Source Geography Estimation for Web Pages

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
[davide.bardone|carotti|demartin]@polito.it

***Abstract*— The problem of inferring geographical information associated to web pages and identifying the geographic scope of their content is gaining increasing attention. However, geographic scope is a concept that can be interpreted in many different ways, ranging from the expected target scope of a specific content to the country where the content originated. The latter, in particular, albeit difficult to address, is of great importance for many reasons, such as, for example, market inquiries or anytime estimates on content production in specific countries are needed. Search engines may also be affected by the knowledge of the various kinds of geographic scopes, to better tune their responses to queries, e.g. according to (but not restricted to) the geographic proximity with the user location. However that information is rarely available and must be inferred in the vast majority of the cases. In this paper we propose a technique, grounded into the machine learning theory, to estimate source geography of web pages by means of a classifier learned on a specially constructed training set. The training set, consisting of a number of features extracted from web pages and the corresponding source-geography label (i.e. the country of origin of the web page) is automatically built by exploiting the wide number of pages with contents licensed under a localized Creative Commons (CC) license. The model thus learned is then used to classify unlabeled records and our tests showed a mean accuracy of* 81% *with a standard deviation of* 0.9.

Keywords*— Geographic information retrieval, Classification.

I. Introduction

Everyday an enormous quantity of information is retrieved from the World Wide Web using search engines, which offer efficient finding and fruition of specific data that is by its nature widely spread across the Internet.

However, most search engines typically exploit simple textual queries to find pertinent documents containing specific weighted *keywords* via an inverted index, which maps terms to web pages. These kind of tools may also take into consideration other parameters such as, for example, language, file type, and usage rights. Moreover, a user might want to restrict searches to pages having a specific geographical origin or scope, which, in general, may be different from the physical location of the server hosting the resources to be searched. In fact, a user may want to search for specific content from a specific country, such as Italian recipes or, say, mp3 files, or search for a certain keyword in the web when within a certain geographic scope.

In recent years the problem of inferring geographical information contained in web pages in order to determine the geographic context of their content gained increasing attention. The extraction of this kind of information is mostly aimed at allowing web applications, such as search engines or intelligent agents, to retrieve information and compute relevance with respect to a specific query including (but not restricted to), for example, the geographic proximity with the user location. Moreover, the possibility to associate a web page with its geographic scope can also allow to estimate and collect location wise statistics in order to understand the global or local importance of a web site considering its geographical popularity, given by the distribution all over the world of other sites linking to it. Last but not least, knowing the geographic scope and the origin of web pages may be of great importance for market analysts seeking information about interests and needs in specific geographic regions.

However, despite its simple definition, the problem is not an easy one to address, especially if specific metadata (which would render the problem solution almost trivial) are lacking. In principle, it is already possible to tag specific resources with proper metadata specifying their geographical origin ([1], [2], [3]), but, like most other initiatives on metadata, it suffers for the chicken and egg problem and thus, to date, most of the available content on the web lacks this kind of information which, at best, has to be inferred from contextual information.

Moreover, the geographic context problem is many-folds: in fact the geographic scope of the audience seeking specific resources might be different from their geographical origin, which, in turn is, in general, different from their physical location.

[4] and [5] already came up with at least two different definitions of geographic context, namely the *target-geography*, or *content-based* geographic context, which relates to the geographical scope of a web page content and the *source-geography*, or *entity-based* geographic context, which, on the other hand, refers to the geographic context where the content was created, which typically means the location associated with its author or the website.

Given the broad implications and aspects of the problem, past literature mostly focused either on determining the geographical location of the hosting services or on extracting geographical information from the web pages' textual content, i.e., estimating geographic context based on the content, the rationale being that the geographical scope of a page is strictly correlated with the locations it refers to.

Buyukokkten *et al.* [6] extracted the area codes from the publicly available phone numbers of all the network administrators for the Class A and B domains; each area code was then mapped to cities, counties and states which were considered to be the geographic scopes of the corresponding IP addresses.

[4] used simple heuristics to infer the geographic information associated with servers and web sites from the output of standard network tools such as `traceroute` and `whois`, and the names, addresses, postal codes, and telephone numbers in the textual content.

