2011

Mining Relational Structure from Millions of Books

David A. Smith
R. Manmatha, *University of Massachusetts - Amherst*
James Allan

Mining Relational Structure from Millions of Books

Position Paper

David A. Smith, R. Manmatha, and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003, U.S.A.
{dasmith,manmatha,dasmith}@cs.umass.edu

ABSTRACT

Existing large-scale scanned book collections have many shortcomings for data-driven research, from OCR of variable quality to the lack of accurate descriptive and structural metadata. We argue that complementary research in inferring relational metadata is important in its own right to support use of these collections and that it can help to mitigate other problems with scanned book collections.

Categories and Subject Descriptors

H.3.7 [**Information System**]: Digital Libraries

General Terms

Design

Keywords

relational metadata, partial duplicate detection

1. INTRODUCTION

Since the advent of the digital computer, scholars have been conducting book digitization projects to support their research. In addition to projects initiated by single researchers, larger undertakings include the Perseus Project, the Walt Whitman Archive, and the Women Writers Project as well as the digital libraries of learned societies such as the ACM and IEEE. Often, these projects have encoded a coherent corpus of interest to one scholar or group of scholars. Scholars reusing these corpora often find that they have to augment them with additional material at considerable effort.

Other digitization projects such as the ones undertaken by Google and the Internet Archive have instead focused on producing a large amount of data by scanning whatever they can find in certain libraries. While these datasets are a great scholarly resource, selecting and organizing the material in such collections for use by researchers presents a challenge given the diversity of the data and the errors in the metadata. In particular, researchers have been hampered by the inaccuracy of descriptive metadata in broad-coverage scanned collections—most obviously, by the lack of accurate dates for many books or sub-parts of books, and even by the lack of accurate language identification. In addition, researchers have worked to remedy the lack of structural metadata, such as chapter and article divisions.

We believe that additional research should focus on extracting **relational metadata**—for instance, the relationships among books that are duplicates or translations of each other, or the citation or quotation of one passage by another. We anticipate several productive research directions for mining relational structure from online books, in particular:

- identifying groups of books that are duplicates, partial duplicates, or translations of each other;
- mapping patterns of text reuse in order to identify canonical texts, distinguish quoted material in later works for better corpus linguistics, and make citation analysis more robust;
- mapping topical similarities and differences across documents from various genres, languages, and time periods; and
- exploiting the graph of text relationships to improve the accuracy of book and sub-book descriptive and structural metadata.

In what follows, we describe several of these research directions and note how they can be applied to other digital library tasks. We first present a rationale and prospectus for large-scale partial duplicate detection (§2). Then, we explore finer grained analyses with full book alignments (§3). Finally, we sketch how relational metadata can feed back into improved descriptive metadata (§4).

2. INFERRING BOOK RELATIONSHIPS

A scholar who wants to study Vergil's *Aeneid* may want to create a corpus containing all versions of the *Aeneid* available in a digital library (for example the Internet Archive). This may include modern or ancient commentaries (for example, Servius' or Donatus') in Latin as well as translations of the poem into other languages. While metadata might appear to be a good way of solving this problem, there are several issues with it. Some versions may be listed with a different title or listed under the commentator's name. Metadata may also be inaccurate or misleading (see Tables 2

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

and 3 below). The Internet Archive has a variety of titles and authors for different versions of Shakespeare's *Othello*. Even if metadata could be used for finding different versions, it cannot solve the problem of finding which portions of the books correspond.

We are exploring a technical solution that consists of (a) finding all books which are partial duplicates of a given book and also (b) detecting which portions of two books are common. Note that a near duplicate—the definition commonly used in information retrieval for "de-duping" search results or for plagiarism detection—is a special case of a partial duplicate. Our solution would enable scholars to create their own corpora—whether it involves collecting all versions of the *Aeneid* or *Othello*—and it can be done almost completely automatically. For this task we are not interested in finding quotations but duplicates at a coarser granularity. Individual sentence or paragraph level quotations can occur in many different books. As an example, a quotation from *Hamlet* may appear in all kinds of books that have nothing to do with *Hamlet* and hence are not relevant to the task of finding partial duplicates.

