University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3
HR-10000 Zagreb, Croatia

Zagreb, June 19th, 2015

Point-to-point responses to Reviewers

We would like to thank the Editor and the Reviewers for their time and effort. The comments we received have been invaluable in making the paper more understandable. In the rest of the document, we individually address each of the reviewers' comments.

Reviewer #1:

1. Please make your data explicit; publishing it (or a virtual collection) in addition to the textual description would help a lot.

Description and links for the datasets generated and used during the evaluation are publicly available at: http://ccl.fer.hr/ds/2015/readme.html. Additionally, the link is also added to the paper in form of footnote on page 1.

2. I am not convinced that the method of clustering chosen adequately finds duplicates in a realistic setting. I would expect the following to hold in reality: 1) you don't know how many of the books (i.e. clusters) there are 2) There is not a fixed number of slightly different versions per book

Thank you for this comment and we agree with the remark. We have updated the paper with more articulated description of what was our primary intention regarding clustering method we used in the first part of the evaluation. More specifically, we divided *Section 5. Experimental results* into two subsections. In the first subsection, we use k-means clustering only to confirm, in a very simple and intuitive way, that coderivative book identifiers group (or "gravitate") well around *known* canonical book identifiers. Thus, clustering on a dataset with fixed number of books and fixed number of book versions is used only for initial, basic feasibility evaluation on the *smaller* dataset. If this baseline clustering did not provide satisfactory results, there would be no point in going further with the proposed approach for similarity-preserving book identifiers. Nonetheless, clustering using some other method on real-world dataset or calculating approximate number of potential clusters in advance are interesting research directions and we would certainly like to include them in the future work.

Since our goal is not to cluster books based on content, but to identify and recommend list of similar books (sorted by identifier distance), for actual retrieval of similar books we have used (efficient, bucket-based) similarity queries, not clustering. In the second subsection of Section 5., similarity queries were evaluated on the larger dataset that resembles more realistic setting – i.e. number of book coderivatives is *not* fixed. Additionally, total number of books is not a (direct) issue for similarity queries.

3. A distance of 10 (out of 128 bits) would allow for quite some collisions in larger collections.

Optimal harmonic mean (F1), presented on Fig. 4. is indeed around distance of 7-10 bits. For comparison, another work, which we reference (Manku et al., 2007), efficiently uses 64 bit fingerprints and optimal distance for their dataset of 8 billion web pages is 3 bits. These results suggest that optimal distance for similarity threshold depends on artifact structure and content, as well as collection size.

As we have noted in the Conclusion section, future explorations will include evaluation of presented approach on real-world dataset. If collision becomes an issue, then book identifier length can be increased (with some performance penalty, as suggested in the paper) in order to increase optimal similarity threshold.

Reviewer #2:

1. My main overall comment is on the overall feasibility and practicality of your approach as a "digital book identifier". How could such a book identifier be established as a central service provided by a supra-national organization? Who determines what the baseline digital book is, is such a notion needed at all? Maybe the one and only reference of a specific book cannot exist and all digital books must be considered variations?

We have added paragraph in the Conclusion section that explains our intentions regarding described book identifier as part of future infrastructure for digital libraries. Our proposed solution consists of some kind of composite identifier where simhash fingerprint will be one of many components. Other components could be derived from metadata (such as ISBN) or even some other information derived from content. In addition to this composite identifier, imagined future infrastructure could utilize peer-to-peer distributed heterogeneous network where these identifiers are shared and derived using distributed consensus algorithms. Other approach could include institutional support to provide some form of centralized service (much like basedata.org). Thus, notion of baseline book is not needed. We fantasize about distributed network of authorities who use, for example block-chain, to infer baseline book. However, composite identifier could be useful in case institutions decide to implement centralized supra-national service. We are willing and ready to join the process, but the process is also political and requires institutional negotiation.

