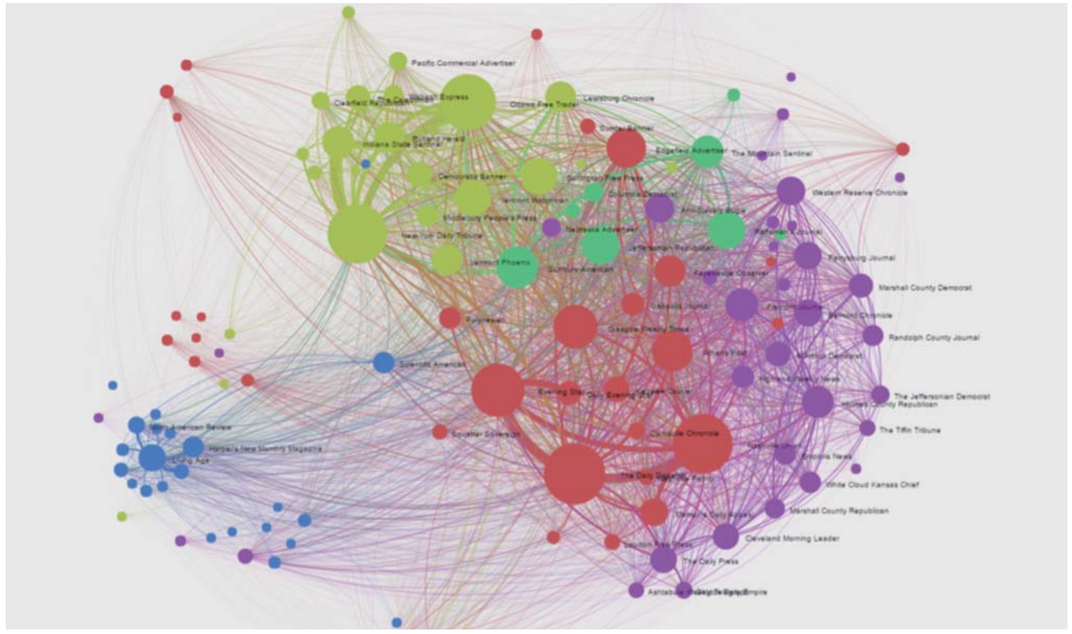
Content Reuse for Text and Multimedia Documents



A network showing text reuse among pre-Civil-War newspapers from the Library of U.S. Congress's Chronicling America newspaper archive [1].

Chuan XIAO

Designated Assistant Professor of the Young Leaders Cultivation Program
Graduate School of Information Science / Institute of Advanced Research, Nagoya University
E-mail: chuanx@nagoya-u.jp

Content reuse is a common practice in an era when electronic documents prevail. For example, when composing presentation slides, people often create new files by reusing the materials in existing slides instead of starting from scratch. Investigating content reuse in electronic documents may assist in a wide range of applications, including plagiarism detection, near-duplicate Web page removal, text summary generation and presentation slide composition. This study covers two tasks: content reuse detection in text documents and content reuse for presentation slide composition. We addressed several important technical issues and proposed effective methods with high efficiency. On the basis of the proposed methods, prototype systems with user-friendly interfaces were developed for practical use.

INTRODUCTION

One of the main issues accompanying the growing popularity of electronic documents is the existence of reused contents. For example, reused text may exist in academic papers, dissertations, etc. People may plagiarize others' work by copying text segments from various sources and making a few modifications to avoid detection. Another example is that when composing presentation slides, 97% of people compose presentation slides by reusing existing materials rather than starting from scratch [2]. One of the primary reasons for such reuse is to repurpose the contents of existing slides for different audiences, events, formats, etc. In this study, we looked at content reuse for two common types of documents: text and multimedia, and focused on the following two tasks: (1) content reuse detection in text documents, and (2) content reuse for presentation slide composition.

Content reuse detection in text documents

Although there have been many existing solutions to the detection of reused contents, e.g., [3, 4], they can be easily fooled by minor modifications (Figure 1), such as reorganizing sentences, replacing words with synonyms, etc. Consequently, reused contents with modifications are often missed by these methods, and hence the quality of detection is not satisfactory.

Seeing the limitations of prior solutions, we propose a new approach by detecting similar text segments. Our method is not only insusceptible to word order or sentence structure, but is also tolerant of a small number of differences. Hence text reuse can be detected in spite of minor modifications. We evaluated our method for the purpose of plagiarism detection. The experiment results show that our method detects more than 90% of plagiarism in benchmark datasets, significantly outperforming existing methods, which detect up to 70% of plagiarism. In addition, our method is able to detect templates and boilerplates, which are commonly used in newspapers and Web pages.

...Now, then, if we are interested in the problem of democracy, we have two courses open it of finding out why it is that the laws we already have are not enforced, and are PRACTICAL difficulties which confront the Negro choice in many parts of the South. The point I wish to make is that legally the Negro is the political equality of the white man as well. The doctrine laid down in the Fifteenth is neither for some both white and colored to poorer category of white man, nor for the women of now expired by limitation. Let me then lay down this general: Nowhere in the South today is neither for the factories and shops, nor for the well within the principle laid down by the Fifteenth Amendment the Negro cut off LEGALLY, as a Negro, from the ballot...