Ding *et al.* [7] described both an estimation technique to exploit the geographic position of sites linking a specific page thus analyzing its area of interest, and an alternative method to extract all the references to geographic locations from the textual content using a named-entity tagger; the scope was then computed by assigning a different weight to each reference to disambiguate and rank them. Both [8] and [9] used an ontology to model geographic areas and relationships between locations to compute scopes by means of some measures on the semantic links between the references; moreover [9] put a specific emphasis on context computation via a graph ranking algorithm.

Particular attention to geographic terms disambiguation and false-positive avoidance was given in [5], where a gazetteer and confidence scores are used for each reference.

All the above techniques focus on computing and exploiting the geographical location in the broadest sense, allowing for more efficient information retrieval processes, thus offering more powerful tools than plain queries relying on simple *"keywords + location"* matches, instead leveraging on the geographical scope of a resource to rank results according to some proximity metrics.

However, it is noteworthy to say that ambiguities can lead to wrong classifications when no distinction is made between the *target* geography and the *source* geography or under the assumption that albeit being different concepts they always bear the same value. Most previous works strongly relied on the analysis of the web page textual content, thus focusing only on the former kind of context, i.e., target geography.

In this paper we focus on *source* geography, thus aiming at estimating the geographic *provenience* of a web page irrespective of its specific content which is not assumed to be always correlated with the source-geography. Interestingly enough, source-geography scope is always present and well defined, albeit difficult or impossible to determine, unlike the *target* one, because a web page may not contain any geographical references or may not be aimed at a specific target area, but it is nearly always created and published from a precise geographical region. However, source-geography is usually unrelated with the specific content or subject of a web page. Thus, source-geography context is more difficult to infer reliably.

In this paper we use well-known machine learning techniques to infer the source-geography of a page. More specifically, a classifier is learned on a training set of pre-labeled data and then used to perform online classification. Since harvesting and hand-labeling a properly sized training set would be impractical, we exploit the widespread adoption of localized Creative Commons licenses [10] (for the sake of simplicity, CC in the following).

The contributions of this paper are, thus, two-fold: first, we devise a way to build a proper training set, and secondly we propose a method to infer the source geography of a web page.

The rest of this paper is organized as follows: key ideas and the proposed technique are introduced in Section II, and results are presented in Section III, finally, conclusions are drawn in Section IV.

II. Proposed technique

The source geography scope tagging problem can easily be cast into a classification task where, given a web resource, one wants to infer the class to which it belongs, where the class here is the country where the resource was produced. We experimented with two simple classifiers, i.e., Naive Bayes [11] (due to its simplicity and low complexity) and Hidden Naive Bayes [12] (an improvement over the simpler Naive Bayes which takes into account interdependency across the features by conditioning on a latent unobservable variable).

A. Training set

As a first step, a set of features has to be chosen and collected to train a proper classifier on a set of pre-labeled data. However, harvesting a proper training set and hand-labeling each resource with the proper class is impractical and time-consuming due to the large number of classes (countries) and the need to have a sufficiently diverse collection of resources for each class.

To overcome this problem we propose an automatic and more efficient technique exploiting the page licensing information (if present).

CC licenses have gained momentum and are being adopted by content producers across the world, due to their simplicity and the wide range of readily available licensing possibilities. These copyright licenses provide a set of predefined options (and the corresponding legal code) to grant some rights to the public, allowing to share, reuse and remix creative works, also regulating commercial uses. They typically represent a convenient solution for non-professional contents creators, i.e., for some free and open contents based business models. Thus, many different kinds of digital objects have currently been put under a CC licensing scheme, ranging from entire blogs, and books, to music, film footage and paintings, a significant part of which has been published on the web.

