d-lib magazine

The SCAM Approach to Copy Detection in Digital Libraries

Narayanan Shivakumar and Hector Garcia-Molina

Department of Computer Science
Stanford University
Stanford, CA 94305, U.S.A
{shiva, hector}@cs.stanford.edu

Scenario 1

Your local publishing company Books'R'Us decides to publish on the Internet its latest book in an effort to cut down on printing costs and book distribution expenses. Customers pay for the digital books using sophisticated electronic payment mechanisms such as DigiCash, First Virtual or InterPay. When the payment is received, the book distribution server at Books'R'Us sends a digital version of the book electronically to the paying customer. Books'R'Us expects to make higher profits on the digital book due to lower production and distribution costs, and larger markets on the Internet.

It turns out, however, that very few books are sold since digital copies of the Books'R'Us book had been widely circulated on UseNet newsgroups, bulletin boards, and had been available for free on alternate ftp sites and Web servers. Books'R'Us retract their digital publishing commitment blaming the ease of re-transmission of digital items on the Internet, and return to traditional paper based publishing.

Scenario 2

Sheng wants to buy a new Pentium portable, and hence wants to read articles on the different brands available and their reviews before choosing a brand to buy. She searches information services like Dialog, Lycos, Gloss and Webcrawler, and follows UseNet newsgroups to find articles on the different portables available and finds nearly 1500 articles. When she starts reading the articles, she finds that most articles are really duplicates or near-duplicates of one another and did not contribute any new information to her search. She realizes this is because most databases maintain their own local copies of different articles in perhaps different formats (Word, Postscript, HTML), or have perhaps mirror sites that contain the same set of articles. Sheng then trudges through the articles one-by-one wishing that somebody would build a system that can remove exact or near-duplicates automatically so that she only needs to read each distinct article.

Around article number 150, Sheng decides not to buy a certain brand since from the articles she learns that that brand had had problems with its color display since its release. But she has to continue looking at articles on that model since they are already a part of the result set. She adds to her wish list a dynamic search system in which she could discard any articles that have more than a certain overlap with some article she had previously discarded.

In this article, we will give a brief overview of some proposed mechanisms that address each of the problems illustrated by the above two scenarios.

In Copy Guarantees for Digital Publishers , we consider mechanisms that make it harder to redistribute or republish digital documents or their components with impunity. In Duplicate Detection in Information Retrieval , we discuss mechanisms that can remove near-duplicates (such as multiple formats) in sets of retrieved documents. We will then present the SCAM Registration Server that can assist in detecting illegal copies or copies within retrieved document sets.

Copy Guarantees for Digital Publishers

Some publishing entities such as the Institute of Electrical and Electronics Engineers (IEEE) have sought to **prevent** illegal copying of documents by placing the documents on stand-alone CD-ROM systems, while others have chosen to use special purpose hardware [PoKl79] , or active documents [Gr93] (documents encapsulated by programs). We believe that such prevention techniques may be cumbersome, may get in the way of the honest user, and may make it more difficult to share information. In fact, the software industry has noticed that measures to prevent software piracy may actually reduce software sales; hence software manufacturers currently prefer to try piracy detection rather than piracy prevention [BDGa95].

Drawing on our analog from the software industry, we advocate detecting illegal copies rather than the copy prevention approach. In the copy detection approach, there are two important orthogonal questions.

1. Is a document at a certain Web site or an article posted on a newsgroup an illegal copy of some pre-existing document?
2. If the document is an illegal copy, who was the originator of the illegal copy?

We will now look at some popular schemes to address each of the two questions.

Registration Server

One answer to the first question (that we also pursue) is to build a registration server: documents are registered into a repository, and query documents are checked with the documents in the repository to detect any possible copies [PaHa89, BDGa95, ShGa95]. In Figure 1 (below), we show an example of a generic registration server which consists of a repository (may be distributed) of documents. A stream of documents arrives to be registered into the database or to be checked against the current repository of documents for duplication or significant overlap.

