

Activities and Findings:

Research and Education Activities:

A text collection such as a newswire archive or Web crawl typically contains a great deal of repeated information. Different authors may each present versions of a story or event, the same event may get presented in different ways for different audiences, and the facts of an event may get recapitulated each time it is presented. Sometimes such presentations have little in common with each other; at other times one may be a copy of the other with minor edits. Given a topic of interest, then, a sufficiently extensive archive may contain much of the history of the topic. In particular, it can plausibly be used to identify when particular ideas or statements originated. It can also be used to check facts or the sources of statements made about the topic. Our interest is in exploring whether we can identify alternative versions of the same information in order to reconstruct the information flow.

The extent to which passages of text are considered similar to each other can be thought of as falling somewhere on a *similarity spectrum*. At one end of this spectrum is identity; two documents that are the same as each other in every way clearly have the highest level of similarity possible. The other end of the spectrum is the standard task of information retrieval: two documents are a match if they are about the same information need. We are exploring the similarity spectrum in the context of analyzing information flow or reuse. The objective of this project is to develop methods for tracking and analysis of facts and concepts through a text corpus. In order to create such methods, we need to develop retrieval and representation methods that are appropriate for queries made up of long pieces of text, such as sentences or paragraphs.

This year we hired a number of new students and were able to make progress on the main task of the project as well as some related tasks. These are described below.

Finding Text Reuse on the Web. We have developed a novel technique that addresses the unique challenges of finding original sources of information on the Web. This research is based on a scenario where the user wants to find out where a statement on a Web page comes from, and how it might have changed over time. The new technique segments the statement into interesting “chunks”, searches the Web using those chunks, and then applies text reuse detection algorithms to the pool of documents that are retrieved. In addition, we examined two methods for reconstructing the “information flow”, which is a timeline showing when the statement (or a variant) was used and where.

Most of the previous work on text reuse detection has been focused on relatively homogeneous collections, usually including documents of the same type (e.g., news or blogs). On the other hand, web search returns documents of varying types (electronic newspaper sites, personal blogs, wikipedia articles, news aggregator sites, and so on), formats, sizes, writing styles and quality. Some of the returned pages contain only a short portion of text reuse (e.g., a news story headline with a link to the original story), while others aggregate complete news stories by a certain event or topic. Another salient property of the web is its size. Previously, text reuse detection methods were tested on collections of a much smaller scale than the web. Computationally expensive

techniques such as *sentence retrieval* cannot be efficiently applied in a general web search setting for the entire collection. We address this problem by applying these techniques for a small subset of documents retrieved by a query.

Due to the inherent heterogeneity of the information on the web, information extraction from web pages is a hard problem. For the task of text reuse, information extraction is interesting from several perspectives. It can be used for identifying the major actors involved in the statement or event of interest, geographic location and date and time of its occurrence. Similarly, information extraction techniques can be applied not only to the original statement itself, but to reports related to it as well, aiding in constructing the previously discussed information flow.

In this work, we focused on two applications of information extraction to text reuse: extraction and weighting of *key concepts* related to the event or statement of interest and dating of the statement itself and the web pages related to it. Statement and document dating are interesting in the web setting as pages usually have no pre-assigned dates, and, as we showed, timestamps returned by the web search engine can often be misleading. In addition, in long heterogeneous documents containing multiple entries (e.g., news reports, blog posts, headlines, etc.), different entries will have different dates associated with them.

An additional challenge that often arises with user-generated content (rather than centrally published content, like newswire collections) is determining source quality and trustworthiness. Link analysis using graph-based techniques such as PageRank and HITS has been shown to be successful for a number of tasks, including determining web page quality and determining web page importance for a topic, among many others. In this work we investigated how well link analysis is suited for the task of text reuse, and how it can be employed to leverage information about trustworthy and relevant sources of information.

Local Text Reuse Detection. Text reuse and duplication can occur for many reasons. Web collections, for example, contain many duplicate or near-duplicate versions of documents because the same information is stored in many different locations. *Local* text reuse, on the other hand, occurs when people borrow or plagiarize sentences, facts, or passages from various sources. The text that is reused may be modified and may be only a small part of the document that is being created.