Even more interestingly from our standpoint is the fact that CC licenses have been localized to the diverse jurisdictions of many countries; to date 51 country-specific versions of

	Top level domain	*geoIP*	*whois (IP addr.)*	*whois (FQDN)*	*language (n-grams)*	*lang*	*xml:lang*	*Cont. Lang. (HTML)*	*Cont. Lang. (HTTP)*	*Charset HTML*	*Charset HTTP*	**Source Geography**
Top level domain	**3.23**	0.86	0.84	1.60	0.74	0.33	0.21	0.09	0.04	0.09	0.06	1.12
geoIP	0.86	**2.97**	2.69	1.44	0.99	0.49	0.31	0.13	0.05	0.14	0.10	1.54
whois(IP addr.)	0.84	2.69	**3.11**	1.39	0.95	0.46	0.29	0.12	0.04	0.12	0.09	1.49
whois(FQDN)	1.60	1.44	1.39	**3.77**	1.21	0.54	0.33	0.14	0.04	0.14	0.10	1.94
language (n-grams)	0.74	0.99	0.95	1.21	**2.91**	0.72	0.42	0.18	0.04	0.12	0.07	1.89
lang	0.33	0.49	0.46	0.54	0.72	**2.21**	0.87	0.10	0.05	0.11	0.10	0.75
xml:lang	0.21	0.31	0.29	0.33	0.42	0.87	**1.49**	0.10	0.05	0.05	0.04	0.45
Cont. Lang. (HTML)	0.09	0.13	0.12	0.14	0.18	0.10	0.10	**0.65**	0.01	0.04	0.03	0.18
Cont. Lang. (HTTP)	0.04	0.05	0.04	0.04	0.04	0.05	0.05	0.01	**0.20**	0.01	0.01	0.05
Charset (HTML)	0.09	0.14	0.12	0.14	0.12	0.11	0.05	0.04	0.01	**1.34**	0.39	0.15
Charset (HTTP)	0.06	0.10	0.09	0.10	0.07	0.10	0.04	0.03	0.01	0.39	**1.20**	0.10
Source Geography	1.12	1.54	1.49	1.94	1.89	0.75	0.45	0.18	0.05	0.15	0.10	**3.90**

Table 1. Mutual Information (in bits) shared between each pairs of attributes, and between each attribute and each class label (which corresponds to the *source geography*.

the licenses [13] are available by volunteers's team. These countries comprise most of the main contributors of online contents and can be considered sufficient and significant for our application. Localized licenses cover almost 20% [14] of all the CC licensed contents (which currently amounts to more than 100 millions [15]), thus representing a valid source of information.

Under the reasonable assumption that a localized CC license is usually applied only to a content "*belonging*" to the country it refers to, it can be safely used to label freely available licensed content. Moreover, we assume that pages with CC-licensed contents, while diverse and heterogeneous, are not different (feature-wise) from pages with no reference to a CC license.

Thus, we developed a web spider to crawl the web collecting pages containing CC licensing information. This kind of information is usually encoded in the page by means of either RDF [16] expressed in XML and inserted into HTML comments, or via RDFa [17] as recommended by Creative Commons [18]. If no metadata are present it is still possible to simply check for back-links to the license deeds hosted on the CC website. Each time a web page is analyzed, it is labeled according to its license and, along with a number of specific features (irrespective of its licensing scheme) it is added to the training set.

B. Feature selection

The relevant features to infer source geography scope can be roughly divided into two categories, i.e., information about the server hosting the content and information extracted from the content itself.

The first set of features includes the physical location of the server hosting a particular page and can be obtained by checking its Fully Qualified Domain Name (FQDN), especially the top level domain part, and its IP address. When a page is crawled and analyzed, the WHOIS [19] protocol is used to obtain from the WHOIS official databases the information about the country of the site domain name and the corresponding IP address owners. Analogously, the IP address of the server hosting the page can be mapped to a country by means of the MaxMind GeoIP's APIs [20].

Class	*Number or records*
ARGENTINA	142
AUSTRALIA	679
BRAZIL	830
CANADA	244
CHINA	444
FRANCE	798
GERMANY	1310
ITALY	1274
JAPAN	866
MEXICO	113
SPAIN	2527
SWEDEN	202
SWITZERLAND	166
UNITED KINGDOM	495
UNITED STATES	1270
OTHER	856
TOTAL	12216

Table 2. Total number of unique records per country in the training set.