The registration server can be used in two ways. The first is to have authors and publishers of original works register them at the server. In this case, publishers can use the server to check a to-be-published document for originality. Similarly, "crawlers" can automatically sample bulletin boards, news groups, mailing lists, and Web sites, and check those documents against the registered ones. The second option is to automatically collect and register large numbers of articles from the open sources. (They would only be kept at the server for a few days, else the registration database may be too large.) In this case, publishers or authors who are worried that their works are being copied illegally can check them against the registered database. Notice that in both cases, when the server reports a duplicate or near-duplicate, it only reports a potential problem; a human would have to examine the matching documents to determine if it is an actual legal or ethical problem.

There have been several approaches to building efficient registration servers [BDGa95, Man94 ShGa95] , scalable to hundreds of thousands of documents. In Architecture of SCAM , we report on the **Stanford Copy Analysis Mechanism**, one of the registration servers from our research group currently available on the Internet . This server can detect duplicates in a variety of ways, for example, by looking for matching sentences or matching sequences of words. It can not only detect identical documents, but can also detect documents that overlap significantly, where "significantly" can be defined in a variety of ways.

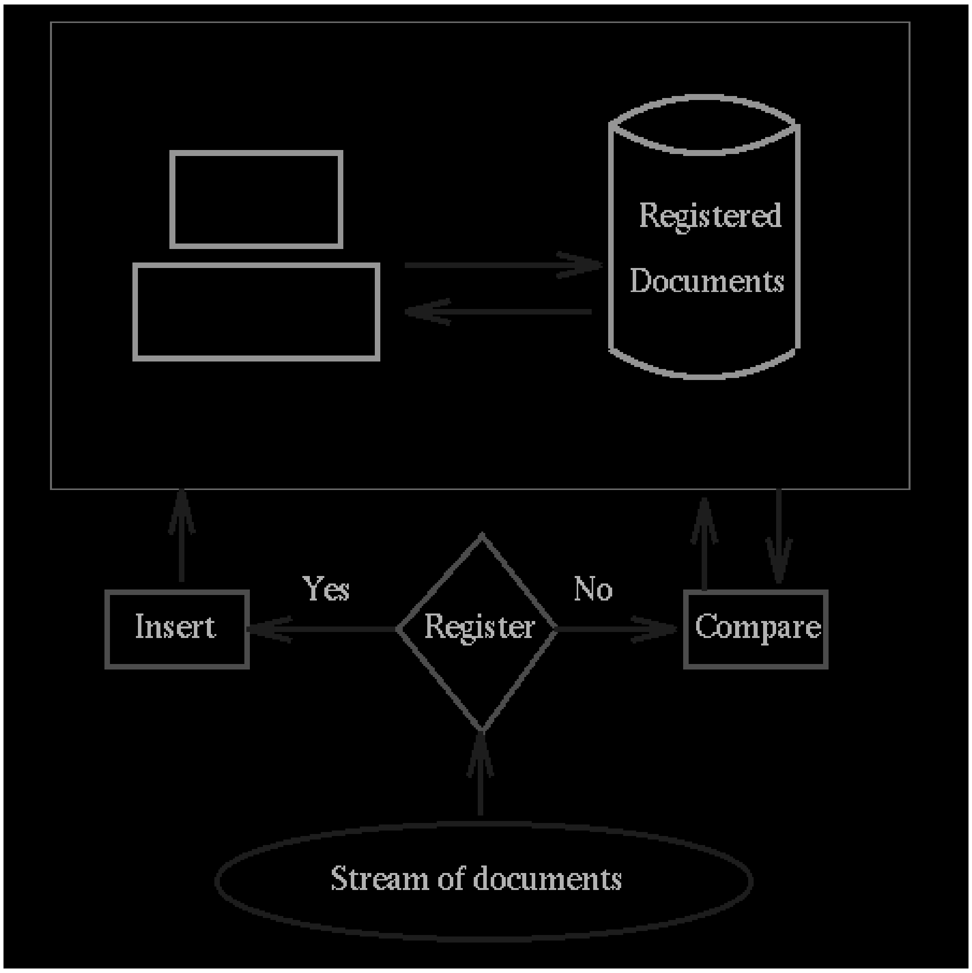

Figure 1

Document Signatures

Once a document is suspected or known to be an illegal copy (through one of the registration server schemes outlined above), it is sometimes useful to know who was the originator of the copy. For instance, Books'R'Us would like to know which one of its few paying customers is re-transmitting its books for commercial advantage (or for free). In signature based schemes, a "signature" is added to each document, and this signature can be used to trace the origins of each document. One approach is to incorporate unique *watermarks* encoded in word spacings or images into documents [BoSh95, BLMG94a, BLMG94b, CMPS94]. The unsophisticated user is not aware of the watermarks, but when an illegal copy is found, the book distribution server at Books'R'Us can determine who purchased the book with that particular signature.

We believe that by combining the notion of a registration server to detect possible illegal copies, and by using document signatures to trace the originator of the copy, one can detect and discourage much of the illegal copying that occurs on the Internet. Of course, none of these schemes is perfect. For example, a user can print and distribute multiple paper copies of a digital document. However, this involves more effort than electronic