Most work on detecting duplicates is focused on finding near duplicates for web pages using chunking techniques [1, 3]. Two documents are near duplicates if they only have a few small differences. Chunking involves creating a fingerprint using n-grams. In other words, a numerical fingerprint replaces every n-consecutive words. Two documents are compared by seeing if they have a lot of n-grams or chunks in common. It is very expensive to use all the chunks in a document so the usual process is to sample the chunks, and there are a large number of techniques based on different sampling techniques [5, 10, 9, 2, 12, 1]. One common technique is to use mod p sampling. Each chunk is hashed (given a random number key) and only chunks whose hash value is divisible by p are preserved; p is typically 25 or 50. This approach achieves efficiency at the cost of accuracy. Another technique for comparing computer files is implemented by the widely-used utility `diff` [7]. This is based on creating a fingerprint from each line, and then comparing the two sequences of fingerprints. However, `diff` fails for most of the texts we are interested in due to the fact that OCR errors or changes in formatting will cause the fingerprints to be different in the two sequences.[1] Another class of techniques use the assumption that the relative frequencies of words in duplicated documents must be similar [6, 11].

The problem we propose to solve is different in several respects. First, we use books which are generally much longer than web documents (most web pages are 1 or 2 pages in length). Second, the books used here have OCR errors which sometimes affect a significant fraction of words, especially for languages such as Latin. Third, there may be significant formatting differences between the two versions. Finally, the problem here involves finding not only near duplicates but partial duplicates. Since many of our books are different editions or versions of the books, the differences can be substantial. Figure 1 shows a book with the text of *Othello* on the left and a critical edition containing several plays including *Othello* on the right. Only 30% of the material on the page is common (the red boxed areas) and even these have spelling and other changes. Chunking methods do not work as well when partial duplicates are involved.

Dataset	Proposed		baseline	
	P	R	P	R
Duplicate Test Set	0.99	1.0	0.91	0.88

Table 1: Precision (P) and Recall (R) scores on a 294 book dataset for alignment technique vs. a chunking baseline. See [13] for further details and evaluations.

Duplicate detection techniques also need to be very efficient. For example, 80% of the Internet Archive consists of English books. Given that the Archive currently has about 2.5 million scanned books, this means that to find a version of Shakespeare's *Othello* one needs to compare it against 2 million books, which can be very expensive if not done efficiently.

We have implemented a method that involves detecting duplicates using a new representation [13]. We first find words which are unique in the vocabulary of a book. Each book is represented as a sequence of its unique words. Two books are considered to be duplicates if there is a significant overlap in their unique word sequences. This works because the order of ideas in partial duplicates must be the same for some portion even if there is intervening commentary or other material. To find this overlap we need to align the sequences. This can be done using an algorithm to find the longest common subsequence (LCS).

The particular representation used has several advantages. First, this is a compact representation since the number of unique words is about 4% of the total number of words in a book. However, there are enough unique words to ensure a good match even with OCR errors. It is much less likely that a unique word will be in error than a chunk of n-grams. Unique words can be computed very efficiently since all that is necessary is to count words.

