GEOSPATIAL CONTEXT EXTRACTION FOR CREATIVE COMMONS LICENSED DIGITAL CONTENTS

Davide Bardone, Elias S. G. Carotti, Juan Carlos De Martin

NEXA Center for Internet & Society
Dipartimento di Automatica ed Informatica
Politecnico di Torino
c.so Duca degli Abruzzi, 24 — 10129, Torino, Italy
phone: +(39) 011 564 7219, fax: + (39) 011 564 7216
email: [davide.bardone|carotti|demartin]@polito.it
web: http://nexa.polito.it/

During recent years many authors and content creators around the world published their creations under Creative Commons (CC in the following) licenses. In fact, many diverse kinds of digital objects have currently been put under a CC licensing scheme, ranging from entire blogs, books, to music, film footage and paintings. However, despite the widespread and ever increasing adoption of these licenses, it is still difficult to estimate not only the geographic origin but also simply the quantity of the licensed works. In fact, because of the lack of a centralized repository or directory for CC-licensed works – which, however, would require every creator to advertise his creations – and the relatively infrequent introduction of metadata to describe the digital objects in the web pages hosting them – which would allow for a crawler to simply browse the web and *count* the objects – reliably estimating the quantity and the geographical distribution of commons-based digital contents and the relative distribution of each CC license (attribution, share alike, no derivatives, non-commercial, and their combinations) is still an open challenge.

In the seminal work [1] Cheliotis *et al.* estimated the number of CC licensed contents by querying the Google and the Yahoo search engines for pages containing backlinks to the CC-licenses' deeds, or containing specific metadata describing the license and the content (such search option, specifically tuned for CC-licensed content, exploiting RDF descriptions, is offered by both Google and Yahoo.) They chose to perform the search using also 75.000 random english words to query the search engines instead of not well documented wildcards. Moreover, besides merely estimating the total number of CC-licensed digital objects, the authors used the same method to estimate the geographic location of the digital objects, by means of their licenses jurisdiction.

Unfortunately, these approaches are not exempt from estimation errors due to the difficulty to avoid double counting or omitting to count certain digital objects. For example, many different ones might be present on a single web page where just one single backlink to the license is present, covering all the objects included in the same page. Moreover, results are clearly affected by the search engine algorithm which might result in (difficult to correct) estimation biases.

In fact, associating digital objects with geographical information is not an easy task and often it is not even possible to deterministically estimate the correct region. In [2] different features are taken into account to infer the geographical location of provenience of a web page content. For example, information about the web server hosting the page is considered, such as those coming from the WHOIS [4] database, or the way the traffic is routed on the Internet (i.e. the neighbouring routers), the association between IP addresses and countries (which is readily available via online databases), and, finally, the country code top level domains (e.g. .uk, .it, .fr, etc.). Other important sources of geospatial context can be extracted from the *content* of a web page. For example, sometime addresses, postal codes, telephone numbers, geographic feature names are present, and even the language used for the content might bear some useful information. Obviously, often these clues are not present or might not be available at the same time, and, in any case, are characterized by varying degrees of reliability and precision. The most problematic aspect is that often the geographic context of the *content* of a web page can be very different from the one associated to the machine which hosts it, for example a French user can write a CC licensed blog hosted by a server in the United States.

However, it is still important to check the backlinks, because the choice of a specific localized version of the license might bear some information about the provenience of the digital content, although not every country has a specific version of CC licenses maintained by a local volunteer team and the use of the generic version of these licenses is still widespread (estimated to about 80% in [1].)

In this paper we propose a method to achieve more precise and more consistent estimates of the geographic context of CC-licensed digital objects.

The proposed technique collects some of the *features* described above for every web page containing some Creative Commons licensed digital objects, to determine a soft, probabilistic association between the content and the geographical location, according to a statistical model which has to be learned offline on a large training data set.

First a significant set of web pages containing digital objects has to be collected and hand labeled to associate them with the country, then this corpus is used as ground-truth to train a number of Gaussian Mixture Models (one for each country considered) by means of the Expectation Maximization algorithm (EM). Each GMM thus encodes the statistical dependence across the features for a different country and can be used to compute the *likelihood* that a specific content originates from a specific country.

For this purpose we designed and developed a web crawler which identifies and collects all the pages where an explicit choice of a CC license has been done, by means, for example, of the RDF metadata or the backlinks to the licenses' deeds. Moreover, the crawler extracts and stores, for each visited web page, some geospatial context features in a proper database. The features considered include the IP address, the WHOIS database information, its top level domain, the character set adopted to write the HTML page, the language and the URL of the page who linked to the current one.

The IP address of the server hosting the page is first mapped to a country by means of the MaxMind [5] APIs, then, the WHOIS database is queried for the same IP address to identify the server's geographical location. Also, the top level domain is considered (as it can bear relevant information, especially if it corresponds to a country code top level domain (ccTLD).) Moreover, the crawler parses the HTML code of each page and extract, if present, information about the character set used, and analyzes the text content to determine the language by means of the algorithm described in [3]. This algorithm is able to extract a *language profile* of the text and determine the language by finding the best match with a number of precalculated language profiles. The URL of the page linking the current one is recorded by the crawler in order to exploit any possible correlation between the nationality of the respective contents.

This approach can also attempt to evaluate Creative Commons licenses adoption by nationality for all the digital contents contained in great communities like Flickr, Jamendo, etc., if more features are considered, including, for example, users' profiles.

Future work includes the ability to distinguish between different types of CC licensed contents within web pages, also taking into account the nature of the content itself.

REFERENCES

[1] Giorgios Cheliotis, Ankit Guglani and Giri Kumar Tayi, "Measuring the Commons: Quantifying Global Online Licensing Behavior," *3rd Symposium on Statistical Challenges in E-Commerce Research*, May 15-17 2007, University of Connecticut

[2] Kevin S. McCurley, "Geospatial Mapping and Navigation from the Web" *Proceedings of the 10th International World Wide Web Conference*, 2001, Hong Kong

[3] William B. Cavnar and John M. Trenkle, "N-Gram-Based Text Categorization" *Proceedings of SDAIR- 94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994

[4] IETF, "RFC 3912", `http://tools.ietf.org/html/rfc3912`

[5] MaxMind, "GeoIP APIs", `http://www.maxmind.com/app/api`