Biodiversity Informatics: Mining Untapped Resources
February 3, 2016
696G Linguistics Seminar
Douglass 211

P. Bryan Heidorn
Director
University of Arizona
School of Information
Overview

- Show history of scholarly data management
- History of publication
- Need for digitization
- Using HMM for Museum Data
- Location parsing
- Discussion of Conditional Random Fields as an option
Biodiversity Information Diversity

1) Wrongly perceived as bioinformatics and two sets of base-pairs

2) Biodiversity = Data Complexity
   a) Requires new information theory and cyberinfrastructure
   b) Largely unrecognized as an interesting problem are in computer science
The problem

- Information is not in accessible
- Computer Science, Information Science and Technology has not addressed the problem
Where to find dark data

- Literature/Biodiversity Heritage Library
- Museum Specimens
- Field notes
- (Un)Experimental data sets
- Citizen Observations
What is dark data good for? Why NLP and text mining

- Ecological Niche Modeling
- Climate Change niche change prediction
- Taxonomic Name Resolution
- Literature Search Support
  - Taxonomic intelligence
  - Key-like – character searching
- Phenology and Phenology change
- Food-web / trophic level
Global Biodiversity Information Facility has 100s of millions of specimen records
Global Earth Observation System of Systems (GEOSS) and Historical Data

Historical and Current Data need to be in a form that allow for use and reuse.

- Unpublished observations of flowing time in Concord by Alfred Hosmer from 1888 to 1902
- Photographs of Flowers
- Blue Hill Observatory meteorological data

The problem with Museum Specimens

- >1 Billion Natural History Specimens
- Collected over 250 years / many languages
- No publishing standards
- Near infinite classes
  - Your high school teacher lied
- 6 min / label * 1B labels = 100M hours
- Saving 1 min = 16.7 Million hours
- $10/hr = $167,000,000
- 1/4790 of U.S. deregulation financial bailout
Automatic Metadata Extraction (Darwin Core) From Museum Specimen Labels

...<co> Curtis, </co><hdlc> North American Pl
<hdlc><cnl> No. </cnl><cn>503*</cn>
<gn> Polygala</gn><sp> ambigua,</sp><sa> Nutt.,</sa><val> var.</val>
<hb> Coral soil, </hb><lc> Cudjoe Key, South Florida.
</lc><col> Legit</col><co> A. H. Curtiss.</co><dt>February</dt>...

With Qin Wei, Univ of Illinois
Sample records

Herbarium of Yale University
Plants of San Luis, Peten, Guatemala

No: 301  Family: Boraginaceae
Scientific Name: Heliotropium
Mopan Mayan Name: u p'ot k'ix
Colloquial Spanish Name: moco de cabeza
Location: in pueblo (village)
Date: 29 May 1976
Comments: herbaceous plant with small yellow flowers

C. G. PRINGLE,
PLANTÆ MEXICANÆ.
1890.
—STATE OF SAN LUIS POTOSI—

3119 Acacia micantha, Benth.
Mountains, San Jose Pass.
13, July; 11, October.
Yale University Herbarium

YU.001300

Curtisb, North American ANTS,

No. 503* "^ Polygala ambigua, Nutt., var.

Coral soil, Cudjoe Key, South Florida.

Legit A. H. Curtiss.
Label Labels

- bc - barcode
- bt - barcode text
- cm - common/colloquial name
- cn - collection number
- co - collector
- cd - collection date
- fm - family name
- ft - footer info
Label Labels

- gn - genus name
- hd - header info
- in - infra name
- ina - infra name author
- lc - location
- pd - plant description
- sa - scientific name author
- sp - species name
Example Training Record

<?xml version="1.0" encoding="UTF-8"?>
<labeldata>
<bt>Yale University Herbarium</bt>
<ns>~r^-^"""" r-n------</ns><bc> YU.001300
</bc><co cc="Curtiss"> Curtisb, </co><hdlc cc="North American Plants"> North
American Pl
</hdlc><ns>C^o.nr r^-n
ANTS, </ns>
<cnl> No. </cnl><cn> 503* </cn><ns>"^</ns>
<gn> Polygala</gn><sp> ambigna,</sp><sa> Nntt.,</sa><val> var.</val>
<hb> Coral soil,</hb><lc> Cudjoe Key, South Florida.</lc>
</lc><col> Legit</col><co> A. H. Curtiss.</co>
</labeldata>
Structure of a simple HMM

HD: Header part
FM: Family
GN/SP: Genus, Species
LC/HB: Locality, Habitat
PD: Description
CO/FT: Citation part, Footer
Supervised Learning Framework

**Training Phase**

- **Unclassified Labels**
  - Human Editing
  - Gold Classified Labels
  - Machine Learner
  - Trained Model

