Partial Duplicate Detection for Large Book Collections

I. Zeki Yalniz, Ethem F. Can, R. Manmatha

Dept. of Computer Science
University of Massachusetts
Amherst, MA, USA
Book Collections

- Large scanned book collections
  - Scanned page images with noisy OCR output
  - Metadata (sometimes incorrect / incomplete)

- Our aim – find different versions, editions and prints of the books in the collection
  - Partial duplicate detection problem
Problem characterization

• Finding partial duplicate book pairs
  o Most books do not have any duplicates
  o Books are almost never full or near duplicates:
    - extra/missing text
    - different editions, volumes, prints, compilations
    - scanning errors
    - OCR errors, spelling differences etc.
  o Do it at large scale

• Which portions of the book are duplicated?
  o Not dealt with here
ACT I. SCENE III.

By being in his eye, Most gracious duke,
To my unfolding lend your prosperous ear;
And let me find a charter in your voice,
To assist my simplicity.

Duke. What would you, Desdemona?

Desdemona. That I did love the Moor to live with him,
My downright violence and storm of fortunes
May trumpet to the world: my heart's subdued
Even to the very quality of my lord:

I saw Othello's visage in his mind,
And to his honours and his valiant parts
Did I my soul and fortunes consecrate.
So that (deere Lords) if I be left behind
A Moth of Peace, and he go to the Warre,
The Rites for why I love him, are bereft me:
And I a heavy interim shall support
By his dear absence. Let me go with him.

Othello. Let her have your voices.

Vouch with me, heaven, I therefore beg it not,
To please the palate of my appetite,
Nor to comply with heat—the young affects
In me defunct—and proper satisfaction,
But to be free and bounteous to her mind:
And heaven defend your good souls, that you think
I will your serious and great business scant
For she is with me. No, when light-wing'd toys
Of feather'd Cupid seal with wanton dulness
My speculative and offic'd instruments,
That my disports corrupt and taint my business,
Let housewives make a skillet of my helm,
And all indig and base adversities
Make head against my estimation!

Duke. Be it as you shall privately determine,
Either for her stay or going. The affair cries haste,
And speed must answer it.
England and Christendom

CONTENTS.

THE CROWN IN COUNCIL ON THE ESSAYS AND REVIEWS 3
THE CONVOCATION AND THE CROWN IN COUNCIL 33
THE WORKINGS OF THE HOLY SPIRIT IN THE CHURCH OF ENGLAND 81
THE REUNION OF CHRISTENDOM 137
APPENDIX 227

REUNION OF CHRISTENDOM.

A PASTORAL LETTER TO THE CLERGY ETC.

BY HENRY EDWARD MANNING
ARCHBISHOP OF WESTMINSTER

LIBRARY ST. MARY'S COLLEGE
Applications

• Finding different versions, editions, prints
  o Scholars in social disciplines
• FRBRization
  o New library metadata standards requires knowing which books are related.
• Propagating annotations
• Eliminating noisy versions of books
• Improve text search
• OCR error correction etc.
Related work

• Most work on web documents assumes near duplicates
  o Fingerprinting (shingling) techniques sub-samples n-grams of words [Broder’97, Brin’95, Bernstein’04, Manber’94, Shivakumar’99]
  o Relative frequency techniques: using words with similar document frequencies [Shivakumar’95, Hoad’03]
  o I-Match: uses certain words in the vocabulary of the document [Chowdhury’02]

• Global alignment has been used to find plagiarized passages [Clough’03]

• Problems with existing techniques:
  o OCR errors, spelling variations corrupt shingles
  o Vocabulary and term frequencies differ if the content overlap is small
  o Global alignment - too slow for books
Our methodology

• Represent each book by the sequence of words which appear only once in the book.
  o Descriptive of the content and flow of ideas in the book
  o Compact representation
  o Robust to OCR errors

• Compare books by aligning the lists of unique words using Longest Common Subsequence (LCS) algorithm
  o LCS is normally $O(n^2)$
  o For unique word sequences it can be done in $O(n \log \log n)$

• Do it for all $n(n-1)/2$ pairs of books in the collection
To purchase peace and rest.
It’s no in makin muckle, mair,
It’s no in books, it’s no in lean.

To purchase peace and rest.
It’s not in making much more,
It’s not in books, it’s not in learning,
To make us truly blessed.
Similarity measures

• Similarity is measured by the length of LCS(X, Y)
• LCS length depends on the length of the books aligned, therefore it has to be normalized

\[
DUPNIQ - cs(X, Y) = \frac{|LCS(X, Y)|}{\sqrt{|X||Y|}}
\]

\[
DUPNIQ - its(X, Y) = \frac{\log (|LCS(X, Y)|)}{\log (|X| + |Y| - |LCS(X, Y)|)}
\]
Baselines

Unique word overlap (UWO): the set of unique words in the book

Vocabulary overlap (VO): the set of words in the vocabulary of the book

Shingling: n-grams of words with sampling 0 mod p
(n = 4, p = 50)

Overlap is defined by Jaccard distance

\[ \text{Jaccard distance} = \frac{|A \cap B|}{|A \cup B|} \]
## Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Language</th>
<th>#books</th>
<th>#pairs</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>English</td>
<td>151</td>
<td>67</td>
<td>manual</td>
</tr>
<tr>
<td>1K</td>
<td>English</td>
<td>1000</td>
<td>258</td>
<td>manual</td>
</tr>
<tr>
<td>3K</td>
<td>French</td>
<td>2884</td>
<td>483</td>
<td>manual</td>
</tr>
<tr>
<td>100K</td>
<td>English</td>
<td>103455</td>
<td>45485</td>
<td>Semi-automatic</td>
</tr>
<tr>
<td>Partial set</td>
<td>English</td>
<td>458</td>
<td>460</td>
<td>manual</td>
</tr>
</tbody>
</table>

- Books are downloaded from the Internet Archive website
- Semi-automatic annotation – please see the paper for details
Partial duplicate dataset

1K

3K French

100K

precision and recall for different datasets and methods.
Evaluation

• Characteristics of false matches in the 100K dataset (0.004% of total matches):
  - 14% dictionaries and encyclopedias
  - 16% religious books
  - 12% literary books
  - 6% technical books

• Relatively more robust to high rates of OCR errors

• Speed: 30 minutes to process all book pairs in the 100K collection using 350 cores (without term indexing)
Thank you!

Questions?