Reviewer #3:

1**. The related work is a little too narrow, as I see it. It does not relate to the domains of blocking and record linkage. I propose that the authors check the following publication for related work: http://dx.doi.org/10.1109/TKDE.2011.127**

Thank you for the related work suggestion. We have added mentioned reference to the Section 2. and described how blocking methods for record linkage relate to our work. To be more specific, blocking methods for record linkage operate on metadata (record) level. However, for this preliminary study we explicitly and consciously decided not to embed metadata into simhash fingerprint and rely only on book content. In the future work, we plan to address this issue with some kind of composite identifier that will include metadata in some form (in addition to existing contextual information). Once composite identifier is available, record linkage could be utilized for identifying coderivative books. We have updated future work subsection in the Conclusion with information about mentioned *composite identifier*. Additionally, our bucket-based similarity queries, as described in Section 5., can be considered as implicit blocking technique for book identifiers.

2. I would expect even some early comparison with other methods or a baseline, to indicate whether there exists any added value in the method. E.g., why not use (truncated) suffix trees or some other bucket-based method? What is the level of gain in performance/space/accuracy?

We did not provide comparison with baseline algorithms because, in our opinion, a comparison would be unfair. Simhash is the only method we identified that reduces high dimensionality inputs to a couple of bytes of data without TF-IDF on the input text. To best of our knowledge, all other methods use much more data to represent single book and that makes them more efficient in terms of accuracy. For example, baseline bag-of-words is superior to simhash. Additionally, in our distributed setting it is not possible to make use of TF-IDF analysis, as outlined in the Section 4., and other approaches usually have access to the whole collection and use IDF. Thus, we decided to sacrifice accuracy for space and practicality – simhash fingerprints are very small (128 bits was good enough for our dataset), can be easily shared and efficiently compared – much like ISBN numbers, with the benefit of embedded contextual information.

3. I feel that the figures are a little bloated, offering too much information. If possible, I would propose breaking them down (again, I am aware of the space limitations).

Figures were dense due to space limitations of the short paper. However, we managed to generate and include new graph for the embedded figure (which used to be Figure 2). Instead of single multi-graph figure there are now two figures that present clustering accuracy (Fig. 2) and fingerprint generation execution time (Fig. 3), respectively.

4. I understand the limitations of a short paper, but I would expect a more theoretic approach to the expected error of the fingerprint. Then, the evaluation would be useful to support the theoretic estimates. Otherwise, there is no clear guarantee to the performance of the system.

Since we were primarily interested in the practical usage of the described book fingerprints, we carried out experiments and provided empirical standard measurements from the information retrieval to evaluate accuracy. Specifically, precision and recall were calculated to assess accuracy of the fingerprint as a book identifier. More theoretic approach is out of scope for this preliminary study in the form of short paper. However, we agree that more formal analysis of the expected error (for example, collision rate or probability of false positive) is needed and it will be addressed in the future work.

5. In the evaluation, it was not very clear to me whether the results are based on the bucket-based search for matches or not. Furthermore, recall, precision and F1 are not trivial when evaluating clustering. The evaluation, thus, needs an improved explanation to indicate the actual measurement.

Following your remarks, text describing the experimental results (presented in the Section 5.) was restructured to clearly define evaluation methodology. First section describes initial feasibility experiments using clustering on the smaller dataset. Standard accuracy measures for clustering were calculated by following academic best practices – by comparing gold cluster membership known in advance (since dataset is synthesized) with cluster membership derived using k-means clustering algorithm. We agree that the metrics are not ideal, but our opinion is that they illustrate well selected aspects of the proposed method. Second part of the Section 5. describes experiments on the larger dataset using similarity queries (without clustering). Additionally, bucket-based approach is implemented to increase performance by following the locality-sensitive hashing approach described in (Rajaraman and Ullman, 2011). Again, precision, recall and F1 are calculated by following standard information retrieval formulas for the mentioned measurements, as described in the paper.