Furthermore, often some sort of load-balancing at the name server (DNS) is performed to distribute the load more fairly across different servers. As a consequence, the same name is resolved at each request to a (possibly) different list of IP addresses, which are ranked either at random or according to a sequential "round robin" policy over the set of available servers. Clients typically choose to use the first address of the list. Thus, we had to take into considerations all the IP addresses corresponding to a given FQDN, to map them to their geographic location and, in case they did not belong to the same country, consider the more frequent one in the set.

Some relevant information is also provided by certain features of the page content and the Uniform Resource Locator (URL), especially from the top level domain name if it matches a country code.

In fact the content itself, along with its language, are also very important to assess the source geography, although they may still lead to some ambiguity, especially for certain languages which are widely spoken in different countries. However, language is often strongly related to the information we seek. Thus, both the character set and an estimate of the language in which the page is written, along with explicit language declarations, if any, are considered as important features.

[21] describes how to express the language information with HTTP headers and into the HTML documents in many ways and with slightly different meanings. In fact, the intended audience language about the document as a whole, for high level processing purposes, is typically described the `Content-Language:` HTTP header. HTML may also include information about the text-processing language, as internal language declarations in the HTML metadata to specify the language for portions of the text and allow tools such as voice browsers and spell checkers to handle the content appropriately.

Moreover, language itself can be declared in various ways in documents written either in the HTML or the XHTML dialects, the latter being compliant to XML. In fact, while on one hand language can be specified by using the `lang` and `xml:lang` attributes of XHTML elements and is inherited by their descendants; if the attribute is used on the `<html>` tag it sets the language for the whole document, however the declaration can be overridden by a similar one in a descendant element. Typically, `lang` is used for HTML pages, `xml:lang` for XHTML pages used for derive an XML document, while both can be used in XHTML served as a text/HTML document, as in the Web scenario. On the other hand, it is also possible to use a meta element which sets the tag `http-equiv` to `Content-Language` or, albeit being less common, using the Dublin Core `language` element [22].

The `Content-Language` meta element and the HTTP header are specifically designed to express the language of the intended audience and they typically consists of set of values (i.e. "de, fr, it") but they can also bear information about the source geography.

Lastly, language information can be inferred by applying natural language processing techniques such as N-Gram based text categorization [23] to the textual content of the requested resource. N-Gram based text categorization extracts a text language profile which is then used to find the best match with a number of pre-calculated profiles of different languages.

Character encoding is usually present in the HTTP headers, more specifically in the `Content-Type` line [24] and should also always be specified, as recommended by the World Wide Web Consortium (W3C), into the `<head>` portion of HTML (or XHTML) documents.

Although language cannot be directly inferred from the character encoding, because there is not a one-to-one mapping, the choice of a specific encoding could give hints about the language and the region of provenience.

C. The classifier

It should be noted that none of the above features, if taken alone, can unequivocally determine the source geography scope of a web page, but all of them bear some information. This can be easily assessed by estimating the *mutual information* exchanged between class (i.e., the source geography scope) and each feature. Mutual information between attribute values A defined over alphabet $\mathcal{A}$ and classes C taking values in $\mathcal{C}$ is computed as:

$$I(A;C) = \sum_{a\in\mathcal{A}}\sum_{c\in\mathcal{C}} p_{AC}(a,c)\log\left(\frac{p_{AC}(a,c)}{p_A(a)p_C(c)}\right),$$

where $p_A = Pr\{A=a\}, a\in\mathcal{A}$ and $p_C = Pr\{C=c\}, c\in\mathcal{C}$ are, respectively the probability mass function of the discrete random variables A, C, while $p_{AC} = Pr\{A=a, C=c\}$ is the joint probability mass function of A and C. This value has to be compared with the entropy $H(C)$, which is a measure of the amount of information of C.

Entropy can be estimated by summing over the entire training set as follows:

$$H(C) = -\sum_{c\in\mathcal{C}} p_C(c)\log\left(p_C(c)\right),$$

where p_C is the probability mass function over the country of origin of the resources in our training set. Both entropy and mutual information here are expressed in bits, which means that we used base-2 logarithms.