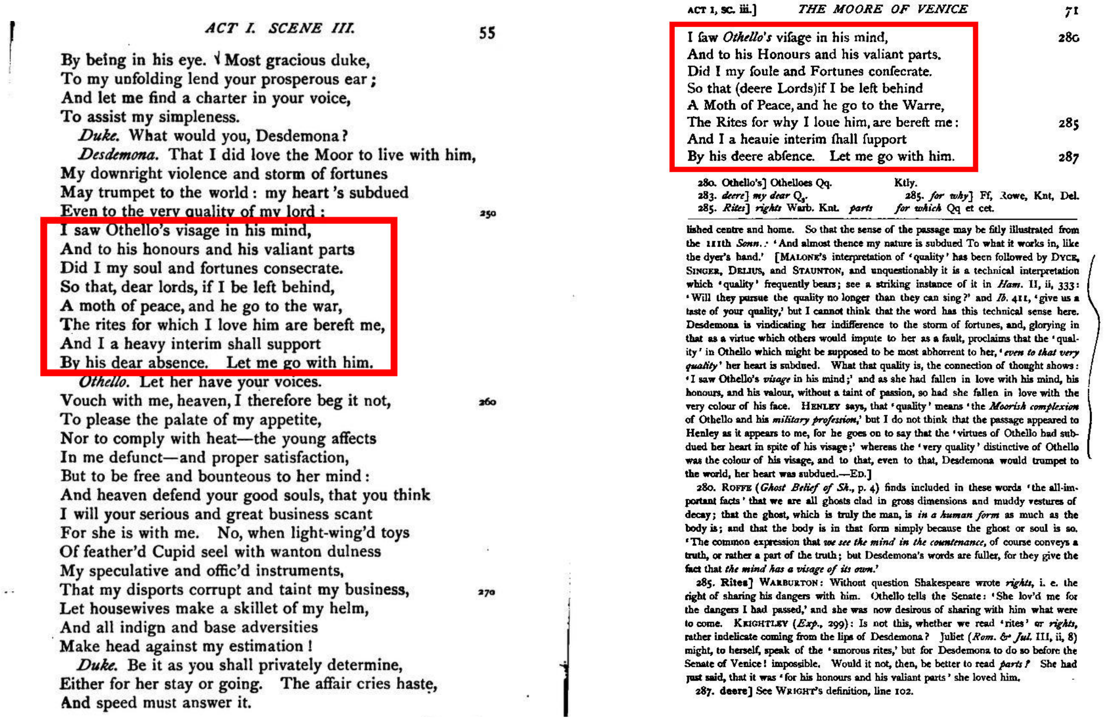
We show some initial results on a dataset of 294 English books where we compare every pair of books (roughly 900,000 comparisons) and classify duplicates (Table 1). A chunking technique with a 4-gram and a sampling factor of 25 is used as a baseline. As can be seen our proposed technique is more accurate and actually faster to compute. For further discussion, see [13, 8].

3. ALIGNING BOOKS FOR STRUCTURE DISCOVERY

We show some preliminary results for duplicate detection and also the resulting alignments. These were derived by taking just the text of one edition of Vergil's *Aeneid*, containing books 1–8, as a template and automatically checking about 24K Latin books for duplicates. To clarify: For the template or original we extracted the portion that corresponded to *Aeneid* books 1–8. We matched it against the entire content of the 24K books including possible introductions, commentaries, etc. The duplicate detection appears to be fairly robust, but the part which identifies which portions of a book are duplicated is still not completely solved. There are also some issues about how to visualize this. With that caveat, we generated some images to show what the alignment looks like.

For books which were classified as duplicates, a complete

[1] An original, and still popular, use for `diff` is in comparing computer language source code, where typographic lines are more significant than in most prose texts.



ACT I. SCENE III. 55

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 simpleness.
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, dear lords, if I be left behind,
A moth of peace, and he go to the war,
The rites for which 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 seel 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 indign 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.

(a)

ACT I, SC. iii.] *THE MOORE OF VENICE* 71

I ſaw *Othello's* viſage in his mind,
And to his Honours and his valiant parts.
Did I my ſoule 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 loue him, are bereft me:
And I a heauie interim ſhall ſupport
By his deere abſence. Let me go with him.

280. Othello's] Othelloes Qq. Ktly.
283. *deere*] *my dear* Q₃.
285. *Rites*] *rights* Warb. Knt. *parts*
285. *for why*] Ff, Rowe, Knt, Del. *for which* Qq et cet.

lished centre and home. So that the sense of the passage may be fitly illustrated from the 111th *Sonn.*: 'And almost thence my nature is subdued To what it works in, like the dyer's hand.' [MALONE'S interpretation of 'quality' has been followed by DYCE, SINGER, DELIUS, and STAUNTON, and unquestionably it is a technical interpretation which 'quality' frequently bears; see a striking instance of it in *Ham.* II, ii, 333: 'Will they pursue the quality no longer than they can sing?' and *Ib.* 411, 'give us a taste of your quality,' but I cannot think that the word has this technical sense here. Desdemona is vindicating her indifference to the storm of fortunes, and, glorying in that as a virtue which others would impute to her as a fault, proclaims that the 'quality' in Othello which might be supposed to be most abhorrent to her, '*even to that very quality*' her heart is subdued. What that quality is, the connection of thought shows: 'I saw Othello's *visage* in his mind;' and as she had fallen in love with his mind, his honours, and his valour, without a taint of passion, so had she fallen in love with the very colour of his face. HENLEY says, that 'quality' means 'the *Moorish complexion* of Othello and his *military profession*,' but I do not think that the passage appeared to Henley as it appears to me, for he goes on to say that the 'virtues of Othello had subdued her heart in spite of his visage;' whereas the 'very quality' distinctive of Othello was the colour of his visage, and to that, even to that, Desdemona would trumpet to the world, her heart was subdued.—ED.]

