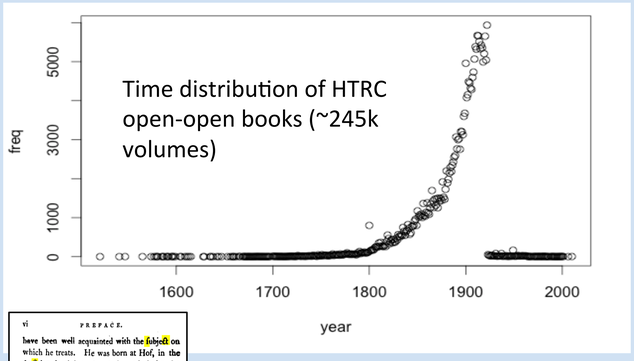
Toward Enhanced Metadata Quality of Large-Scale Digital Libraries: Estimating Volume Time Range

Siyuan Guo, Trevor Edelblute, Bin Dai, Miao Chen, Xiaozhong Liu
Indiana University

The Big Challenge

It is common that metadata records are incomplete in large-scale digital libraries. E.g. 13% publish dates are missing in HT's non-Google digitized public domain collection



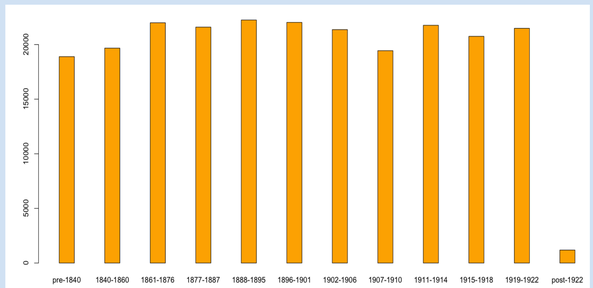
vi PREFACE.

have been well acquainted with the subject on which he treats. He was born at Hof, in the Austrian dominions, in 1748; and, before he embraced the monastic life, was known by the name of John Philip Wefdin. He was feven years Profeffor of the Oriental Languages in the Propaganda at Rome, and fince his return from India has publifhed feveral works relating to that country.

In regard to the prefent work, Dr. Forfter, in his Preface to the German Edition, fays:

Questions

When is a book published?
Can we tell it by looking at the book content?

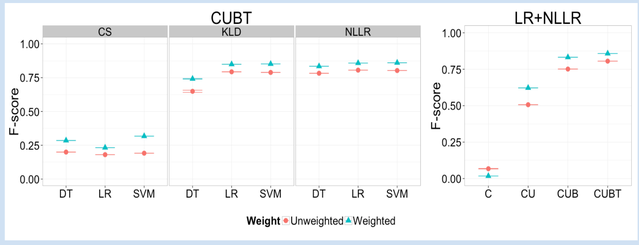


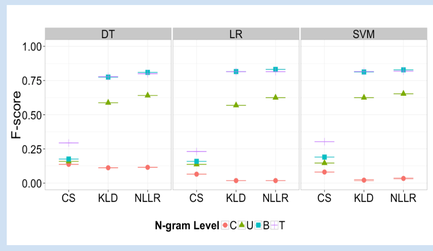
12 Chronons (bins) of books

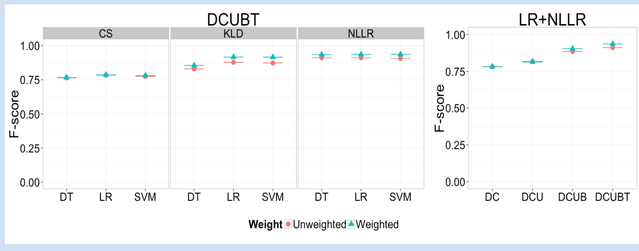
Experiment Design & Results

- ✓ Compare between classifiers
- ✓ Compare between n-grams
- ✓ Compare between similarities
- ✓ Compare between combinations of the above
- ✓ Observe baseline (FD feature only)

	BL	LR	DT	SVM
F-score	74.8	78.2	78.7	77.0
Precision	79.1	79.9	78.8	79.3
Recall	73.2	77.7	78.7	76.4







Feature Sets

Feature set options =

Use Weight Adjuster based on Temporal Entropy (TE)
Yes
No

×

N-grams
Characters
Uni-grams
Bi-grams
Tri-grams

×

Document-Chronon Similarity
Cosine Similarity (CS)
KL Divergence (KLD)
Normalized Log-Likelihood Ratio (NLLR)

+ **First-Date-In-Text (FD)**

+ **Date-Distribution-In-Text (DD)**

TE measures the sensitivity of a term to different chronons

Similarity between a document vector and each chronon vector

Machine Learning

12 Variables for Each Option
Similarity (Document, Chronon 1)
…
Similarity (Document, Chronon 12)

Features → training → Classifiers

Classifiers
Logistic Regression (LR)
Support Vector Machine (SVM)
Decision Tree (DT)

Takeaways

- Higher-order n-grams like bigrams and trigrams are more effective than unigrams
- Proposed bag-of-character to capture OCR errors as a feature
- Temporal entropy (TE) is an effective term weighting strategy
- The methodology works reasonably well on HT digital library data
- Can generalize to other large-scale digital library metadata records

Acknowledgement to HathiTrust Research Center for the NGPD data set and metadata.

Poster based on Guo, S., Edelblute, T., Dai, B., Chen, M., & Liu, X. (2015). Toward Enhanced Metadata Quality of Large-Scale Digital Libraries: Estimating Volume Time Range. iConference 2015. Newport Beach, CA.