Via Po, 53 – 10124 Torino (Italy)
Tel. (+39) 011 6704917 - Fax (+39) 011 6703859
URL: http://www.eblacenter.unito.it/

WORKING PAPER NEW SERIES

GEOGRAPHIC MAPPING OF CULTURAL COMMONS ON THE WEB

Bardone D., Carotti E. and J. C. De Martin

Dipartimento di Economia “S. Cognetti de Martiis”

International Centre for Research on the
Economics of Culture, Institutions, and Creativity
(EBLA)

Centro Studi Silvia Santagata (CSS)

Working paper No. 10/2010

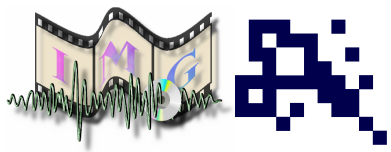
Università di Torino

Geographic Mapping Of Cultural Commons On The Web

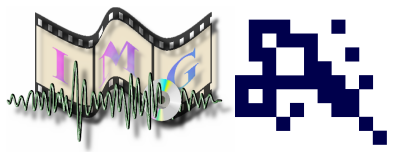
Davide Bardone, Elias S.G. Carotti, Juan Carlos De Martin
[davide.bardone|carotti|demartin]@polito.it

Internet Media Group
NEXA Center for Internet & Society
Dipartimento di Automatica e Informatica
Politecnico di Torino - Italy

First International
Cultural Commons Workshop

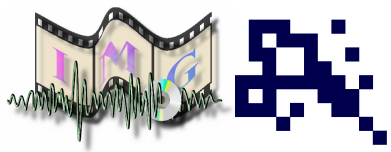


Outline



Web 2.0

- Web 2.0 main characteristics:
  - collaboration;
  - easy content sharing;
  - easy content creation;
  - interaction.

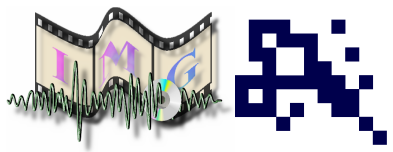


Web 2.0

- Web 2.0 main characteristics:
  - collaboration;
  - easy content sharing;
  - easy content creation;
  - interaction.
- Huge aggregator of diverse and heterogeneous user created contents.

Web 2.0

- The distributed and collaborative nature of the Web 2.0 allows for easier and spontaneous virtual communities creation.
  - expression or source of specific cultural commons.

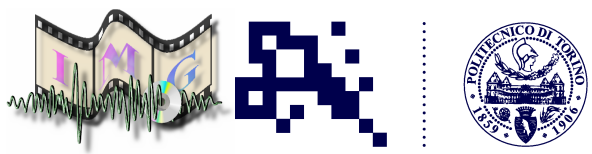


Web 2.0

- The distributed and collaborative nature of the Web 2.0 allows for easier and spontaneous virtual communities creation.
  - expression or source of specific cultural commons.
- Communities may refer to user belonging to a specific country or sharing some other characteristics.

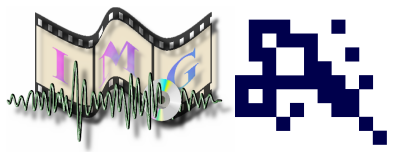
Web 2.0

- The distributed and collaborative nature of the Web 2.0 allows for easier and spontaneous virtual communities creation.
  - expression or source of specific cultural commons.
- Communities may refer to user belonging to a specific country or sharing some other characteristics.
- Geographically classify web contents could help:
  - looking for virtual expressions of location specific cultural commons;
  - make location-wise studies in existing content aggregators.



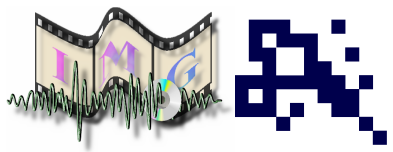
Web Geography

- Association of a web page with its geographical scope.



