [the long tail?]

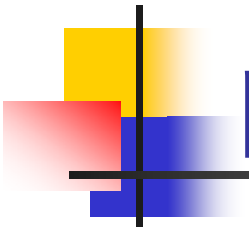
How appraise?

- Reject outright
- Accept everything
- Accept partially (reductive appraisal)
  - Accept by initial agreement only part of materials offered (front-end)
  - Perform granular “processing-appraisal” (back-end)

What are the effects of reductive appraisal?

- The ideal: appraisal accurately chooses
  - the best selection of materials
  - for informational and evidentiary uses
  - according to best knowledge of the time
  - ...and without going broke
- How (short of living forever) to test how closely this ideal is reached?

Reductive appraisal as preemptive IR



- Is there a fit between appraisal and conventional IR methods?
- Appraisal as preemptive information retrieval
  - IR selects desirable records, but can always come back to original corpus
  - Appraisal selects desirable records, discards the rest; original corpus is gone
- Can evaluation methods and measures borrowed from IR be used?

Testing appraisal effectiveness against digital corpora

- Digital corpora permit digital tools, so digital corpora permit complete testing
- Sources of digital tools:
  - Corpus linguistics
  - Literary analysis tools (e.g., style, authorship)
  - Text mining/clustering
  - Information retrieval evaluation tools

Proposed experiment

- Begin with corpus
- Use simple appraisal model to reduce corpus
- Measure “information loss”

Example: PKG's MDAH email

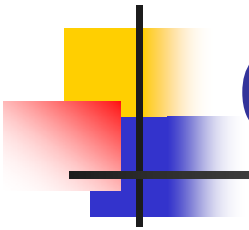
- PKG's MDAH email, 1997
  - Sent only
  - Attachments removed
- Consistent class of records
- No privacy concerns
- Potential for classification
  - Topics
  - Correspondents

Appraisal modelling

- Can appraisal methods be modelled formally?
  - Maybe not simply (cf. Gilliland's results)
- Selection constraints
- Selection as decision tree
- Appraisal as data-reduction process

Appraisal data reduction

- Implementing data reduction procedure
- Reduction of corpus
  - FBI appraisal decision tree (based on results of the FBI appraisal)
  - Selection profiles applied to Sent97: systematic sample, “fat files”



Corpus preprocessing

- Isolating one message per file for entire corpus
- Tokenization of all messages, including
  - Removal of headers
  - Stopword removal
  - Stemming
- Derivation of reduced versions of corpus
- Preparing term/document incidence and term/document frequency matrices
- Calculating distance and similarity measures

Analysis of reduced versions against original corpus

- Internal semantic structure (clustering of tokens)
- Network structure (correspondents, dates)
- Similarity in vector space
- Information gain (or loss)

Larger project

- Creating multiple formal appraisal models from case studies in archival literature
- Testing appraisal models against appropriate digital corpora and each other

Characterizing appraisal with more elaborate formal models

- Operationalizing implied model of record production (provenance)
- Appraisal modelled explicitly as data reduction process
- Discovering and specifying formal effects on content through automated content analysis

Characterizing specific appraisal contexts

- Actual digital record corpus
  - Formal methods for characterizing corpus
- Stakeholder/corpus actor-network as provenance specification
  - Correspondence analysis
- Effects of assumptions on selection procedures

Evaluating appropriate appraisal procedures

- Choosing a digital corpus
  - As generic model for similar digital collections
  - For analogous behavior to some paper collection of interest
- Testing against formal appraisal models
- Comparing results to appraisal goals
- Formally defining appraisal method choice as a function of “acceptable loss”