Non-Verbatim Copyright Infringement Detection for Text

Ozlem Uzuner & Boris Katz

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Copyright infringement is a serious problem which threatened authors of creative works even in the non-electronic world. In the electronic world, easy access to electronic documents and the ease of reproducing and distributing these documents have made copyright infringement an even bigger problem. In general, the ease of reproducing and distributing online documents threatens not only the economic interests of authors but also the growth of the Internet as an online resource.

Motivation: Given the importance of the Internet as an online resource used both in education and in research, it is desirable that online libraries enable authorized people to access valuable information, while safeguarding the interests of information providers who agree to make their information available in these libraries. Maintaining such a safe and fair information resource requires protection against copyright infringement as well as other problems.

Related Work: There have been two kinds of attempts to solve the problem of copyright violations on the web. The first kind is based on protection of the work by making it difficult to copy. For example, placing the conference papers on a CD-ROM rather than putting them on the web is a form of copy protection. Copy protection methods make copying difficult, but not impossible.

The second kind of attempts can be generally called copy detection. These systems detect copyright violations by finding partial overlaps among documents. Copy detection methods do not make copyright infringement impossible either. However, they might be effective deterrents.

There have been a few attempts to identify "verbatim" copyright violations, i.e., violations where part of the document has been copied character for character. These systems prepare fingerprints of documents from the web in order to form a repository and match target documents against the fingerprints in their repository.

Depending on their main goal, each of the currently existing systems like CopyCatch [1], turnitin.com [2] and SCAM [4] calculate fingerprints in different ways and define similarity differently. For example, SCAM uses different chunking strategies to break document text into smaller pieces. In order to detect partial overlaps of sentences, SCAM's authors experiment with overlapping k-word chunks among others. The authors mention that slight modifications by infringers to otherwise identical documents, e.g., inserting an extra space, result in completely different chunks for two otherwise identical documents defeating the copyright infringement detection software.

None of the currently existing detection programs forms a **conceptual** fingerprint of documents. With string based fingerprinting, alternative ways of communicating the same concepts end up being represented differently, and this makes slightly paraphrased but conceptually equivalent documents look completely unrelated.

Identifying conceptually significant units and finding common representations for conceptually equivalent expressions among documents is an important part of document fingerprinting, and is necessary in order to identify subtle infringements.

Approach: As an attempt to solve the non-verbatim copyright infringement detection problem, we focus only on paraphrase detection.

Our proposed solution to this problem is based on the claim that words may differ in form and/or be related differently in syntax due to paraphrasing, and yet hold the same relationship to each other with respect to meaning. Consider the following statements:

1. An engineering methodology to identify profitable market segments for the use of new materials is presented

[7]...

2. We describe a utility-based method which can be used for identification of market segments for the use of new materials.

3. This paper presents a method for identifying profitable markets for new materials.

In all of the above sentences, **method** and **identify** have the same kind of relationship with each other; the concept described by these two words is the same in all three sentences. We claim that the "relationship" between these two words remains constant even when the sentence they appear in is paraphrased. In contrast, **method** and **identify** in 4 do not have the same relationship as their counterparts in sentences 1-3.

4. We identified a useful method which can be applied to many cases.

The 'relationship' between two words that constitute a concept can be defined as a combination of lexical attraction, association ratio, mutual information and syntactic relationships like subject-verb-object patterns.

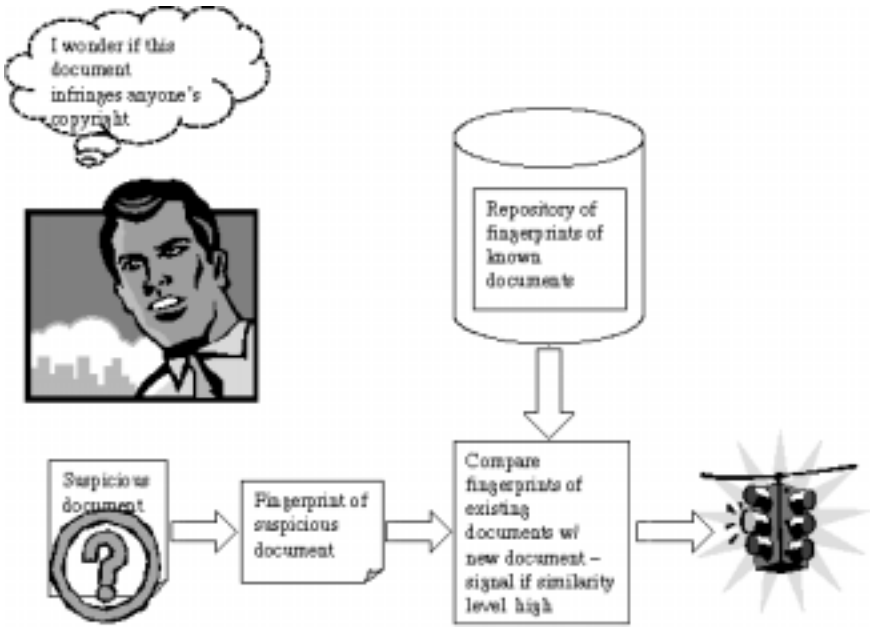

Figure 1: Overview of Copyright Infringement Detection Process

Future Work: We are currently experimenting with existing statistical measures that can help identify conceptual units within a document. These units make up the fingerprint of the document and can be used to compute similarities among documents. The problem of solving the rate of similarity between two expressions of the same concept will be based on the results of fingerprinting.

Research Support: This research is funded by DARPA under contract number F30602-00-1-0545 and administered by the Air Force Research Laboratory.

References:

[1] http://www.copycatch.freeserve.co.uk/vocalyse.htm.

[2] http://www.turnitin.com.

[3] C. G. E. Mangin, R. de Neufville, F. Field III, , and J. Clark. Defining markets for new materials: Engineering methodology with case application. *Resources Policy*, (3):169–178, 1995.

[4] N. Shivakumar and H. Garcia-Molina. Scam: A copy detection mechanism for digital documents. *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries*, 1996.

[7] Sentence taken from Mangin [3].