Özlem Uzuner
Identifying Expression Fingerprints Using Linguistic Information
Dissertation Abstract

This thesis presents a technology to complement taxation-based policy proposals aimed at addressing the digital copyright problem. The approach presented facilitates identification of intellectual property using expression fingerprints.

Copyright law protects expression of content. Recognizing literary works for copyright protection requires identification of the expression of their content. The expression fingerprints described in this thesis use a novel set of linguistic features that capture both the content presented in documents and the manner of expression used in conveying this content. These fingerprints consist of both syntactic and semantic elements of language. Examples of the syntactic elements of expression include structures of embedding and embedded verb phrases. The semantic elements of expression consist of high-level, broad semantic categories.

Syntactic and semantic elements of expression enable generation of models that correctly identify books and their paraphrases 82% of the time, providing a significant (approximately 18%) improvement over models that use tfidf-weighted keywords. The performance of models built with these features is also better than models created with standard features used in stylometry (e.g., function words), which yield an accuracy of 62%.

Current approaches to the digital copyright problem employ technological protection mechanisms in the form of Digital Rights Management (DRM) systems that limit the uses of a digital work to those that are authorized by its copyright holder. These systems enable the copyright holder to control all uses of digital works, including some fair uses that are in the best interest of the public but that conflict with the interests of copyright holders. In addition, these protection mechanisms enable copyright holders to inhibit others’ speech and to collect private information about the ways individuals use digital works. These and similar side effects contravene the stated goal of U.S. copyright law: promotion of “Progress of Science and the Useful Arts”.

Digital tracking technologies enable alternate solutions to the digital copyright problem. Expression fingerprints facilitate digital tracking even when literary works are DRM- and watermark-free, and even when they are paraphrased. Given methods for accurate metering of popularity of works, we can encourage large-scale dissemination and unrestricted use of digital works by society while protecting the revenues of copyright holders, for example through taxation-based revenue collection and distribution systems.