R·I·T Printing Industry Center

October 2010 eReview

Center Spotlight

Print in the Mix

Print in the Mix is "a unique site demonstrating the role of print as a viable information medium in the marketing mix." This **free** resource is published by the Printing Industry Center.

Sample Fast Fact:
North American consumers in the 18-34 year-old demographic prefer to learn about marketing offers via postal mail and newspapers rather than online sources such as social media platforms, according to survey research from ICOM, a division of Epsilon Targeting.
Read the full fast fact here.

Have you visited **Print in the Mix** yet? Find out how this site can help you 'make the case' for print!

printinthemix.rit.edu
Funded by The Print Council

Transclusion of Dynamic Document Fragments - Part B

This month's research summary is the completion of an RIT School of Print Media graduate thesis entitled *Transclusion of Document Fragments from Dynamic Text*, by Manu Choudhury. Last month, we began with a summary of the introduction and literature review (see Part A). This month, we will finish with a look at the research objective, methodology, and research results.

Research Objective

Develop an algorithm based on encryption of text that can be used to transclude text from dynamically changing source pages, such that the transcluded document would accurately reflect the changes made in the source document.

Methodology

Design of the Algorithm

An algorithm was required that would ensure that changes in the source document would automatically and accurately reflect in the transcluded user document.

The basic requirement for transclusion is that the user document should not include the transcluded text, but only contain some form of reference to the document fragments of the source page.

Overview of the Algorithm

The algorithm was based on encryption of the transcluded text and its positional relationship. When retrieving the transcluded text from the source document the source document was encrypted using the

exact same encryption technique, and then the encrypted transcluded text was compared with the encryption of the source document.

The fragment of the source document whose encryption matches closest to the encrypted transcluded text was returned as the transcluded text.

For a step-by-step description of the algorithm, please see** the complete thesis**.

Analysis of the Algorithm

Since the accuracy of the returned transcluded text depended on the original source file, the portion of the text transcluded from the source file, the amount, number, and nature of changes in the source file, an automated script was written to test the accuracy of the algorithm. The steps that were taken are as follows:

1. A collection of 839 documents was downloaded from the Web: http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes/20_newsgroups.tar.gz

2. The script selected one file at a time, and treated it as a source file.

3. It then randomly selected a portion of that file as the transcluded text.

4. A random number between zero and two was chosen. This number denoted the number of additions to the original source file to be performed.

5. For each addition a random file was selected, and a random portion of this new file was picked and inserted at a random point in the original source file. The source file, with this addition, was treated as a new source file. Depending on the number of additions to be performed, new text was added onto this file.

6. A random number between zero and two was chosen again. This number denoted the number of deletions to the source file.

7. For each deletion a random point in the source file was

of the Printing Industry Center at RIT for registered Affiliate companies. Articles are also published in the quarterly printed publication *PrintReview.*

Forward this eReview.

Subscriptions

You are receiving this newsletter because you registered as an Affiliate of the Printing Industry Center.

Update your profile.

Unsubscribe from this list.

Contact the Center

Director:
Patricia Sorce

Communications Coordinator:
Ashley Walker
(Web site, publications, general info)

Mailing Address:
RIT Printing Industry Center
College of Imaging Arts & Sciences
Rochester Institute of Technology
55 Lomb Memorial Dr
Rochester, NY 14623

Ph: 585-475-2733
Fax: 585-475-7279
Web: http://print.rit.edu
Email: printing@rit.edu
Twitter: RITprintcenter

selected, and a portion of the text was deleted. Portions of text were further deleted from it based on the number of deletions.

8. Since the script knew the changes that went through the original source file and the original transcluded text, the script calculated the expected transcluded text.

9. The new source file and the previous transcluded text was provided to the algorithm; and the algorithm returned the most probable transcluded text that reflected the changes performed on the source file.

10. This predicted transcluded text was then compared to the expected transcluded text, and thus, the accuracy of the algorithm for that situation was recorded by dividing the length of Longest Common Substring between the expected transcluded text and the predicted transcluded text, by the length of expected transcluded text.

11. The script recorded the percentage of change to the source document, number of deletions, number of additions, length of the source document, length of the original transcluded text, and the accuracy of the algorithm for that situation.

12. The script did the same for all documents.

13. The script was executed several times for all of the documents, just to ensure that the accuracy of the algorithm has been recorded for most different sizes of the source document, different sizes on the portions transcluded from the source documents, and different amounts, numbers, and the nature of changes in the source document.

Results

Percentage Change to Source Documents

Before the accuracy of the algorithm is presented it is important to understand the percentage change to the source documents. The researcher hypothesized that the greater the changes, the less effective the algorithm.

Figure 1: Histogram of percentage change of the source

About the Center

Dedicated to the study of major business environment influences in the printing industry precipitated by new technologies and societal changes, the Printing Industry Center at RIT addresses the concerns of the printing industry through educational outreach and research initiatives.