280. ROFFE (*Ghost Belief of Sh.*, p. 4) finds included in these words 'the all-important facts' that we are all ghosts clad in gross dimensions and muddy vestures of decay; that the ghost, which is truly the man, is *in a human form* as much as the body is; and that the body is in that form simply because the ghost or soul is so. 'The common expression that *we see the mind in the countenance*, of course conveys a truth, or rather a part of the truth; but Desdemona's words are fuller, for they give the fact that *the mind has a visage of its own.*'

285. **Rites**] WARBURTON: Without question Shakespeare wrote *rights*, i. e. the right of sharing his dangers with him. Othello tells the Senate: 'She lov'd me for the dangers I had passed,' and she was now desirous of sharing with him what were to come. KEIGHTLEY (*Exp.*, 299): Is not this, whether we read 'rites' or *rights*, rather indelicate coming from the lips of Desdemona? Juliet (*Rom. & Jul.* III, ii, 8) might, to herself, speak of the 'amorous rites,' but for Desdemona to do so before the Senate of Venice! impossible. Would it not, then, be better to read *parts?* She had just said, that it was 'for his honours and his valiant parts' she loved him.

287. **deere**] See WRIGHT'S definition, line 102.

(b)

Figure 1: Example scanned versions of "Othello". The common text is within the red boxes. Notice that text 1 is basically Othello while text 2 is a variorum version of Othello with extensive notes—every page has more notes than the main text. Variations in spelling, typography (both upper-cased nouns in the second version and also the use of f for s) and parentheses occur. Line numbers don't match up.

alignment between the original and each book was generated and used to create an image. Each image in Figure 2 shows the starting-query edition on top and the other edition below. The length shows the relative lengths of the two book pairs. The books are binned and if the alignment density of a bin is greater than 50% a green (lighter) bar is shown; otherwise it is red (darker). Clearly, this is only an approximate representation; much more work could be done to improve this visualization. For instance, vertical lines could link aligned portions of the two books, and extracted section information could label different portions of each image.

A complete alignment requires that every word in a book be matched against every word in another book. Our current implementation uses a version like the hierarchical Hidden Markov Model in [4]. Since some words are frequent (e.g., "the", "and") in our current implementation these tend to line up even in regions where there is no duplication. While the alignment is in general fairly reliable, since there are a lot of these stopwords this can cause some problems. There are also some issues with the visualizations, and further research could investigate how this can be done better. For example, the current version shows bars without indicating which specific portions line up and a user may be interested in this aspect. It is also important to show alignments not only at the macro level but at the micro level—or the level of sentences.

We envision an interactive tool to allow scholars to use these innovations. For example, the tool will allow the researcher to specify a clean work of say the *Aeneid* that the system will use to build a corpus. The user may then interactively visualize different aspects of the corpus to ask questions such as, given a pair of books how are they related? The user could specify a specific page or section of a book and see the corresponding portion in the other book as in Figure 1. The tool could also find a book with a commentary so that the researcher can look up the commentary for the portion she is reading. The user may also want to see corresponding sentences lined up one above the other to see if there are for example differences in spelling (as in the Shakespeare figure). Corpus- or book-derived statistics may also be of interest. For example: How many words are there in a section of the corpus? How often are the words in a section of the corpus repeated in other parts?

4. PROPAGATING METADATA

The relations among books and parts of books has further applications beyond corpus creation and structural analysis. Improved information about the relations among texts can help improve the accuracy of the descriptive metadata about texts.