duplication, and furthermore, paper copies may not have the same "value" as digital copies (e.g., no hyperlinks, cannot be searched on-line). Users could also transmit documents to a "small" number of friends without being caught. But in terms of lost profits, this problem is not as serious as the one where copies are widely circulated. Moreover, copy detection and tracing mechanisms can discourage even small scale duplication because one will never know when a "friend" will give a copy to another "friend" who will then post to NetNews, eventually getting the original purchaser into trouble and thereby creating a disincentive for anyone potentially initiating such a chain of events.

Duplicate Detection in Information Retrieval

In this age of information overload, an important value-added service that search engines and databases can perform is removing duplicates of articles before presenting search results to users. For example, SIFT is an information filtering service where users subscribe to information of interest [YaGa95a] . (SIFT is currently used by over 11,000 users from all over the world.) Subscribers periodically receive (via email or "personal" Web pages) set of documents that match their interests. Within these sets, we found that duplicates or near duplicates were a common problem for users. They arise, for instance, because the same document appears in multiple formats, because articles are cross-posted in different newsgroups, or because articles or substantial parts of them are forwarded in other posting to newsgroups or mailing lists.

In [YaGa95b] , we show how a copy detection be used to automatically remove multiple copies of the same article, and how a user may dynamically discard certain classes of articles that have sufficient overlap. In that paper, we discuss how users can specify what new documents trigger removal of their copies in the future, how significant overlap should be to trigger removal, and how long a document needs to be remembered for duplicate removal purposes. For example, a user who sees a Call for Papers for a conference he dislikes, may specify that we does not want to see any partial or full copies of this call forever. (The system may actually force a smaller time window for performance reasons.) Another user who sees a report on a Pentium bug may indicate he does not what to see identical copies (knowing that this report will be copied widely), but would like to see partial copies since they may include interesting commentary. The user may also specify default values for the duplicate removal parameters. In Architecture of SCAM (below), we present the underlying registration mechanism of SCAM that could be used as a "plug-in" module for copy detection in a system such as SIFT.

Architecture of SCAM

We first show SCAM from the user's perspective as a black-box entity on the Internet. We will then briefly describe the underlying implementation of SCAM.