Near-duplicate document detection has been a major focus of researchers because of the need for these techniques in Web search engines. These search engines handle enormous collections with a great number of duplicate documents. The duplicate documents make the system less efficient in that they consume considerable system resources. Further, users typically do not want to see redundant documents in search results. Many efficient and effective algorithms for near-duplicate document detection have been described in the literature.

Local text reuse detection requires different algorithms than have been developed for near-duplicate document detection. The reason for this is that, in the case of local text reuse, only a small part (or parts) of a document may have been taken from other sources. For example, state-of-art near-duplicate detection algorithms like the locality sensitive hash assume a transitive relation between documents. That is, if a document A is a near-duplicate of document B, which

is a near-duplicate of document C, then document A should be a near-duplicate of document C. A text reuse relationship based on parts of documents, however, violates this assumption, as shown in figure 1.

We developed algorithms for detecting local text reuse based on parts of documents. In this work, we expanded on the idea of local text reuse by introducing categories of reuse. These categories are the basis of our experimental evaluation. We also introduced a novel algorithm for local text reuse detection called *DCT fingerprinting*.

Discovering Key Concepts in Verbose Queries. Current search engines do not, in general, perform well with longer text passages or more verbose queries. One of the main issues in processing these longer pieces of text is identifying the key concepts that will have the most impact on effectiveness. In this work, we developed and evaluated a technique that uses query-dependent, corpus-dependent, and corpus-independent features for automatic extraction of key concepts from text passages or verbose queries. This method achieves higher accuracy in the identification of key concepts than standard weighting methods such as inverse document frequency. We also proposed a probabilistic model for integrating the weighted key concepts identified by our method into a query, and demonstrated that this integration significantly improves retrieval effectiveness for a large set of natural language description queries derived from TREC topics on several newswire and web collections.

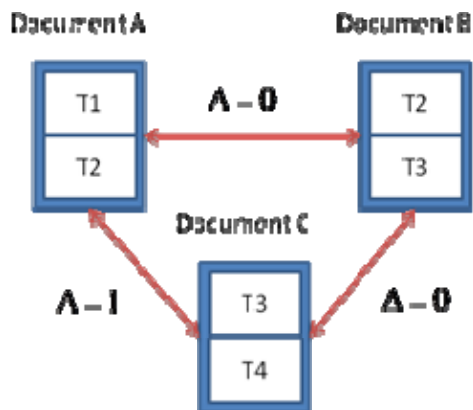


Figure 1. Text reuse by partial text of documents (Local text reuse), where Δ is an indicator function that is 1 if two documents have no text reuse relationship. A document A and a document B have a text reuse relationship by partial text T2. Document B and a document C have a relationship by partial text T3. But, Document A and C have no such relationship.

Search techniques for Blogs. As part of our work studying text reuse techniques for blogs, we have looked at how search techniques to identify relevant blog homepages. A blog homepage consists of many individual blog postings. Current blog search services focus on retrieving postings but there is also a need to identify relevant blog homepages. In this paper, we focus on finding the relevant blog homepages rather than the postings. Blog homepage search is similar to resource selection of distributed information retrieval in that the target is to find relevant collections. We introduced resource selection techniques for blog homepage search and compared the performance of each technique. Further, we proposed a factor describing the topic-

diversity of each blog homepage based on the properties of blog homepages. Our results showed that the appropriate combination of the resource selection techniques and the diversity factor can achieve significant improvements in the retrieval performance compared to our strong baseline.

Findings:

In this section, we focus on the major results related to text reuse. Results for the related projects can be found in the articles listed in the publications section.

Finding Text Reuse on the Web.

Table 1 summarizes the main result in this work. It shows that the methods we developed were effective at finding a pool of web pages containing text reuse (the IC method), and identifying the specific sentences that text reuse occurred in (the MX method).