...Legally, today, any Negro who can meet the comparatively slight requirements as to education, or property, or both, can cast his ballot on a basis of equality with the white man. I have emphasized the word legally, for I know the PRACTICAL difficulties which confront the Negro votes in many parts of the South. The point I wish to make is that legally the Negro is essentially the political equal of the white man; but that practically, in the enforcement of the law, the legislative ideal is still pegged out far beyond the actual performance. Now, the, if we are interested in the problem of democracy, we have two courses open to us. We may think the laws are unjust to the Negro, and incidentally to the 'poor white' man as well. If we do, we have a perfect...

Figure 1. Text reuse (left, in yellow) by copying from another document (right) and making minor modifications (in red).

Besides improving the result quality, we also developed efficient indexing and optimization techniques to speed up our method for the purpose of handling large volumes of text data. The experimental evaluation shows that our method equipped with these techniques is up to 12 times faster than alternative solutions.

Content reuse for presentation slide composition

We designed a platform to help users compose slides by reusing existing materials. The platform consists of three modules: (1) slide element search, (2) slide management, and (3) slide auto-generation.

For slide element search, due to the existence of different types of elements in presentation slides, e.g., textual elements such as titles and sentences, and graphical elements such as images, charts, and diagrams, we develop a series of techniques to handle the variety of presentation slide elements. For textual elements, users can input keywords or sentences to search, like when using a Web search engine. For graphical elements, users can select an image on their disks or drag a rectangle area in a slide as a query (Figure 2). Then the module efficiently searches in their presentation files and shows relevant materials. It also supports the feature of approximate search so that users do not have to remember the materials exactly.

Slide management is a module with which users can manage their presentation files in terms of their relationships, e.g., multiple versions, summaries, etc. The files are visualized in a network (Figure 3). Two files are

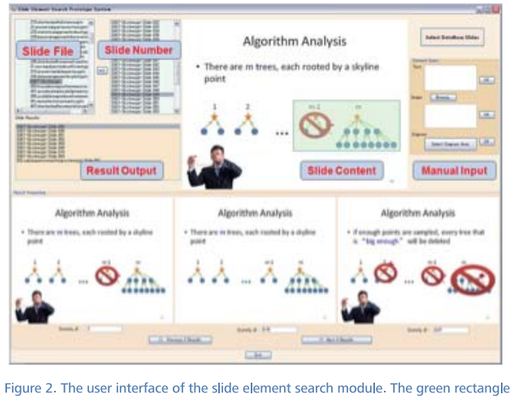


Figure 2. The user interface of the slide element search module. The green rectangle contains the query diagram. The results are shown at the bottom.

Figure 3. Two files are connected if one file reuses material from the other. The grey square on the left indicates a summary. The green circle on the right indicates multiple versions of the same presentation.

connected if common materials are identified. Users can easily find out which files reuse materials from others and which elements are reused. In addition, users can drill down the reused elements to see their timelines, i.e., which file they originate from and in which files their content has been reused and when.

The slide auto-generation module saves users from making slides page by page. First, the users specify the titles for each slide. They may input a title by the keyboard or choose from a range of common titles, such as "related work," "experiments," or "conclusions" for academic presentations. They then select elements using the slide element search module, and assign these elements to the pages. Finally, they adjust the slide layouts, e.g., the positions of the text and images. Presentation slides are automatically generated after these steps. A set of example slides generated by this module is shown in Figure 4.

On the basis of the above modules, we design prototype systems with user-friendly interfaces that can easily be used in a company with common a slide composition tool such as Microsoft PowerPoint.

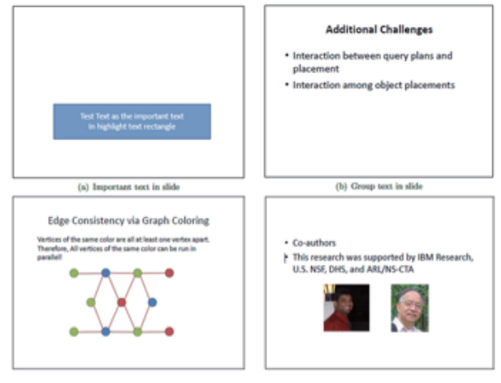


Figure 4. Example slides generated by the slide auto-generation system.

Acknowledgements

I would like to thank the co-authors of this study: Jie Zhang, Toyohide Watanabe, Yoshiharu Ishikawa, Pei Wang, Jianbin Qin, Xiaoyang Zhang and Wei Wang.

References

(1) R. Cordell, D. A. Smith, et al. The Viral Texts Project. Northeastern University.
(2) M. Sharmin, L. Bergman, J. Lu, and R.B. Konuru, "On slide-based contextual cues for presentation reuse." International Conference on Intelligent User Interfaces, 129–138 (2012).
(3) J. Seo and W. B. Croft. "Local text reuse detection." ACM SIGIR Conference, 571–578 (2008).
(4) Y. Sun, J. Qin, and W. Wang. "Near duplicate text detection using frequency-biased signatures." International Conference on Web Information Systems Engineering, 277–291 (2013).