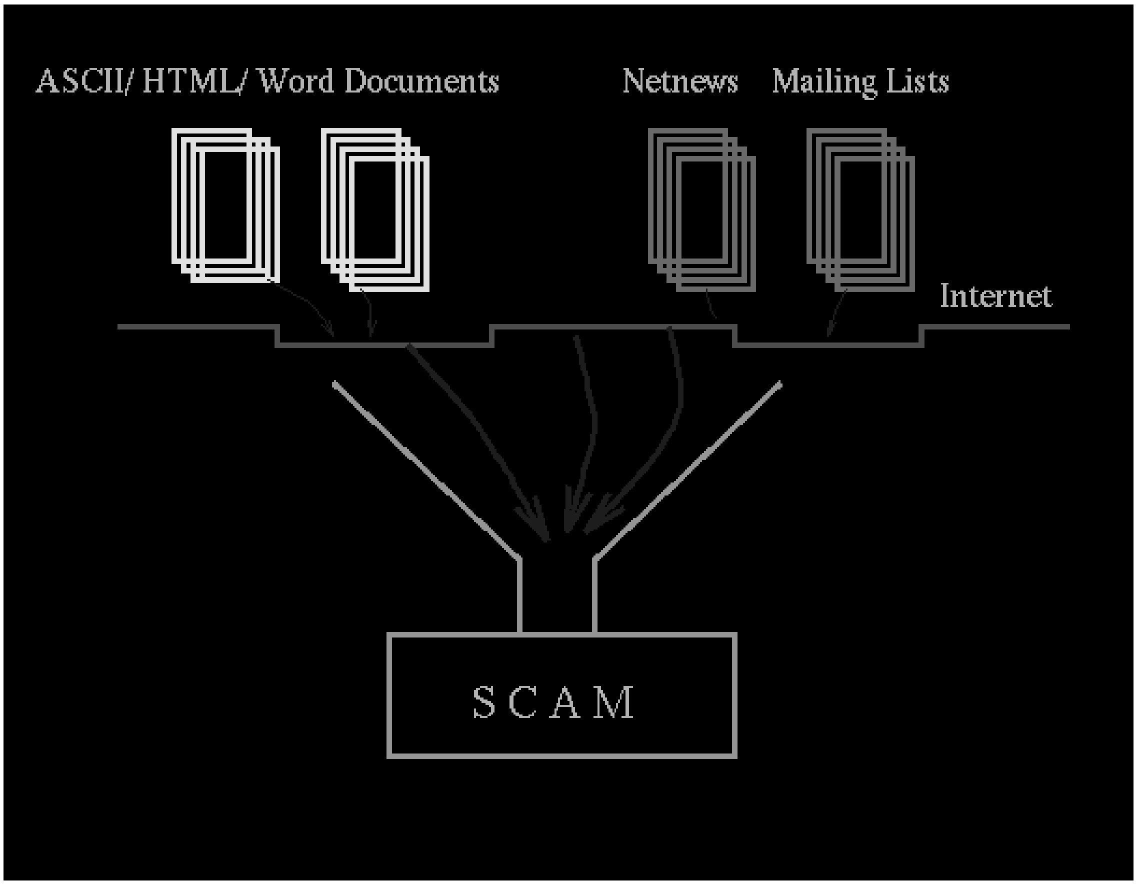

Figure 2

In Figure 2, we show conceptually how our current implementation of SCAM is accessible through the Internet. Netnews articles from the Usenet groups, and from certain large mailing lists are currently being registered into SCAM on a daily basis into a publicly accessible repository. (After a few days, the documents are purged from the repository.) We have also developed a form-based web client, and a bulk loader so that users across the Internet may send documents of different formats (such as ASCII, HTML, Postscript) to be registered into their private databases, or to be queried against their private databases or the public repository.