Metadata can be propagated along the links of textual relationship discovered by the methods above. An important application of this metadata propagation is improving the date information about individual chunks of text. Systematic errors in dating are a common problem in large, automatically constructed archives. For example, multiple republications of canonical works can obscure the date to

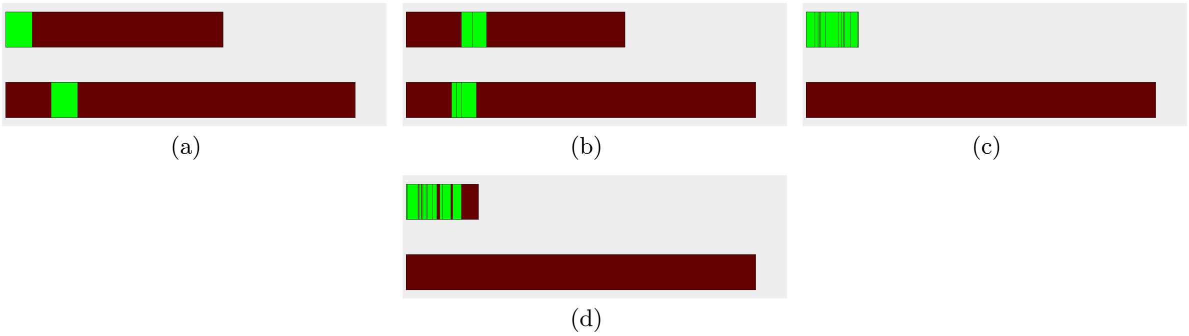


Figure 2: Books 1–8 of the Aeneid used as the template—the top bar in each figure. Green (lighter) regions shows matches; red (darker) portions which are not matched (or below the 50% threshold). (a) Second book contains Aeneid book 1, (b) Second book contains Aeneid book 3, (c) One bar is green while the other red. This indicates that there is a strong overlap but the second book has a verse on each page followed by commentary and hence falls below the 50% threshold—this is an artifact of the visualization, (d) The first 6 books of Aeneid are duplicated, the second book is an ancient commentary on the Aeneid which has a few lines or words followed by commentary, the red in the second book is an artifact of the visualization.

which we assign linguistic features of those works. If *The Anatomy of Melancholy*, from 1621, is republished in 1921, we do not want Burton's 17th century turns of phrase to be attributed to searches or statistical summaries of 20th-century language. Using the duplicate detection techniques from §2, a system could infer a single date for an entire cluster of similar documents. That consensus date could be the date of the earliest dated member of the cluster or simply the average of the dates in the cluster weighted by a prior distribution.[2]

Besides date information, we can exploit the relations among books to pool information about authors, titles, subjects, etc., which can improve search within these collections. Table 2, for example, lists books in the Internet Archive's collection that match Bernard Mandeville's *Fable of the Bees*. Book #3, a reprint from 1962, erroneously lists the editor, Irwin Primer, as the only creator, but apart from a brief introduction, its text is a good match for the 1729 edition. The two editions (#1 and #4) that list the authors as Mandeville would overrule this attributation. Book #2, which consists primarily of a later author's writing against Mandeville's argument, also contains the text of the original poem in an appendix. This appendix, when aligned to the original *Fable*, would inherit the author and date metadata from the rest of the cluster. Similarly, Table 3 shows a range of reprint dates for Mary Shelley's *Frankenstein* from 1823 to 1993. The preponderance of author information points to Mary Wollstonecraft Shelley although some outliers like "Wordsworth Collection" and "Kenneth Branagh" slip in.

In addition to inferring metadata for whole books, we can propagate information about partially aligned texts and quotations. When Plato quotes Homer, that doesn't mean Plato is speaking an early Greek literary dialect for a few lines, nor does Mark Twain archaize when quoting Sir Walter Scott. Similarly, a modern edition of Cicero that mentions "America" in its commentary does not imply that Cicero knew about America. When part of a book is aligned to another work (see §3), the aligned portions are likely to share metadata. Passages quoted from another work will also share metadata—such as date, authors, and title—with the source.

