**STANFORD UNIVERSITY**
Department of Computer Science
Stanford, California 94305-2140

NARAYANAN SHIVAKUMAR

## Summary of Thesis Research

The World Wide Web is a content publisher's nightmare come true. Currently, any small time cyber-pirate can make copies of music CDs and books available on the web in digital format to a large audience at virtually no cost. Hence content producers (such as artists, musicians and writers) and content publishers (such as record companies and book publishers) expect to loose several billions of dollars in royalty revenues.

In my thesis, I focus primarily on building a *copy detection system* (CDS) for copyright violation detection. In this system, content publishers (such as Disney and Times Warner) *register* their valuable digital content into the registration repository of the CDS. The CDS then *polices* publicly accessible sources on the web, identifies illegally copied content by comparing the web content to the registered content and notifies the content owners of illegal copies. Some of the challenges in building such a system include

- **Accuracy:** The CDS should have low false negative errors in that it should not miss illegal copies. Similarly, the CDS should have low false positive errors in that it does not wrongly classify some object to be an illegal copy of a registered object.
- **Resiliency:** The illegal copy of a registered object may not be an exact copy. For instance, the cyber-pirate may post on his web site only a section or a paragraph of a registered book, or a 1-minute clip from a 5-minute registered song. Similarly, the illegal copy may be a "noisy" version of the original copy. For instance, the illegal copy may be sampled at a different rate, use a different format, or may use different quantizations. The CDS should be resilient to such forms of tampering and identify the above to be illegal copies as well.
- **Scalability:** The number of web pages and servers is large and is growing at a rapid pace. In this context, the techniques the CDS adopts should scale in terms of the crawl strategy as well as how it identifies the illegal copies.

To address the above challenges, in my thesis I have developed a core architecture that can be used to build copy detection systems for a variety of data types. As proof of concept, I have built two prototype CDS: (1) SCAM (*Stanford Copy Analysis Mechanism*) for finding copies of textual documents (such as PostScript, HTML, ASCII documents), and (2) FRAUD (*Finding Replicas of AUDio*) for finding copies of audio files (such as WAV and MP3 files). Some of the key modules that constitute my prototypes include:

- **Crawler:** The CDS prototypes use the following types of crawlers to roam the web, retrieve objects and compare these objects with registered objects.

1. **Traditional:** The CDS uses traditional crawlers like those servicing search engines (such as AltaVista, Excite) to roam the web and retrieve HTML pages and sound clips.
2. **Probe:** Often, digital objects are stored in a searchable database. For instance, INSPEC is a compendium of abstracts of conference papers and journal articles that can be searched using boolean queries. In such cases, the crawler automatically generates queries to retrieve text documents in the searchable database based on the registered text document. Similarly with audio files.
3. **Streamed:** Many web sites push content across to clients using streaming technologies such as PointCast, RealAudio and NetVideo. The crawler processes such streams and delivers these objects to the CDS.

- **Feature extraction:** Given a set of objects retrieved from the web, the CDS should efficiently check if the retrieved objects are illegal copies of registered objects. Recall that the comparison procedure should be resilient to tampering by a cyber-pirate. For instance, we need to identify retrieved textual documents that significantly overlap with some registered textual document. Similary we need to identify audio files that are noisy versions of some registered audio file. For this, we extract important features from digital objects (registered as well as crawled), and compare these features. The key challenge in extracting these features is that they should be resilient to tampering, fast to extract, fast to compare and scale to several hundreds of gigabytes of data. In my thesis, I have studied many different varieties of feature extraction techniques and incorporated the appropriate techniques into my prototypes.
- **Indexer:** Once the CDS extracts its features, these features are stored into indices that I specifically designed and built for identifying illegal copies. These indices are a key aspect of my prototypes in that they are a factor of hundred faster and a factor of hundred more space efficient than the indexes built in current database systems for query processing.

The reader may be interested to know that we have used the SCAM prototype to find actual occurrences of academic plagiarism. In a renowned case, a graduate student had copied and partly modified technical reports of researchers in computer science and resubmitted them under his own name to different journals and conferences. When ACM was compiling a case against him, they had access only to his abstracts. But in order to compile the case, ACM needed to find the set of original technical reports he had copied from. We used SCAM to track these initial technical reports using abstracts from INSPEC, a repository of abstracts of technical papers in computer science. The success of SCAM was reported in several mailing lists and newsgroups on the Internet, as well as magazines such as Communications of the ACM, Forbes Magazine, The Recorder and the Intellectual Property Magazine.