Radical Transparency on the Web

Terrance Goan
Stottler Henke Associates, Inc.
Seattle, USA
email: goan@stottlerhenke.com

***Abstract*—Web users face massive and ongoing challenges in ascertaining the legitimacy and originality of information they discover. Intellectual property is commonly misappropriated; false news is propagated virally; and the original source of information is typically very difficult to determine. Recent technological advances have opened the door to supporting real-time transparency for all web content. In this paper, we outline a way forward, and open a discussion of the potential impact of radical transparency on the Web.**

Keywords- plagiarism detection; web search; intellectual property; source discovery; fact checking.

I. INTRODUCTION

As much as the Web is a resource for valuable and legitimate information and services, it has also become increasingly riddled with copyright violations, urban legends, rumors, fraud, and misinformation. Further, the vast scale and scope of the Web makes it ungovernable by any centralized authority. The only means of combatting this challenge is by empowering Web users by revealing evidence of credibility and sourcing. There are, of course, individuals and organizations that seek to fact-check hoaxes and scams, but the processes of source-discovery and fact-checking are laborious, and the products of these investigations typically reach only a small percentage of those who could benefit.

What is needed now is a complement to traditional Web indices—one that makes it easier for users to follow information "bread crumbs" back to original source materials so that they can assess for themselves validity and originality. The technical solution may appear to the end-user as a plagiarism detection system that can highlight all passages on a Web page that share a common origin with one or more other Web pages. Such a solution would not only reveal potential illegal reuse and fraud, but also could serve as a foundation for a new form of Web navigation that allows users to navigate amongst closely related materials that have not been explicitly hyperlinked (e.g., news stories that share significant quotations). This radical form of transparency would fundamentally change the way content is created and consumed on the Web.

In Section 2 we present related research, and in Section 3 we describe a key technology that could pave the way forward. In Section 4 we then describe some initial steps that we have taken. In Sections 5 and 6 we then discuss the potential impacts of radical transparency on the Web and our conclusions.

II. RELATED WORK

The proposed approach to Web transparency is most closely related to research conducted in the areas of plagiarism detection and author identification. Over the past decade, numerous useful tools have been developed to detect plagiarism [3]. State-of-the-art plagiarism detection systems rely on matching similarities between documents, and thus the effectiveness of these tools is limited by the scope and characteristics of the text collections they index.

While details vary, contributions to the annual Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN) competition [4] follow a two-stage strategy that was originally proposed by Stein et al. [6]. First, the contents of a suspicious document are analyzed to generate queries that may return potential source documents. Each of these candidates is then compared with the suspicious document to identify matching passages.

What is important to note is that while these methods may sufficiently scale to the challenge of detecting plagiarism in college essays, they are wholly inadequate for application at Web-scale [5][7]. The primary challenges in applying existing methods are the costs of finding and retrieving potential matches and in doing pairwise comparisons. However, new approaches to text indexing and comparison are available that could be implemented at Web-scale. In the next section we outline one of these possible approaches along with evidence of its sufficiency.

III. A PATH TO WEB TRANSPARENCY

In 2011, Mansour et al. published a paper describing a parallelizable string indexing scheme that can index the human genome in just 19 minutes on an ordinary desktop computer [2]. Since then, we have witnessed continued progress in the rapid indexing, search, and analysis of huge text sequences (e.g., string similarity calculations) (e.g., [8]). While this research originated with a focus on bioinformatics, the potential applications for these new algorithms are far-reaching and include the potential for full-text comparisons on the Web.

The conceptual foundation for Mansour's work is the suffix tree [1]. A suffix tree is a data structure that indexes all possible suffixes of a string (e.g., a Web page). Both the construction, and querying, of a suffix tree can be done in linear time (in the length of the input). Further, the suffix tree

 ISBN: 978-1-61208-636-1

allows us to search a very large corpus (e.g., the Web) for *all* occurrences of a String S in $O(|S|)$. This ability to conduct full text comparisons against an arbitrarily large corpus of text in time that grows linearly with the size of the query string opens the door for the Web transparency we propose. Finally, Mansour's approach allows for parallel disk-based processing on commodity hardware.