Table 1 presents the mutual information between each possible attribute pair and between each attribute and class. Since mutual information is a symmetric operator the matrix is symmetric as well and the values on the diagonal are the entropies of the attributes (corresponding to the self mutual information), since $I(A;A) = H(A)$.

It is evident how the most informative attribute is the domain name owner country, whose mutual information with respect to the class only amounts to almost half the class's entropy. This can be explained by the fact that nowadays many websites are containers for contents generated by a great number of people from different countries, such as blog publishing systems.

Moreover, it should also be noted that part of the information brought to bear by the attributes is redundant, and the contributions of different features partly overlaps.

In addition, even lesser informative attributes are affected by similar problems: the country which hosts the server often has no relationship with the users who use it to publish contents, language metadata are often missing, certain top level domain are either too generic, such as the case of *.com* and *.org* or are misused (*.tv* and *.tk*), thus reducing their relationship with the source geography.

Lastly, Table 1 confirms that source geography can not be deterministically decided simply by looking at a single feature. Thus to obtain a robust and reliable estimate a probabilistic approach should be used, because no feature can completely discriminate across the classes.

Assume that P is a probability mass function, $F_1, F_2, ..., F_n$ are n features and an unseen record R is a vector $(f_1, f_2, ..., f_n)$, then the optimal probabilistic conditional model of class variable C for a classifier is:

$$P(C|F_1, F_2, ..., F_n),$$

which, by Bayes's theorem

$$P(C|F_1, F_2, ..., F_n) = \frac{P(C)P(F_1, F_2, ..., F_n|C)}{P(F_1, F_2, ..., F_n)}$$

leads to the Bayesian classifier Γ defined as:

$$\Gamma(R) = \arg\max_{c \in \mathcal{C}} \{P(c) \cdot P(f_1, f_2, ..., f_n|c)\}.$$

Learning the Bayesian model therefore implies the computation of the conditional probability for every possible combination pair of the features values, which, in general, leads to high computational cost; thus a simpler model called Naive Bayes, is adopted. The Naive Bayes approach assumes that all the features are conditionally independent, leading to a classifier defined as:

$$\Gamma(R) = \arg\max_{c \in \mathcal{C}} \left\{ P(c) \cdot \prod_{i=1}^{n} P(f_i|c) \right\}.$$

Obviously the assumption of conditional independence across the features is a rough approximation which deliberately ignores the mutual information *between* the features, i.e., their relative redundancy which, in our case, is always greater than zero as shown in Table 1. To overcome this problem we also experimented with the *Hidden Naive Bayes*[12] which relies on weaker assumptions about independence across the features by introducing a latent, hidden parent feature for each observable one.

The resulting classifier is thus defined as:

$$\Gamma(R) = \arg\max_{c \in \mathcal{C}} \left\{ P(c) \cdot \prod_{i=1}^{n} P(f_i|f_{hp_i}, c) \right\},$$

where

$$P(F_i|F_{hp_i}, C) = \sum_{j=1, j \neq i}^{n} W_{ij} \cdot (F_i|F_j, C), \qquad (1)$$

with $\sum_{j=1,j\neq i}^{n} W_{ij} = 1$, and where F_{hp_i} is the latent unobservable feature.

The hidden parents in Eq. (1) are defined by means of weighted one-dependence estimators, where the weights W_{ij} encode the *importance* of each attribute, thus playing a key role in the learning process. We chose to define weights as in [12], i.e., for any given two features i, j, the corresponding weight is defined as their mutual information normalized over the sum for each possible pair (self mutual information excluded):

$$W_{ij} = \frac{I(F_i; F_j|C)}{\sum_{j=1,j\neq i}^{n} I(F_i; F_j|C)},$$

where:

$$I(X; Y|Z) = \sum_{x,y,z} P(x, y, z) \log \frac{P(x, y|z)}{P(x|z)P(y|z)}.$$

Model	*Mean Accuracy*	*Standard deviation*
Naive Bayes	**79.395%**	0.951
Hidden Naive Bayes	**80.675%**	0.916

Table 3. Mean accuracy of the tested classifiers is presented along with the corresponding standard deviation.

