Ethnea -- an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database

Vetle I. Torvik and Sneha Agarwal
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Abstract: We present a nearest neighbor approach to ethnicity classification. Given an author name, all of its instances (or the most similar ones) in PubMed are identified and coupled with their respective country of affiliation, and then probabilistically mapped to a set of 26 predefined ethnicities. The dominant ethnicity (or pair of ethnicities) is assigned as the class. The predictions are also used to upgrade Genni (Smith, Singh, and Torvik, 2013) to provide ethnicity-specific gender predictions for cases like Italian vs. English Andrea, Turkish vs. Korean Bora, Israeli vs. Nordic Eli, and Slavic vs. Japanese Renko. Ethnea and Genni 2.0 are available at http://abel.lis.illinois.edu

Methods: Existing approaches have focused on machine learning techniques that extract features of names, with known ethnicities, harvested from online sources (e.g., TextMap: Ambekar et al. 2009, and EthnicSeer: Treeratpituk and Giles 2012). TextMap provides for hierarchical classification with 12 leaves, while EthnicSeer has a flat set of 12 slightly different classes (e.g., it excludes Jewish, Nordic, and African, and includes Chinese, Korean, and Vietnamese in place of EastAsian). Our approach differs in several respects. First, it is instance-based. That is, no machine training or feature selection occurs, we just perform a look-up of author name instances previously geocoded and mapped to countries worldwide (Torvik, 2015). A temporally weighted multiclass logistic regression model then probabilistically maps the country distribution to ethnicities, reducing the undesirable effects of outliers and highly unbalanced classes. In order to enable fast partial name matching, all 3- and 4-character n-grams of each author name was indexed using MySQL + Sphinx which has rankers that allow for higher weighting of e.g., name-endings. Partial matching only kicks in when the name under question occurs fewer than 100 times in our database. Instance-based classifiers are often more capable than feature-based classifiers at capturing highly non-linear classification boundaries but they rely more heavily on a large, dense set of instances. Our database has tens of millions of author name instances distributed across 200+ countries over 20+ years. Second, we picked the 26 ethnic classes (see Table 2) to be as specific possible, yet separable, and to broadly cover the ethnicities observed with a significant frequency in PubMed. As a result, some countries were pooled regionally (e.g., non-Arab African countries map to African), and countries with no single super-majority were mapped to multiple classes (e.g., Canada maps both to French and English).

Results: We have pre-computed predictions for millions of last names (and first names, separately) observed in a variety of bibliographic databases. This enables fast lookup of probabilities for a first + last name pair and combining the two respective probabilities into one. Table 1 compares the predictions of Ethnea to TextMap and EthnicSeer on a small sample of names taken from DBLP. The first 11 were identified by Wu et al. (2014) as the most prolific of their ethnic class. Table 2 compares the rate of agreement between Ethnea and EthnicSeer on the names of all 4.7M authors with first and last names in the Author-ity 2009 dataset (Torvik and Smalheiser, 2009). EthnicSeer agrees with Ethnea for 78% of cases, if we set aside the ~10% cases mapped to classes EthnicSeer does not explicitly capture (Nordic, Dutch, Turkish, etc.) Ethnea provides a contemporary reflection of ethnicity, perhaps not surprisingly, leaning more towards nationality rather than distant ancestry, compared to TextMap and EthnicSeer. It also captures dual ethnicities which are not unusual e.g., due to marriage or migration and assimilation.

Table 1. Illustrative examples: a sample of DBLP authors.

Name	Ethnea	TextMap	EthnicSeer
Kang Shin	KOREAN	EastAsian	KOR
Scott Shenker	ENGLISH	British	JAP
Philip Yu	CHI-ENG	EastAsian	CHI
Anil Jain	INDIAN	Indian	IND
Lotfi Zadeh	ARAB	Muslim	ARA
Tomaso Poggio	ITALIAN	Italian	ITA
Robert Tarjan	HUNGARIAN	African	RUS
Hector Garcia-Molina	HISPANIC	Italian	SPA
Terrence Sejnowski	ENGLISH	EastEuropean	FRN
Herbert Simon	GERMAN	Jewish	GER
Ian Foster	ENGLISH	British	ENG
Vetle Torvik	NORDIC	EastEuropean	RUS
Evangelos Triantaphyllou	GREEK	Indian	FRN
Pucktada Treeratpituk	THAI	British	IND

TextMap: http://www.textmap.com/ethnicity/
EthnicSeer: http://singularity.ist.psu.edu/ethnicity
Ethnea: http://abel.lis.illinois.edu/cgi-bin/ethnea/search.py

Table 2. Distribution of 4.7M authors in the Author-ity 2009 dataset.

Ethnea	Proportion	EthnicSeer	Agreement
ENGLISH	31.52%	27.82%	76%
HISPANIC	9.99%	6.85%	59%
CHINESE	9.30%	9.87%	94%
GERMAN	8.57%	16.27%	84%
JAPANESE	8.16%	8.56%	98%
FRENCH	5.55%	8.49%	69%
ITALIAN	4.44%	7.38%	86%
SLAV	4.09%	3.88%	48%
INDIAN	3.25%	5.79%	94%
ARAB	2.99%	3.13%	68%
KOREAN	1.65%	1.73%	87%
VIETNAMESE	0.08%	0.23%	93%
SUBTOTAL	**89.58%**	**100.00%**	**78%**
NORDIC	3.29%	GER	48%
DUTCH	2.10%	GER	57%
TURKISH	1.18%	GER	32%
ISRAELI	0.97%	RUS	24%
GREEK	0.86%	FRN	27%
AFRICAN	0.56%	FRN	20%
HUNGARIAN	0.42%	GER	34%
THAI	0.41%	IND	57%
ROMANIAN	0.19%	ITA	32%
BALTIC	0.10%	RUS	27%
INDONESIAN	0.03%	IND	35%
CARIBBEAN	0.007%	IND	44%
MONGOLIAN	0.003%	IND	31%
POLYNESIAN	0.001%	IND/GER	35%
UNKNOWN	0.29%		

Acknowledgements: NSF 1348742 and NIH P01AG039347

References:

Ambekar A, Ward C, Mohammed J, Male S, Skiena S (2009). Name-ethnicity classification from open sources. In Proceedings of the 15th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (pp. 49-58). KDD '09. Paris, France

Smith BN, Singh M, Torvik VI (2013). A search engine approach to estimating temporal changes in gender orientation of first names. Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 199-208). JCDL '13. Indianapolis, IN, USA.

Torvik VI (2015). MapAffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide. D-Lib Magazine, 21 (11/12).

Torvik VI, Smalheiser NR (2009). Author name disambiguation in MEDLINE. ACM Transactions on Knowledge Discovery from Data 2009, 3(3):11.

Treeratpituk P, Giles CL (2012). Name-Ethnicity Classification and Ethnicity-Sensitive Name Matching. Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (pp. 1141-1147). AAAI-12. Toronto, ON, Canada.

Wu Z, Yuan D, Treeratpituk P, Giles CL (2014). Science and Ethnicity: How Ethnicities Shape the Evolution of Computer Science Research Community. http://arxiv.org/abs/1411.1129

Presented at the *International Symposium on Science of Science March 22-23, 2016 - Library of Congress, Washington DC, USA*