**Application Phase**

- **Unclassified Labels**
- Segmentation
- Segmented Text
- Machine Classifier
- Silver Classified Labels
Herbis Experimental Data

- 295 marked up records
- 74 label states
- 5-fold cross-validation
Performances of NB and HMM
Element Identifiers

<table>
<thead>
<tr>
<th>Element</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curtis,</td>
<td>North American Plants, No. 503</td>
</tr>
<tr>
<td>Polygala ambigua, Nutt., var.</td>
<td></td>
</tr>
<tr>
<td>Coral soil, Cudjoe Key, South Florida.</td>
<td></td>
</tr>
<tr>
<td>Legit A. H. Curtis.</td>
<td></td>
</tr>
<tr>
<td>February.</td>
<td></td>
</tr>
<tr>
<td>Collected by Pierre Venturi, Yale Department of Anthropology</td>
<td></td>
</tr>
<tr>
<td>small yellow flowers</td>
<td>(det. L. Brown)</td>
</tr>
</tbody>
</table>
Improved Performance With Field Element Identifiers
Polygala ambigu, Nutt., var.

Coral soil, Cudjoe Key, South Florida.

Legit A. H. Curtiss.

February.
Learning w/ pre categorization

Gold Labels

Categorization

Class 1 Labels → Machine Learner → Model 1
Class 2 Labels → Machine Learner → Model 2
Class n Labels → Machine Learner → Model n

Unclassified Labels

Categorization

Class 1 Labels → Machine Classification → Classified Labels
Class 2 Labels → Machine Classification → Classified Labels
Class n Labels → Machine Classification → Classified Labels
FIG. 5. Improved Performance of Specialist Model

Specialist100 Curtiss VS 100 General

Percentage

0

100

0

200

Iterations

Specialist

Random

FIG. 5. Improved Performance of Specialist Model
Conditional Random Field

- Conditional random fields (CRFs) are a class of statistical modelling method often applied in pattern recognition and machine learning, where they are used for structured prediction.
Machine Learning in BioGeomancer’s Locality Specification

P. Bryan Heidorn\textsuperscript{1}, Hong Zhang\textsuperscript{1}, Eugene Chung\textsuperscript{2} and BGWG
\textsuperscript{1}Graduate School of Library and Information Science, \textsuperscript{2}Linguistics, University of Illinois

- Worldwide collaboration of natural history and geospatial data experts
- Maximize the quality and quantity of biodiversity data that can be mapped
- Support of scientific research, planning, conservation, and management
- Promotes discussion, manages geospatial data and data standards, and develops software tools in support of this mission
Participants
# Example Locality Types

<table>
<thead>
<tr>
<th>Record #</th>
<th>Specification of Location</th>
<th>Locality Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>dario 7 mi wnw of; RIO VIEJO</td>
<td>FOH; F</td>
</tr>
<tr>
<td>86</td>
<td>near Aleutian Islands; S of Amukta Pass</td>
<td>NF; FH</td>
</tr>
<tr>
<td>100</td>
<td>INDIAN CREEK, 11 MI. W HWY 160</td>
<td>P; POH</td>
</tr>
<tr>
<td>109</td>
<td>TIESMA RD, 1.5 MI NW EDGEWATER; OFF LAKE MICHIGAN R</td>
<td>P; FOH; NP</td>
</tr>
<tr>
<td>160</td>
<td>WALTMAN, 9 MI N, 2.5 MI W OF</td>
<td>FOO</td>
</tr>
<tr>
<td>181</td>
<td>0.4 mi N Collinston on LA 138</td>
<td>FPOH</td>
</tr>
<tr>
<td>200</td>
<td>Seward Peninsula; vic. Bluff, S coast</td>
<td>F; NF; FS</td>
</tr>
</tbody>
</table>
FRAMES

- JOH: offset from a junction at heading
  - e.g. 0.5 mi. W Sandhill and Hagadorn Roads

  [FEATURE [CITY = Sandhill]]

  [FEATURE [ROAD=Hagadorn Roads]]

  OFFSET VALUE = 0.5
  DIRECTION = W
  UNIT = mile

  JUNCITON
    [FEATURE [CITY = Sandhill]]
    [FEATURE [ROAD=Hagadorn Roads]]
Xiaoya Tang and P. Bryan Heidorn

Different vocabularies in queries and documents

User query

Long leaves

Description of leaf Length in texts

…… Leaves 20–75, many-ranked, spreading and recurved, not twisted, gray-green (rarely variegated with linear cream stripes), to 1 m × 1.5–3.5 cm, …….. Inflorescences: …… spikes very laxly 6–11-flowered, erect to spreading, 2–3-pinnate, …….
Leaf blade obovate to nearly orbiculate, 3--9 × 3--8 cm, leathery, base obtuse to broadly cuneate, margins flat, coarsely and often irregularly doubly serrate to nearly dentate, ...