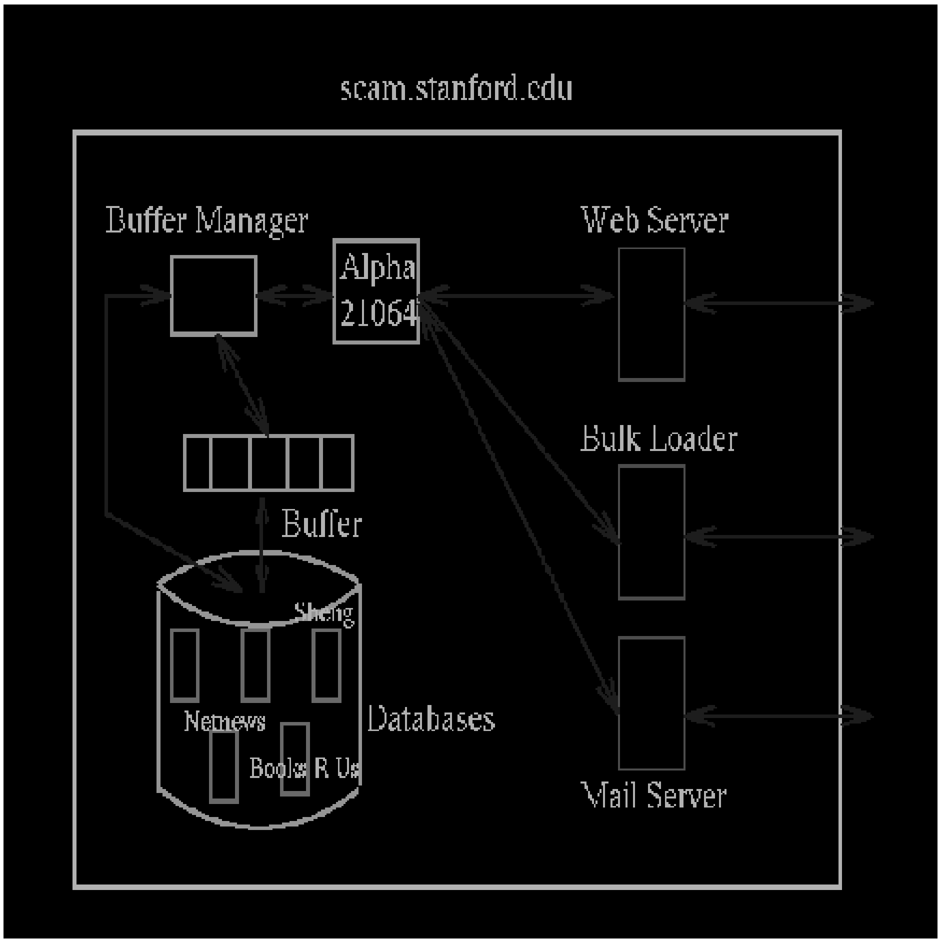

Figure 3

In Figure 3, we show the underlying architecture of SCAM that provides the functionality of Figure 2. SCAM currently runs on a Dec 3000 at scam.stanford.edu. It has the traditional database components of a buffer manager controlling a buffer (10 Megabytes), and a disk (1 Gigabyte). There are several databases on the disk which may be part of the public repository (such as Netnews, Mailing Lists) or may be personal user-owned databases (such as Sheng's or Books'R'Us). Different servers (like the Web Server) have been implemented to provide interfaces for users accessing SCAM, and different parsers are employed to handle multiple formats (HTML, postscript, ASCII etc.).

Possible Applications of SCAM

There are several possible ways in which SCAM may be used. We now outline some of the more interesting applications of SCAM.

1. Book companies, authors, commercial UseNet groups and professional societies (like ACM, SIGMOD) that have valuable digital documents may create their own personal databases with SCAM, and register their digital documents (through our Web server, mail server or bulk loaders) into their databases. SCAM will then check UseNet newsgroups, mailing lists and some Web sites

on a daily basis for full or partial overlap (similar sentences, paragraph chunks etc.) and will notify the appropriate user of the overlap and the source.
2. Other users can probe the public databases and check if some document they are interested in overlaps with some article in one of the public sources (UseNet newsgroups, web pages, mailing lists).
3. Class instructors may create their own databases and store into those any articles they may find relevant for classes they teach, and also register digitally submitted homeworks into the database. When the class is offered again, he may use the database to check for any significant overlap to previous homeworks and other registered articles. Similarly, Program Committee Chairs of Conferences and Editors of Journals may use databases specific to their field to check if any new submission overlaps significantly with some previous paper in the field. (Actually, SCAM was successfully used in this mode in early 1995. It helped find several cases of plagiarism of published technical papers. See [details].)