III. Results

A set of about 1.5 million web pages from all over the world was collected by means of our crawler which also recorded the attributes needed to perform classification. However, most of these pages were redundant, often because they belonged to the same site, thus, the set was deeply pruned down to a subset of 12.216 pages not sharing the second level domain name (in order to avoid multiple samples like `bob.blog.com` and `alice.blog.com` with similar characteristics), except for web pages sharing some domain names but with a different class, i.e., which belong to different countries. Table 2 shows the number of records per country we collected this way.

Pruning however left many countries with very few samples, clearly insufficient to perform training properly, so we decided to limit the number of classes by lumping into one single class all the classes corresponding to countries with fewer than 100 distinct samples. We also noticed how the ratios between country samples number in this set became close to the ratio of the number of licensed contents per jurisdiction as estimated by the Creative Commons organization [15] and by some previous works [14]. This suggests that our content harvesting techniques did not introduce any particular bias into the distribution of the training samples.

Our algorithm was tested by means of 10 runs of a 10-fold cross validation, i.e., the training set [1] (after pruning and class lumping) was first divided into 10 almost equally-sized subset, which were in turn used as test sets for a model learnt on the remaining nine. After each run, records were randomly sorted. At the end of the whole evaluation process mean accuracy and standard deviation were computed, both for the Naive and the Hidden Naive Bayes classifiers. In Table 3 results for both methods are shown. The simpler Naive Bayes model achieved a mean accuracy of 79.395%, while Hidden Naive Bayes accuracy increased of 1.3%, up to a mean value of 80.675%. Interesting enough, the standard deviation is about 0.9 for both classifiers, showing that classification performance is quite stable across the training set.

Table 4 shows the confusion matrix after one fold classification. It is noticeable how most misclassifications belong to the UNITED STATES class. This happens because of the great number of contents outside the United States sharing many features typical of pages actually coming from the USA, such as being written in English language and hosted on a server located in the USA.

[1] The training set is available for download at: `http://nexa.polito.it/nexafiles/geoweb_tr_set.zip`

	BR	UK	AU	IT	CN	MX	FR	US	SE	DE	CH	JP	AR	CA	SP	*OTHER*
BR	**81**	0	1	0	0	1	0	0	0	0	0	0	0	1	0	4
UK	0	**23**	1	0	0	0	0	2	0	1	0	0	0	2	1	1
AU	0	0	**47**	0	0	0	1	1	0	0	0	1	0	0	0	0
IT	0	1	0	**101**	0	0	1	2	1	0	0	0	0	1	0	1
CN	0	0	0	0	**31**	0	1	0	0	0	0	0	0	0	1	3
MX	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
FR	0	0	1	1	0	0	**53**	2	0	1	1	1	0	2	0	2
US	7	17	30	2	9	4	3	**113**	6	2	1	4	1	7	6	35
SE	0	0	0	0	0	0	0	0	**20**	0	0	0	0	0	0	0
DE	0	2	0	1	0	1	3	2	1	**109**	4	0	0	0	0	2
CH	0	0	0	0	0	0	0	0	0	1	**12**	0	0	0	0	0
JP	0	0	0	0	4	0	0	0	0	0	0	**83**	0	0	0	2
AR	0	0	0	0	0	0	0	0	0	0	0	0	**4**	0	0	0
CA	0	0	0	0	1	0	0	3	0	0	0	0	0	**12**	0	0
SP	0	0	1	1	1	6	2	3	1	0	0	0	7	1	**254**	5
OTHER	0	0	1	0	0	0	2	0	0	2	0	0	0	0	0	**43**

Table 4. Confusion matrix obtained after one fold classification. The diagonal contains the number of correct classifications per country.

IV. Conclusions

In this paper we presented an algorithm to automatically classify web contents with their country of origin. This classification is made by means of a supervised learning algorithm which is used to build up a probabilistic model starting from a set of already labeled records. For this purpose, we also proposed a technique to build the training set automatically by exploiting Creative Commons licensed web pages in fact, the web was harvested and for each page found we extracted the nationality and a set of geographically meaningful features. Two different probabilistic models, a simpler Naive Bayes and a Hidden Naive Bayes models were trained and used to perform classification,

Results show that the Hidden Naive Bayes model successfully classified unlabeled contents with an accuracy of about (mean $\pm$ standard deviation) $81\% \pm .9$.