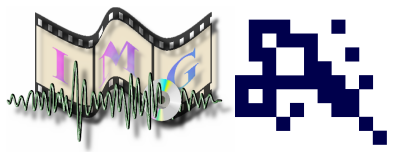
Web Geography

- Association of a web page with its geographical scope.
- Other applications:



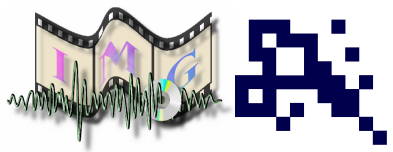
Web Geography

- Association of a web page with its geographical scope.
- Other applications:
  - information retrieval;



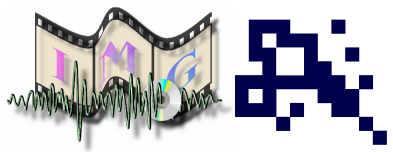
Web Geography

- Association of a web page with its geographical scope.
- Other applications:
  - information retrieval;
  - local market analysis;



Web Geography

- Association of a web page with its geographical scope.
- Other applications:
  - information retrieval;
  - local market analysis;
  - location wise statistics;



Web Geography

- Association of a web page with its geographical scope.
- Other applications:
  - information retrieval;
  - local market analysis;
  - location wise statistics;
  - measurement of web sites diffusion and importance.

Web Geography

- Association of a web page with its geographical scope.
- Other applications:
  - information retrieval;
  - local market analysis;
  - location wise statistics;
  - measurement of web sites diffusion and importance.
- Specific metadata are available but not widely used.

Geographical Scope

Ambiguous concept with different interpretations:

Geographical Scope

Ambiguous concept with different interpretations:

- target-geography or content-based geographic context:
  - the location the page is about;
  - the geographic area of the audience;

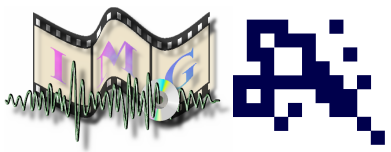
Geographical Scope

Ambiguous concept with different interpretations:

- target-geography or content-based geographic context:
  - the location the page is about;
  - the geographic area of the audience;
- source-geography or entity-based geographic context:
  - the location in which the page was created.

Existing techniques

Existing techniques make use of:



Existing techniques

Existing techniques make use of:

- heuristics on:
  - network related information (position of the physical server);
  - network administrators phone numbers;

Existing techniques

Existing techniques make use of:

- heuristics on:
  - network related information (position of the physical server);
  - network administrators phone numbers;
- information extraction algorithms on web pages' textual content:
  - phone numbers;
  - addresses;
  - location references;

Existing techniques

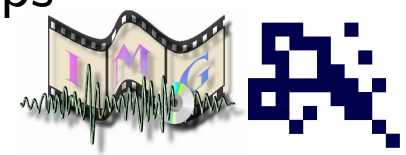
Existing techniques make use of:

- heuristics on:
  - network related information (position of the physical server);
  - network administrators phone numbers;
- information extraction algorithms on web pages' textual content:
  - phone numbers;
  - addresses;
  - location references;
- disambiguation algorithms for location references;

Existing techniques

Existing techniques make use of:

- heuristics on:
  - network related information (position of the physical server);
  - network administrators phone numbers;
- information extraction algorithms on web pages' textual content:
  - phone numbers;
  - addresses;
  - location references;
- disambiguation algorithms for location references;
- ontologies for considering spatial relationships between different scopes.

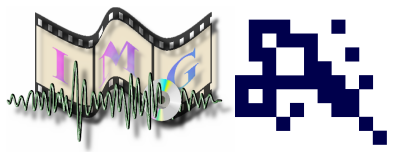


Observations

- No distinction between scope interpretation leads to ambiguities and misclassifications;

Observations

- No distinction between scope interpretation leads to ambiguities and misclassifications;
- most previous works strongly relied on the analysis of web pages textual content;



Observations

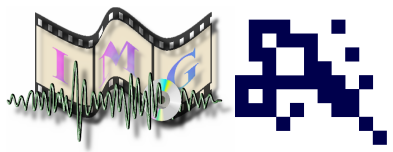
- No distinction between scope interpretation leads to ambiguities and misclassifications;
- most previous works strongly relied on the analysis of web pages textual content;
- geographical references are not always present;

Observations

- No distinction between scope interpretation leads to ambiguities and misclassifications;
- most previous works strongly relied on the analysis of web pages textual content;
- geographical references are not always present;
- it is very difficult to find a robust heuristic to reliably perform this kind of classification.