Support for the Center comes from:

Sloan Foundation
Rochester Institute of Technology

Adobe
Avery Dennison
Democrat and Chronicle
Hewlett-Packard
NewPage Corporation
NPES
Scripps Howard Foundation
VIGC
Xerox Corporation

click to view image larger

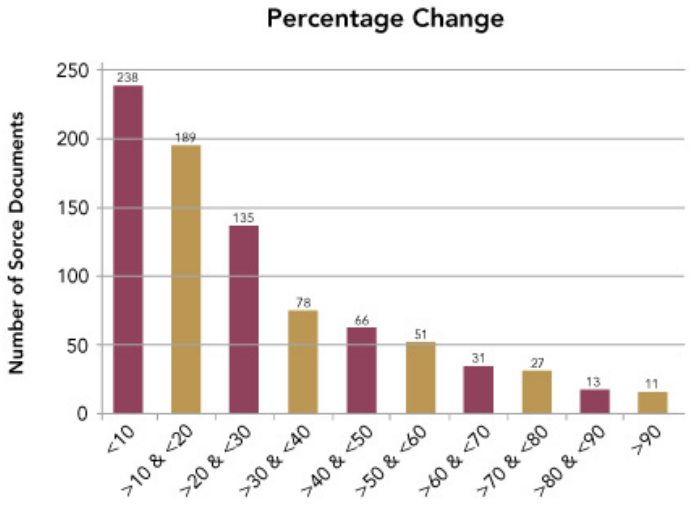

Figure 1: Histogram of percentage change of the source documents

The percentage change was calculated using the following formula:

%Change = (Number of characters added + Number of characters deleted) / Total number of characters in the original source document

Average Accuracy vs. Weights Selected in the Formula to Calculate the Score

The score indicates the quality of the transcluded text that is returned after execution of the algorithm. The algorithm assigned a score to a document fragment based on its probability of being the transcluded text that reflected the change in source document.

Thus, the most important task for the algorithm was to find a document fragment that had the highest score.

For a discussion of the formula used to find the score, please see** the complete thesis**.

Accuracy of the Final Transcluded Text Compared to the Expected Outcome

The original source document, the original selection, and the changes performed on the source document were all recorded for each of the documents in the test set. Using all this information, the expected

transcluded text was calculated.

The predicted final transcluded text was obtained by the researcher's algorithm using the reference to the transclusion (this includes the location of source document and the encryption of the initial selected document fragment).

The predicted final transcluded text was compared with the expected transcluded text and the accuracy of the algorithm for that data set was computed.

The accuracy obtained for all of the documents in the test set was recorded. Figure 2 describes the overall accuracy of the algorithm irrespective of the nature of changes and the percent of change that each of the documents went through.

Figure 2: Histogram of average accuracy
click to view image larger

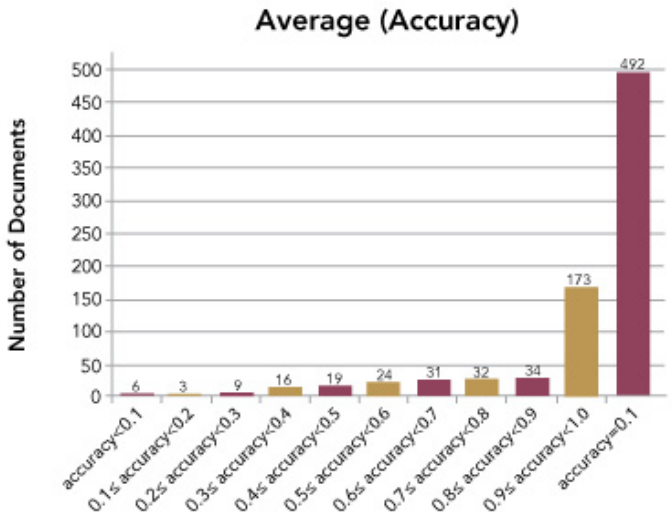

Figure 2: Histogram of average accuracy

The plot indicates that the predicted transcluded text returned from the algorithm matched exactly as the expected result for 492 of the 839 documents comprising the total test set. This indicates that for close to 60% of the test set, the algorithm returned the final transcluded text that matched the expected outcome exactly, i.e., the final transcluded text reflected all of the changes that were made to the original source document.

Figure 2 also shows that 173 times the algorithm returned transcluded text that matched the expected outcome at least 90%, but was not identical. This shows that almost 80% of the results matched the expected outcome to at least 90%.

Accuracy vs. Percent Change to Source Document

The accuracy of the algorithm depended on the percentage of change the original source document has gone through. Figure 3 shows how the algorithm performed with the varying change percentage that the source document had undergone.