5. CONCLUSIONS

While much time could be spent solely on cleaning up the metadata in scanned book repositories, we believe that enlarging the problem to look at relations among their contents will give scholars more powerful tools. This is most obvious in the case of partial duplicates and quotations, which are not often subject to cataloguing. In addition, more refined models for inferring dates and other metadata in clusters of works are possible, and we hope to explore them in future work. The text reuse graph, for instance, provides pairwise ordering relations that can aid dating. Finally, we note that citation analysis has proved to be an important tool both in traditional scientometrics and in web search; enabling similar analyses for genres and time periods without settled citation schemes, and for books whose OCR makes citation parsing difficult, should provide another boost to mining books online.

6. ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0910884. Any opinions, findings, and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] Y. Bernstein and J. Zobel. A scalable system for identifying co-derivative documents. In *SPIRE*, pages 55–67, 2004.

[2] S. Brin, J. Davis, and H. Garcia-Molina. Copy detection mechanisms for digital documents. In *ACM SIGMOD*, pages 398–409, 1995.

[2] E.g., based on author lifespan information. In the absence of highly precise collations of variants among editions in an author's lifetime—again, *The Anatomy of Melancholy* furnishes a good example—such prior distributions can represent our uncertainty.

#	Creator	Title	Date
1	Bernard Mandeville	The Fable of the Bees	1729
2	Law, William, 1686-1761 Maurice, Frederick Denison, 1805-1872	Remarks on The fable of the bees	1844
3	Irwin Primer	The Fable Of The Bees Or Private Vices Publick Benefits	1962
4	Bernard Mandeville	The Fable of the Bees : Or, Private Vices, Publick Benefits: With an Essay on Charity and ...	1806

Table 2: Internet Archive catalogue records for *The Fable of the Bees*

#	Creator	Title	Date
1	Shelley, Mary Wollstonecraft, 1797-1851	Frankenstein	1993
2	Mary Wollstonecraft Shelley	Frankenstein: or, The modern Prometheus	–
3	Shelley, Mary Wollstonecraft, 1797-1851 Wordsworth Collection	Frankenstein; or, The modern Prometheus	1869
4	Mary Shelley Kenneth Branagh	Frankenstein: or, The modern Prometheus	1823
5	Mary Wollstonecraft Shelley	Frankenstein, Or, The Modern Prometheus	1869
6	Mary Wollstonecraft Shelley	Frankenstein: Or, The Modern Prometheus	1891
7	Mary Wollstonecraft Shelley	Frankenstein: Or, The Modern Prometheus	1888
8	Shelley, Mary Wollstonecraft, 1797-1851	Frankenstein, or, The modern Prometheus	1831

Table 3: Internet Archive catalogue records for *Frankenstein*

[3] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. *Computer Networks*, 29(8-13):1157–1166, 1997.

[4] S. Feng and R. Manmatha. A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. In *JCDL*, pages 109–118, 2006.

[5] N. Heintze. Scalable document fingerprinting. In *USENIX Workshop on Electronic Commerce*, 1996.

[6] T. C. Hoad and J. Zobel. Methods for identifying versioned and plagiarized documents. *JASIST*, 54(3):203–215, 2003.

[7] J. W. Hunt and M. D. McIlroy. An algorithm for differential file comparison. Technical Report CSTR 41, Bell Laboratories, Murray Hill, NJ, 1976.

[8] K. Krstovski and D. A. Smith. A minimally supervised approach for detecting and ranking document translation pairs. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 207–216, 2011.

[9] U. Manber. Finding similar files in a large file system. In *USENIX Winter 1994 Tech. Conf*, pages 1–10, 1994.

[10] S. Schleimer, D. Wilkerson, and A. Aiken. Winnowing: local algorithms for document fingerprinting. In *ACM SIGMOD conference*, pages 76–85, 2003.

[11] N. Shivakumar and H. Garcia-Molina. Scam: A copy detection mechanism for digital documents. In *Ann. Conf. on the Theory and Practice of Digital Libraries*, 1995.

[12] N. Shivakumar and H. Garcia-Molina. Finding near-replicas of documents on the web. In *Intl. Workshop on the World Wide Web and Databases*, 1999.

[13] I. Z. Yalniz, E. F. Can, and R. Manmatha. Partial duplicate detection for large book collections. In *Proc. of CIKM*, 2011.