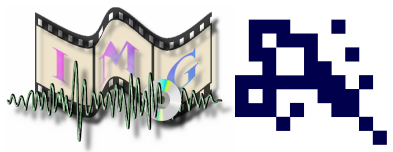
Proposed technique

- We focused on source-geography:
  - it is always present and well defined (no ambiguities);



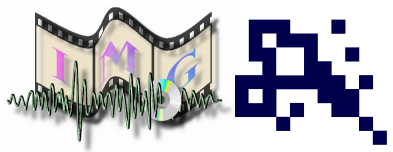
Proposed technique

- We focused on source-geography:
  - it is always present and well defined (no ambiguities);
- we can't rely on location references in text:



Proposed technique

- We focused on source-geography:
  - it is always present and well defined (no ambiguities);
- we can't rely on location references in text:
  - it is more difficult to make inference;



Proposed technique

- We focused on source-geography:
  - it is always present and well defined (no ambiguities);
- we can't rely on location references in text:
  - it is more difficult to make inference;
  - coarser granularity;

Proposed technique

- We focused on source-geography:
  - it is always present and well defined (no ambiguities);
- we can't rely on location references in text:
  - it is more difficult to make inference;
  - coarser granularity;
- we adopt a soft, probabilistic, machine learning approach:

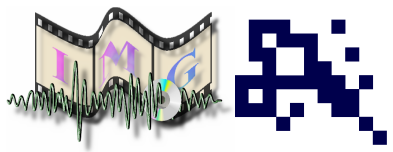
Proposed technique

- We focused on source-geography:
  - it is always present and well defined (no ambiguities);
- we can't rely on location references in text:
  - it is more difficult to make inference;
  - coarser granularity;
- we adopt a soft, probabilistic, machine learning approach:
  - a classifier is trained to classify pages according to the country of origin, given a set of geographically meaningful features.

Training set

We perform a supervised classification:

- we needed a set of pre-labeled data to train a proper classifier;
- collect and label by hand is impractical and time-consuming.



Training set

We perform a supervised classification:

- we needed a set of pre-labeled data to train a proper classifier;
- collect and label by hand is impractical and time-consuming.

We propose an automatic and more efficient technique:

- we exploited the Creative Commons licensing information;
- a web spider collected pages published with localized Creative Commons licenses;
- we made the reasonable assumption that CC licenses adapted for a specific jurisdiction are applied to content created in the country it refers to.

Training set

We perform a supervised classification:

- we needed a set of pre-labeled data to train a proper classifier;
- collect and label by hand is impractical and time-consuming.

We propose an automatic and more efficient technique:

- we exploited the Creative Commons licensing information;
- a web spider collected pages published with localized Creative Commons licenses;
- we made the reasonable assumption that CC licenses adapted for a specific jurisdiction are applied to content created in the country it refers to.

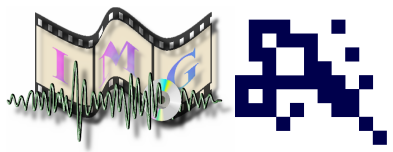
We considered 16 classes: 15 countries and an *OTHER* class.