Pattern:: * <PartBlade> ' ' <leafShape> * ( <leafShape> ) ',' *
Output:: leaf {leafShape $1}

Pattern:: * <PartBlade> * ', ' ( <Range> ' ' * <LengUnit> ) * <PartBase>
Output:: leaf {bladeDimension $1}

Leaf Shape obovate
Leaf Shape orbiculate
Blade Dimension 3–9 × 3–8 cm

PartBlade:
Leaf blade
Blades
blade

......

User log analysis
## Results – System Performance

<table>
<thead>
<tr>
<th>Group</th>
<th>NT</th>
<th>NTH</th>
<th>TSR</th>
<th>SSR</th>
<th>NSST</th>
<th>TST</th>
<th>NDVST</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEARFA</td>
<td>6.75</td>
<td>8.078</td>
<td>0.860</td>
<td>0.210</td>
<td>4.779</td>
<td>338.8</td>
<td>11.16</td>
</tr>
<tr>
<td>SEARF</td>
<td>4.50</td>
<td>3.598</td>
<td>0.568</td>
<td>0.053</td>
<td>9.584</td>
<td>435.2</td>
<td>14.75</td>
</tr>
<tr>
<td>Sig.(ANOVA)</td>
<td>0.005</td>
<td>0.005</td>
<td>0.000</td>
<td>0.011</td>
<td>0.000</td>
<td>0.72</td>
<td>0.162</td>
</tr>
</tbody>
</table>

- **NT**: number of tasks accomplished in total
- **NTH**: number of tasks accomplished per hour
- **TSR**: task success rate
- **SSR**: search success rate
- **NSST**: number of searches to accomplish a task
- **TST**: time spent to accomplish a task
- **NDVST**: number of documents viewed to accomplish a task
"Ridges about 2 miles west of inlet between Peters Lake and Schraders Lake - Drainage Canal, steep north-facing slope. ALASKA: Schraders Lake - Peters Lake area, just northwest of Mount Chamberline, Franklin Mountains, Brooks Range, approx 69 22 N. Lat., 145 03 W Long., about 3000 ft. altitude."
Type: Relation
Location1: Peters Lake
Location2: Schraders Lake
Verification: inlet
Resolution: ______________
Features (1)

- ADDR: Street address
- ADM: Administrative unit
- F: Feature. Anything that could potentially be found in gazetteer
- FS: Subdivision of a feature. “~ part of Feature”
- J: Junction. any intersection of linear feature
- NF/NJ: Near Feature/ Near Junction
Features (2)

- **P**: Path is a linear feature such as a road, trail, boundary or river. A description with a path followed by an offset from feature at a heading should be calculated as a clause of the type rather than as the intersection of a path and a clause.
- **POM**: Path Offset Marker
  - e.g. Mile 49.5 Sterling Hwy.
- **PS**: Subdivision of Path
- **TRS**: Township, Range, Section
- **TRSS**: Township, Range, Section Subdivision
Coordinates

- LL: Latitude and Longitude coordinate
- UTM: Universal Transverse Mercator coordinates
Offsets

- +2P: orthogonal offsets from two paths
- FO: Offsets form a features, no heading
- FOH: Offset form a features at heading
- FO+: orthogonal offsets from a features
- JOH: offset from a junction at heading
- FPOH: offset from a features at heading along a path
- PO: offset along a path, no feature or heading
The following locality types are not found:
- Coordinates: OGS, UTM
- Offsets: +2P
A frame defines general properties hold among a class of objects, called frame instances. Frames contain slots, roughly, attributes.

Some frames are complex in that they refer to sequences of transitions, each of which can itself be separately described as a frame.
Each locality type can be served as a sub-frame.
A sub-frame can be combined with other sub-frames. For example;

[ FEATURE [ CITY = Cansas ] ]
[PATH [ PLACE = Rio Higueron] ]

FOH; P 10 MI SW CANAS; RIO HIGUERON
OFFSET

\[
\begin{align*}
\text{OFFSET} & \quad \text{VALUE} = 10 \\
\text{DIRECTION} & = \text{sw} \\
\text{UNIT} & = \text{mile} \\
\text{HEADING} & \quad \text{FEATURE} [ \text{CITY} = \text{Kansas} ] \\
\text{PATH} & \quad \text{PLACE} = \text{Rio Higueron} 
\end{align*}
\]
JOH : offset from a junction at heading

e.g. 0.5 mi. W Sandhill and Hagadorn Roads

[ FEATURE [ CITY = Sandhill ]]

[ FEATURE [ ROAD= Hagadorn Roads ]]

OFFSET VALUE = 0.5
DIRECTION = W
UNIT = mile

JUNCTION [ FEATURE [ CITY = Sandhill ]]
[ FEATURE [ ROAD= Hagadorn Roads ]]
Structure equals flexibility