(a) Document-Level Retrieval				(b) Sentence-Level Retrieval				
Category	Description	UQ	IC	Category	Description	QL	MX	DM
<i>C3</i>	% (Near) Duplicates	0.29	0.19	<i>C3</i>	(Near) Duplicates	0.30	0.30	0.29
<i>C2</i>	% Text Reuse	0.39	0.42	<i>C2</i>	Text Reuse	0.55	0.58	0.50
<i>C1</i>	% Topical Similarity	0.15	0.19	<i>C1</i>	Topical Similarity	0.13	0.10	0.18
<i>C0</i>	% Non-Relevant	0.17	0.20	<i>C0</i>	Non-Relevant	0.02	0.02	0.03
	Total Judged	373	500		Total Judged	500	500	500
	NDCG@10	0.441	0.464		NDCG@10	0.627*	0.633*	0.625*

Table 1: Performance comparison of the various retrieval methods. Document-level retrieval methods are presented in Table (a): UQ - unquoted query, IC - iteratively “chunked” query. Sentence-level retrieval methods are presented in Table (b): QL - query, MX - mixture models, DM - combination of dependency and mixture models. Tables present both the ratio of categories of text-reuse and the NDCG@10 for all methods. * marks statistical significant difference (two-tailed t-test, $p < 0:05$) between the retrieval methods in Tables (a) and (b).

Our experiments also showed that the methods we developed to predict dates were very effective. Figure 2 gives an example of a timeline constructed using these methods for documents that contained examples of text reuse.

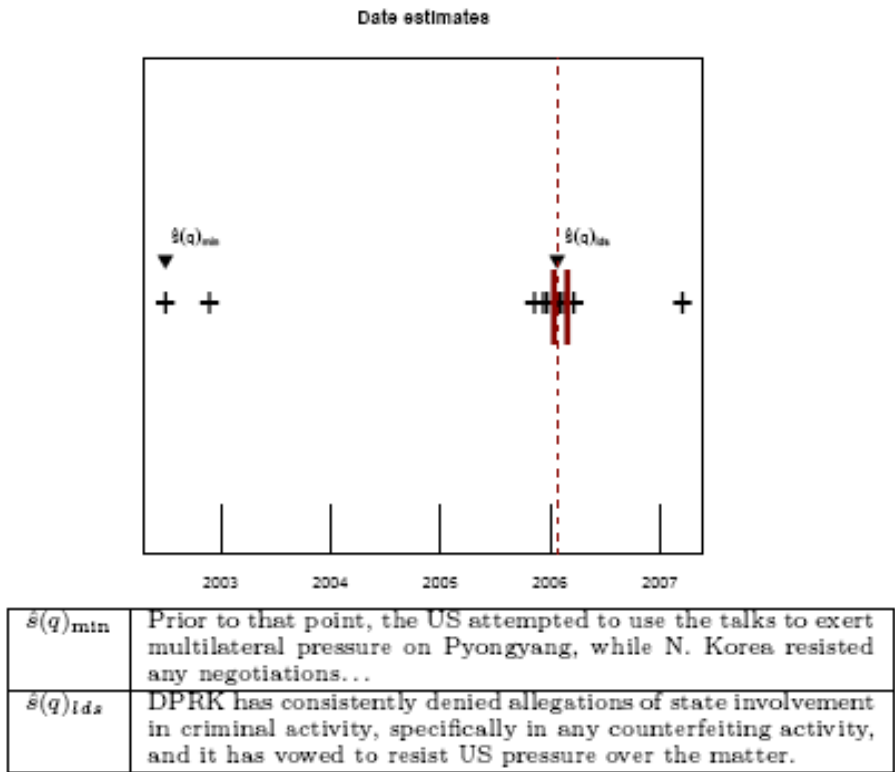


Figure 2: Two different date estimates for an event on a timeline constructed from the 50 top retrieved instances of the text reuse for query *The North Korean government has vehemently denied any hand in counterfeiting and has vowed to resist pressure from the United States over the matter*. The best estimation method is indicated by the two vertical bars. Original source date is indicated by a dashed vertical line.