Figure 3: Average accuracy vs. percent change to source document

click to view image larger

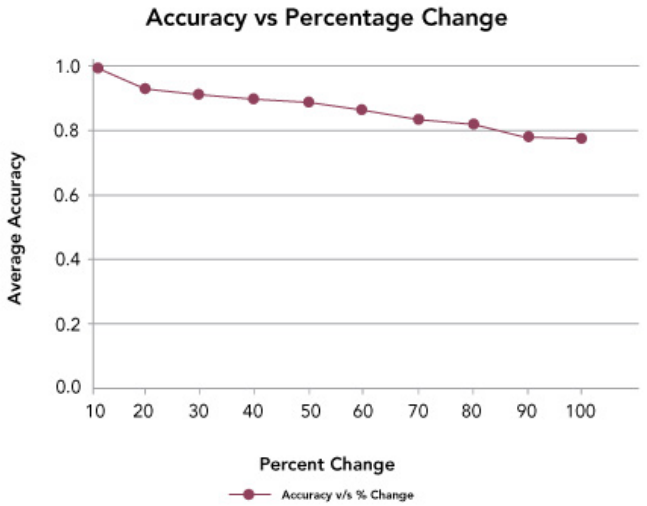

Figure 3: Average accuracy vs. percentage of change

In Figure 3, we can see that when the change was less than 10%, the average accuracy of the algorithm was almost 1, whereas when it was about 50%, the average accuracy of the algorithm was close to 0.9. Even when the source document was changed more than 100%, the average accuracy was still greater than 0.75.

It is important to note that the change could be anywhere in the source document, and not necessarily between the document fragment that was originally selected.

Accuracy vs. Nature of Change

Figure 4 displays the average accuracy of the algorithm depending on the nature and the number of changes that the document fragment went through. Even though the document fragment that was originally selected had not changed, it was possible that other portions of the original document could have changed. A maximum of four changes (two additions and two deletions) were carried on the original source document for the analysis of the algorithm.

Figure 4: Average accuracy vs. nature and number of changes
click to view image larger

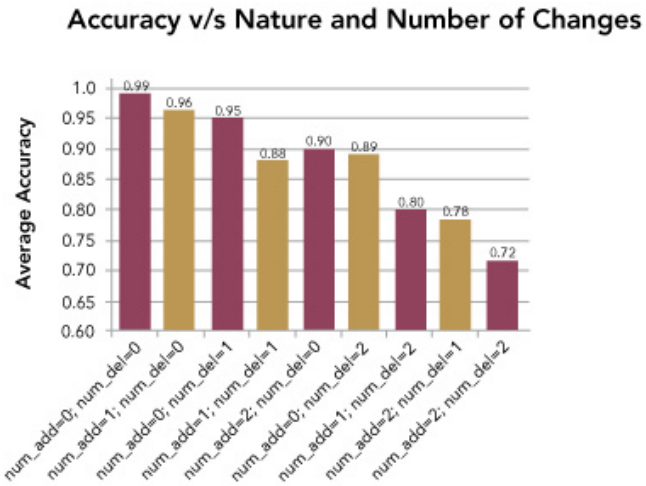

Figure 4: Histogram of average accuracy vs. nature and number of changes

Figure 4 indicates that the average accuracy of the transcluded text returned from the algorithm was 0.99 when number of additions and number of deletions to the initially- selected document fragment was equal to zero.

When the document fragment underwent only one change (either deletion or addition) the average accuracy was between 0.95 and 0.96. When the document fragment underwent more than one change the average accuracy gradually decreased, and finally came down to 0.72 when the document fragment underwent addition and deletions twice each.

Figure 4 also indicates that the nature of change (addition or deletion) had little effect on the accuracy.

Summary and Conclusion

Transclusion is a concept that allows users to reuse document fragments from different source pages, not by duplicating it, but by including a reference to the original work. Therefore, if the source changes, the change is automatically reflected in the final output. This ensures the reader that this new 'transcluded' document always has access to the newest information. Moreover, transclusion can be used for collaborative document creation, where every author writes his own part, and transcludes fragments that other authors have contributed.

Transclusion has been implemented by many researchers from around the globe using various technologies and tools, but most of their implementations made use of static references to the document fragments being transcluded. Thus, if the source page changed, their implementations might not be able to retrieve the document fragment that was originally selected.

The objective of this research was to find an algorithm that could be used to transclude text from dynamically changing source pages, such that the transcluded user document would accurately reflect the changes made in the source document. The algorithm explained in this thesis fulfills the required objective.

Overall, the accuracy of the algorithm was 0.92. For about 60% of the test set, the returned document fragment matched exactly to the expected outcome reflecting all the changes made to the original source document. Moreover, for more than 80% of the test set, the predicted result matched more than 90% to the expected outcome.

It was observed that the accuracy of the algorithm decreased with an increase in the percentage of change to the source document. This result is intuitive as well, since the more the source document changes, the harder it is to predict the final transcluded text, as some changes might be left out, or some extra portions of the document might be returned. The algorithm turned out to be quite forgiving, as even when the source document had changed 100%, the average accuracy still turned out to be more than 75%. It was also observed that the nature of the change (addition or deletion) had little effect on the accuracy.

In conclusion, this research can be used as a framework to transclude document fragments from dynamic source pages, while still ensuring that the changes made to the source page are reflected in the user document. In the researcher's algorithm the connection

between the user document and the source document is retained, and the user document always has access to the newest information, thereby allowing users to present their ideas effectively, and re-use them as a framework for establishing their own ideas.

Research Publications

To read about this research in detail, download the thesis from: https://ritdml.rit.edu/handle/1850/12419

Research publications of the Center are available at: http://print.rit.edu/research/index