References

[1] "W3c semantic web interest group: Basic geo (wgs84 lat/long) vocabulary," http://www.w3.org/2003/01/geo/.

[2] A. Daviel and F. Kagi, "Geographic registration of HTML documents," http://geotags.com/geo/draft-daviel-html-geo-tag-07.txt, IETF Draft, July 2003.

[3] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf, "Dublin Core Metadata for Resource Discovery," http://www. ietf. org/rfc/rfc2413, Sept. 1998.

[4] K. McCurley, "Geospatial mapping and navigation of the web," in *Proceedings of the 10th international inproceedings on World Wide Web*, Hong Kong, 2001, pp. 221–229.

[5] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: geotagging web content," in *Proceedings of the 27th annual international ACM SIGIR inproceedings on Research and development in information retrieval*, 2004, pp. 273–280.

[6] O. Buyukokkten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar, "Exploiting geographical location information of web pages," in *Proceedings of WebDB-99, the 1999 ACM SIGMOD workshop on the web and databases*, 1999, pp. 91–96.

[7] J. Ding, L. Gravano, and N. Shivakumar, "Computing geographical scopes of web resources," in *Proceedings of the 26th International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc., 2000, pp. 545–556.

[8] C. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. Van Kreveld, and R. Weibel, "Spatial information retrieval and geographical ontologies an overview of the SPIRIT project," in *Proceedings of the 25th annual international ACM SIGIR inproceedings on Research and development in information retrieval*, 2002, pp. 387–388.

[9] M. Silva, B. Martins, M. Chaves, A. Afonso, and N. Cardoso, "Adding geographic scopes to web resources," *Computers, Environment and Urban Systems*, vol. 30, no. 4, pp. 378–399, 2006.

[10] "Creative Commons," http://creativecommons.org.

[11] H. Zhang, "The optimality of naive Bayes," in *Proceedings of the 17th Florida Artificial Intelligence Research Society Conference*. AAAI Press, 2004, pp. 562–567.

[12] H. Zhang, L. Jiang, and J. Su, "Hidden naive bayes," in *Proceedings of Canadian Artificial Intelligence Conference*. AAAI Press, 2005, pp. 432–441.

[13] "Creative Commons International," http://creativecommons.org/international.

[14] G. Cheliotis, A. Guglani, and G. Tayi, "Measuring the Commons: Quantifying Global Online Licensing Behavior," in *3rd Symposium on Statistical Challenges in E-Commerce Research*, University of Connecticut, Stamford, CT, May 2007.

[15] "Creative Commons Metrics," http://wiki.creativecommons.org/Metrics.

[16] G. Klyne, J. Carroll, and B. McBride, "Resource description framework (RDF): Concepts and abstract syntax," *W3C recommendation*, vol. 10, 2004.

[17] B. Adida and M. Birbeck, "RDFa Primer," http://www. w3. org/TR/xhtml-rdfa-primer, 2007.

[18] H. Abelson, B. Adida, M. Linksvayer, and N. Yergler, "ccREL: The Creative Commons Rights Expression Language," Creative Commons, Tech. Rep., May 2008.

[19] L. Daigle, "IETF RFC 3912," *Whois protocol specification*, 2004.

[20] MaxMind, LLC, "GeoIP API," http://www.maxmind.com/app/api.

[21] B. Using, "Internationalization Best Practices: Specifying Language in XHTML & HTML Content," http://www.w3.org/TR/i18n-html-tech-lang/.

[22] D. Core, "Dublin Core metadata element set, version 1.1: reference description," http://dublincore.org/documents/dces/, 1999.

[23] W. Cavnar and J. Trenkle, "N-gram-based text categorization," in *In Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, 1994, p. 4001.

[24] "Tutorial: Character sets & encodings in XHTML, HTML and CSS," http://www.w3.org/International/tutorials/tutorial-char-enc/.