Unique records per country

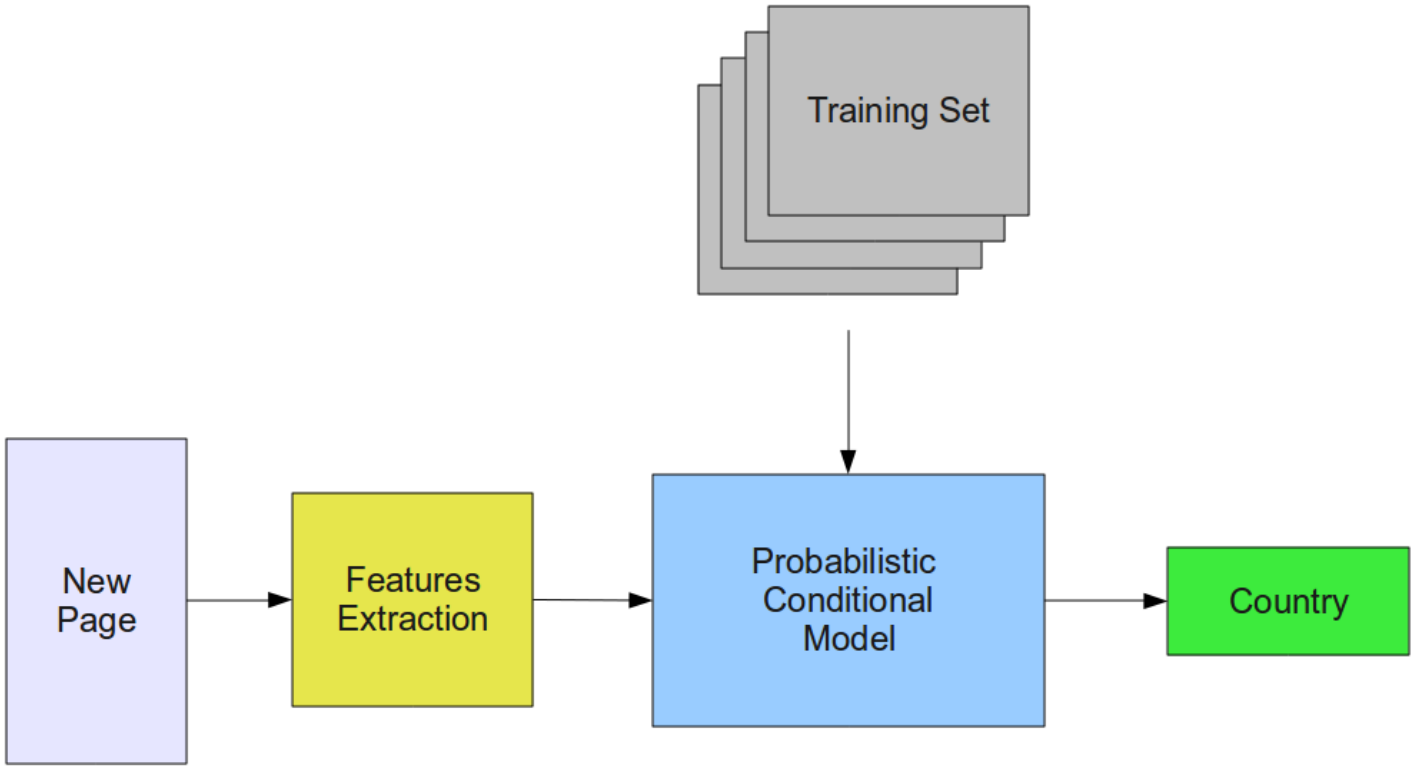
Class	Number or records
ARGENTINA	142
AUSTRALIA	679
BRAZIL	830
CANADA	244
CHINA	444
FRANCE	798
GERMANY	1310
ITALY	1274
JAPAN	866
MEXICO	113
SPAIN	2527
SWEDEN	202
SWITZERLAND	166
UNITED KINGDOM	495
UNITED STATES	1270
OTHER	856
TOTAL	12216

Selected Features

- Top level domain (e.g.: `.it`, `.fr`, `.com`);
- server location;
- IP address owner location;
- domain name owner location;
- textual content language;
- language declarations (in the HTML code and in the HTTP response);
- characters encoding.

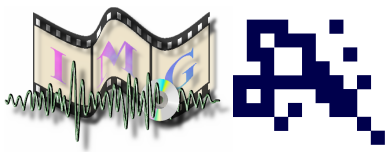


Classifier


Training Set
New Page
Features Extraction
Probabilistic Conditional Model
Country

Results

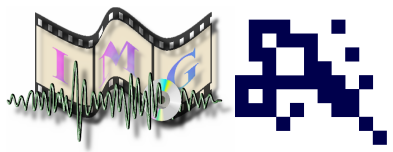
- We tested two different probabilistic models.
- Validation test results:



Results

- We tested two different probabilistic models.
- Validation test results:

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

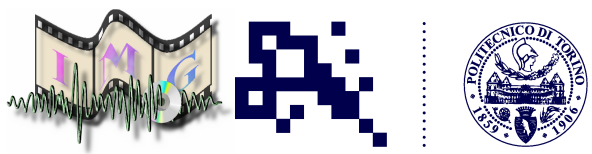


Results

- We tested two different probabilistic models.
- Validation test results:

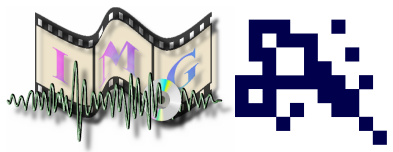
Model	*Mean Accuracy*	*Standard deviation*
Naive Bayes	**79.395%**	0.951
Hidden Naive Bayes	**80.675%**	0.916

- Most misclassifications belong to countries such as the United States, sharing many typical features values with other classes.



Conclusions

- Our algorithm could be used jointly with:
  - web crawlers,
  - semantic information extraction techniques
- to collect large databases of user generated contents or web pages to reveal cultural commons patterns, clusters and dynamics.



Thanks for your attention

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