Identifying Author Heritage Using Surname Data

Maria Karaulova[1,2], Abdullah Gök[1] and Philip Shapira[1,3]

[1] Manchester Institute of Innovation Research (MIoIR), University of Manchester, UK
[2] Institute of Management, Scuola Superiore Sant'Anna, Italy
[3] School of Public Policy, Georgia Institute of Technology (GaTech), USA

Corresponding author: mkaraulova@gmail.com

Background and purpose

This research paper proposes a novel method to identify ethnic or national heritage of authors based on the morphology of their surnames. Ethnic approximation of authors in publication and patenting databases has received some attention in recent years, especially with the purpose to investigate ethnic patenting and publishing, and explore effects of immigration on innovative development (Kerr and Lincoln, 2010).

First name of authors is found to be a reliable indicator of gender (Meng and Shapira, 2010), while dealing with Chinese surname data is a widely recognised problem in bibliometric research that attracts multiple solutions (Tang and Walsh, 2010). However, similar studies for Slavic surnames are virtually non-existent. We develop this line of work by using the morphology of surnames as a key component of information retrieval in large imbalanced datasets (Chawla, 2005). We argue that surname morphology can serve as a reliable approximation of ethnic or national heritage of researchers and demonstrate this by developing a 2-step search procedure for post-Soviet surname data retrieval in nanotechnology publication dataset.

Summary of methods

The source of data for surname-based information retrieval is the particular structure of Russian surnames, namely, their patronymic suffix (Unbegaun, 1972). The most typical *–ov* format has been used in linguistics as a morphological marker of a Russian surname. As post-Soviet space demonstrates heterogeneity of surnames, and, in particular, 'foreign' surnames are overrepresented among Russian scientists, multiple scenarios were tested to develop the optimal configuration of surnames search terms.

A list of exclusion terms was then developed based on commonly met mistakes in author data retrieval. In addition to surname data, given name data field was utilised to identify names with Russian and post-Soviet heritage. A two-step procedure information retrieval format benefits from flexibility to balance recall and precision of the search. The results were validated by manually checking ORCHID data of a randomly selected set of authors.

The data for this research was collected from the Web of Science using nanotechnology search query (Arora et al., 2013; Porter et al., 2008). It was then cleaned and grouped in the VantagePoint software. The seven possible scenarios were tried in a testing set that was composed from the original dataset using undersampling (Liu et al., 2009). Ultimately, the search aims to maximise the weighted average of precision and recall by combining Boolean and string inclusion and exclusion terms.

Findings

In case of Russian surnames, the simplest rule that uses only two most popular patronymic suffixes returns about 80% relevant results. Scenarios that add combinations of other endings and full exception surnames increase the recall rate of the method to 0.95. As the second step, a combination of Boolean string search and full exception names exclude false positive records and increase the precision of the search up to 0.98.

The consistency of Russian heritage surname retrieval is maintained overall, with some national fluctuation. In the countries where the majority of Russian heritage surnames are identified (USA, UK, Germany, France), errors are negligible. In several countries where a minority of Russian surnames were identified (Czech Republic), errors persistently arise. This limitation arises only in testing retrieval results for these specific countries and does not affect broader outcomes of the search.

Discussion and Conclusions

The initial suggestion that surname data can be used for information retrieval in imbalanced datasets, at least in the case of Russian surnames, was found as valid. The findings of this research suggest that surname data can be used to identify communities of scientists or inventors based on shared country of origin (national – or ethnic, in mononational countries).

The method developed and elaborated in this paper is a robust tool that can be used to solve a variety of tasks. Most commonly, tasks related to the structure and composition of co-authored publications can be analysed, as they usually rely on country affiliation of authors. More generally, the use of surname data can be applied to all research problems of transnational research networks in international collaboration, or the effect of uneven coauthorship balances. It also contributes to improving solutions of the classic name disambiguation problem and can be used with little variation for identifying other Eastern European diasporas abroad, such as Czech, Bulgarian authors.

References

Arora, S.K., Porter, A.L., Youtie, J., Shapira, P., 2013. Capturing new developments in an emerging technology: an updated search strategy for identifying nanotechnology research outputs. Scientometrics 95, 351–370. doi:10.1007/s11192-012-0903-6

Chawla, N.V., 2005. Data Mining for Imbalanced Datasets: An Overview, in: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer US, pp. 853–867.

Freeman, R.B., Huang, W., 2014. Collaborating With People Like Me: Ethnic co-authorship within the US (Working Paper No. 19905). National Bureau of Economic Research.

Kerr, W.R., Lincoln, W.F., 2010. The supply side of innovation: H-1B visa reforms and US ethnic invention. National Bureau of Economic Research.

Kissin, I., 2011. A surname-based bibliometric indicator: publications in biomedical journal. Scientometrics 89, 273–280. doi:10.1007/s11192-011-0437-3

Lewison, G., 2001. The quantity and quality of female researchers: A bibliometric study of Iceland. Scientometrics 52, 29–43. doi:10.1023/A:1012794810883

Liu, X.-Y., Wu, J., Zhou, Z.-H., 2009. Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39, 539–550. doi:10.1109/TSMCB.2008.2007853

Meng, Y., Shapira, P., 2010. Women and Patenting in Nanotechnology: Scale, Scope and Equity, in: Cozzens, S.E., Wetmore, J. (Eds.), Nanotechnology and the Challenges of Equity, Equality and Development, Yearbook of Nanotechnology in Society. Springer Netherlands, pp. 23–46.

Porter, A.L., Youtie, J., Shapira, P., Schoeneck, D.J., 2008. Refining search terms for nanotechnology. Journal of Nanoparticle Research 10, 715–728. doi:10.1007/s11051-007-9266-y

Robinson-Garcia, N., Noyons, E., Costas, R., 2015. Can we track the geography of surnames based on bibliographic data? Presented at the 15th International Conference on Scientometrics and Informetrics, Istanbul, Turkey.

Tang, L., Walsh, J., 2010. Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. Scientometrics 84, 763–784. doi:10.1007/s11192-010-0196-6

Unbegaun, B.O., 1972. Russian surnames / B.O. Unbegaun. Clarendon Press, Oxford.