IV. INITIAL STEPS

We recently constructed a limited form of this concept. We implemented Mansour's algorithm and applied it to the challenge of detecting the intentional or unintentional release of sensitive intellectual property from an organization. Essentially, we built an index of proprietary text and then conducted searches against that index with any suspect material. We found that using four commodity desktop machines, our initial implementation could index 1TB of data in four days. Even with this non-optimized implementation, we believe indexing petabytes of Web data in this manner is now feasible.

Searching for matches is more challenging than indexing since identifying *all* overlaps between a query document and a stored corpus typically would require many restarts. The process involves finding a match starting at position X in the query document, then reinitiating the search at position X+1. There are, however, numerous optimizations that are possible, including ignoring short matches and extremely common matches which are likely of little import. In our initial tests, searching 1,000 documents on a single machine took under two seconds.

This approach of course has limitations. Plagiarism detection, for instance, can be complicated by careful obfuscation that would limit the number of longer matches. However, the scale of the index would act to counter this force as numerous variants of the original would also be indexed. Further, many of the useful applications of these indexes would not involve intentional obfuscation.

V. IMPACTS OF WEB TRANSPARENCY

It is worth considering the impact of easily accessible transparency on the Web. While, on balance, the effect of transparency would be positive, there remains a possibility of unfortunate side-effects. Consider the following:

- Intellectual property protection. Given a Web-scale index, unauthorized content reuse would be easily detected. This would likely have the biggest impact on the research community and in journalism, where the material is made public and plagiarism can have dire professional ramifications. Interestingly, the transparency we envision may spark improvements in automated plagiarism approaches where words are strategically replaced with synonyms to avoid detection.
- Implicit reference chaining. Many articles, Web pages, and academic papers legitimately include quotations and share bibliographic references. By revealing these text overlaps, Web browsers could enable users to navigate between related documents without the need for explicit hyperlinks.
- Fact checking and urban legend detection. The proposed approach to Web transparency could enable automated source identification through a combination of text overlaps and timestamps. This would allow information consumers to quickly ascertain for themselves the likely veracity of published reports.
- Change detection. Many legal documents are created from templates or through the reuse of material found on the Web. Examples include usage agreements, licenses, non-disclosure agreements, and the fine print associated with financial instruments, such as credit cards. The proposed web indexing scheme would present the opportunity to highlight differences between a legal document and others that were found online. This in turn would allow users to detect modifications of terms that warrant their attention.

VI. CONCLUSIONS

In this paper, we have proposed the utilization of highly scalable string indexing and search methods to provide an extreme form of transparency on the Web. The implications of revealing shared text during everyday use of the Web are worthy of consideration, particularly given the lack of any governing authority that can act to thwart the growing threat posed by plagiarism, fake news, and government-sponsored propaganda initiatives on the Web. As a next step we anticipate a subject-specific Web-index that will allow us to further explore issues of scalability and utility.

REFERENCES

[1] D. Gusfield, Algorithms on strings, trees and sequences: computer science and computational biology. Cambridge university press, 1997.

[2] E. Mansour, A. Allam, S. Skiadopoulos, and P. Kalnis, "ERA: efficient serial and parallel suffix tree construction for very long strings," Proc. of the VLDB Endowment, vol. 5, no. 1, pp. 49-60, 2011.

[3] V. Martins, D. Fonte, P. Henriques, and D. da Cruz, "Plagiarism detection: A tool survey and comparison," OASIcs-OpenAccess Series in Informatics, vol. 38, pp. 43-58, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2014.

[4] Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN) Workshops http://pan.webis.de/, [retrieved March, 2018]

[5] M. Sanchez-Perez, G. Sidorov, and A. Gelbukh, "A winning approach to text alignment for text reuse detection at PAN 2014," In CLEF (Working Notes), Sep. 2014, pp. 1004-1011.

[6] B. Stein, S. zu Eissen, and M. Potthast, "Strategies for retrieving plagiarized documents," Proc. of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Jul. 2007, pp. 825-826.

[7] N. Unger, S. Thandra, and I. Goldberg, "Elxa: scalable privacy-preserving plagiarism detection," Proc. of the 2016 ACM Workshop on Privacy in the Electronic Society, Oct. 2016, pp. 153-164.

[8] M. Yu, J. Wang, G. Li, Y. Zhang, D. Deng, and J. Feng, "A unified framework for string similarity search with edit-distance constraint," The VLDB Journal, vol 26, no. 2, Apr. 2017, pp. 249-274.

 ISBN: 978-1-61208-636-1