Managing Large Document Collections: The Case of Duplicates

Kyle Williams[1], C. Lee Giles[1], 2
Information Sciences and Technology[1], Computer Science and Engineering[2]
The Pennsylvania State University
University Park, PA, 16802, USA
kwilliams@psu.edu, giles@ist.psu.edu

Large document collections often contain similar or near duplicate documents. This is especially true in the case of document collections that are populated by automatic methods, such as search engines or digital libraries. The detection and potential removal of these duplicate documents is desirable for a number of reasons, such as reducing unnessary storage and computation costs, and providing users with uncluttered search results.

This study describes an investigation into the existence of near duplicates in CiteSeer$^{\mathcal{X}}$, a digital library that contains several million documents of an academic nature, which were automatically found on the Web and added to the digital library via automatic methods.

Using state of the art duplicate detection methods on a random sample of documents in the CiteSeerX collection, it was estimated that between 1 and 2% of the documents were near duplicates of each other. Furthermore, with the rise of the open access movement, it is expected that this number will rise significantly as more documents become freely available online, thus demonstrating the importance of being able to efficiently identify and tag duplicate documents.

This research also seeks to address the ranking of near duplicate documents. That is, once near duplicate documents have been identified, automatically determining which versions are the most useful, based on criteria such as document source, publication venue, etc., and then taking action on documents based on their ranking. Possible actions might include the disposal of all but one document, or the merging of near duplicates into a single digital object.

Lastly, the goal of this research is to generalize approaches for inspecting document contents, ranking them, and taking action on them based on their rankings and, in doing so, develop methods for better managing document collections.