Further details of this work can be found in:
 Bendersky, M. and Croft, W. B., "Finding Text Reuse on the Web", *Proceedings of the International Conference on Web Search and Data Mining (WSDM 2009)*, (2009).

Local Text Reuse Detection.

The graph in Figure 3 summarizes the results we obtained from our new local text reuse detection method (DCT fingerprinting) compared to a number of other duplicate detection methods. This shows that our method is as effective as the best duplicate detection method for finding text reuse and much more efficient.

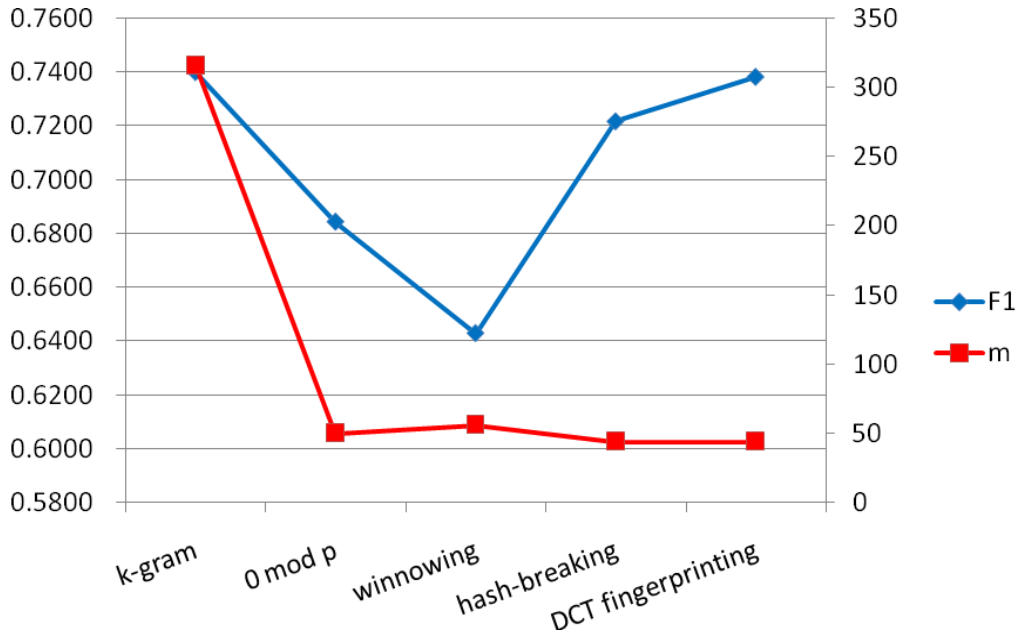


Figure 3. Local text reuse detection performance of fingerprinting techniques on a TREC newswire collection. ‘*F1*’ and ‘*m*’ represents the average of the *F1 effectiveness measure* and the average number of fingerprints of a document, respectively.

Table 2 shows the results of running our text reuse detection methods on a large blog collection (over 3 million postings). Categories C1 through C6 are different relationships between pairs of documents in terms of text reuse. The table shows that although a considerable part of the duplicated text in blog postings comes from spam and site boilerplate text (“Frame”), there is still a considerable amount of borrowing and reusing text in blogs.

Pattern	C1	C2	C3	C4	C5	C6	Total
Text Reuse	16%	20%	20%	6%	12%	18%	15%
Common Phrase	2%	12%	12%	24%	28%	28%	18%
Spam	30%	22%	20%	8%	12%	20%	19%
Frame	36%	46%	48%	62%	48%	34%	46%
URL Aliasing	16%	0%	0%	0%	0%	0%	3%

Table 2. Text reuse categories in the TREC Blogs06 collection.

Details of this work and results are reported in:

Seo, J. and Croft, W. B., "Local Text Reuse Detection", *Proceedings of the 31st Annual International ACM SIGIR Conference (SIGIR 2008)*, vol. , (2008), p. 571-578.