We believe Copy Detection Mechanisms such as SCAM will play an important role in Digital Libraries, making it easier to identify illegal copying and thereby inducing paper-based publishers to switch to digital publishing. Automatic duplicate removal of documents in information retrieval and filtering will also become increasingly important as the number of sources used in searches increases.

Since the number of digital documents is increasing at a fast rate, an important area of research is how to make copy detection mechanisms scale to such large number of articles without losing accuracy in overlap detection. We are currently considering a distributed version of SCAM for reasons of scalability. We are also experimenting with different approaches to copy detection which have different levels of expected accuracy and expense.

Acknowledgements

This research was partially sponsored by the Advanced Research Projects Agency (ARPA) of the Department of Defense under Grant No. MDA972-92-J-1029 with the Corporation for National Research Initiatives (CNRI). The research is also partially supported by the National Science Foundation under Cooperative Agreement IRI-9411306. Funding for this cooperative agreement is also provided by ARPA, NASA, and the industrial partners of the Stanford Digital Libraries Project. . Its content does not necessarily reflect the position or the policy of the Government, CNRI or other sponsoring parties, and no official endorsement should be inferred.

References

[BDGa95]
: S. Brin, J. Davis, H. Garcia-Molina, Copy Detection Mechanisms for Digital Documents *Proceedings of the ACM SIGMOD Annual Conference, San Francisco, CA, May 1995. [PostScript]*

[BoSh95]
: D. Boneh, J. Shaw, Collusion-secure fingerprinting for digital data. *Technical Report 468,*

Computer Science Department, Princeton University, January 1995.

[BLMG94a]
J. Brassil, S. Low, N. Maxemumchuk, L.O. Gorman, Document marking and identification using both line and word shifting. *Technical Report, AT & T Bell Labs, 1994, ftp://ftp.research.att.com/dist/brassil/docmark2.ps*

[BLMG94b]
J. Brassil, S. Low, N. Maxemumchuk, L.O. Gorman, Electronic marking and identification techniques to discourage document copying. *Technical Report, AT & T Bell Labs, 1994, ftp://ftp.research.att.com/dist/brassil/docmark2.ps*

[CKPT95]
D.R. Cutting, D.R. Karger, J.O. Pedersen, J.W. Tukey, Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. *Proceedings of 15th Annual SIGIR Conference, Denmark*

[CMPS94]
A.Choudhury, N.Maxemchuk, S.Paul, H.Schulzrinne, Copyright protection for electronic publishing over computer networks. *Technical report, AT & T Bell Labs, 1994.*

[Gr93]
G.N. Griswold, A method for protecting copyrights on networks. *Joint Harvard-MIT Workshop on Technology Strategies for Protecting Intellectual Property in the Networked Multimedia Environment.*

[Man94]
U.Manber, Finding similar files in a large file system. *Proceedings of USENIX Conference, 1-10, San Francisco, CA, January 1994.*

[PaHa89]
A.Parker, J.O Hamblen, Computer Algorithms for plagiarism detection. *IEEE Transactions on Education, 32(2):94-99, May 1989.*

[PoKl79]
G.J. Popek, C.S. Kline, Encryption and secure computer networks. *ACM Computing Surveys, 11(4): 331-356, December 1979.*

[ShGa95]
N.Shivakumar, H. Garcia-Molina, SCAM: A Copy Detection Mechanism for Digital Documents. *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries, Austin, Texas, 1995. [PostScript]*

[YaGa95a]
T.Yan, H. Garcia-Molina, SIFT - A Tool for wide-area information dissemination. *Proceedings of USENIX, 1995*

[YaGa95b]
T.Yan, H. Garcia-Molina, Duplicate detection in information dissemination. *Proceedings of Very Large Database (VLDB'95) Conference, Zurich, Switzerland, September 1995*

-->

---

hdl://cnri.dlib/november95-shivakumar