Personal Names Popularity Estimation and its Application to Record Linkage*

Ksenia Zhagorina[1], Pavel Braslavski[2,3], and Vladimir Gusev[2]

[1] Yandex, Yekaterinburg Ksenia.Zhagorina@yandex.ru
[2] Ural Federal University, Yekaterinburg
{Pavel.Braslavsky,Vladimir.Gusev}@urfu.ru
[3] JetBrains Research, Saint Petersburg

Abstract. In this study, we investigate several statistical techniques for personal name popularity estimation and perform a record linkage experiment guided by name popularity estimates. The results show that name popularity can leverage personal name matching in databases and be of interest for many other domains.

Keywords: personal name matching · record linkage · name distribution

1 Introduction

Record linkage – the task of matching records referring to the same real-world entity – is a well-studied field within database technology. The task arises when several databases are merged or one is interested in linking duplicate records within a single database. Records referring to people are the most common objects of linkage task. Our study is motivated by an applied record linkage task in a large database, where occurrences of personal names are accompanied with no or only scarce additional information. Under these circumstances, name popularity estimates serve as the main signal for record matching.

Knowing an estimate of people bearing a particular name is beneficial not only for record linkage, but also for social network analysis, people search, information security, and information extraction. Unfortunately, accurate name popularity estimation based on limited number of observations is a hard task. Even very large collections contain many unique names – names are a good example of *large number of rare events (LNRE)* distributions. Therefore, maximum likelihood estimates based even on large name samples are poor predictors, since there are always many unseen names. To address this issue we employ several smoothing techniques that redistribute probability mass from already seen names towards yet unseen ones. Moreover, we use LNRE models to estimate the number of unique names and use this estimate as a smoothing parameter.

* The work was carried out while authors were at Kontur Labs, the research department of SKB Kontur, https://kontur.ru/eng/. The authors benefit from the Russian Ministry of Education and Science, project no. 1.3253.2017, and the Competitiveness Enhancement Program of Ural Federal University.

In our study we used a large dataset of open government data. We conducted two experiments: 1) name popularity estimation and 2) record linkage guided solely by the name popularity estimates. We performed evaluation both for name triples (first, middle, and last) and doubles (first and last). Our results suggest that theoretically informed approaches outperform simple heuristics. The main contribution of our study is a thorough comparative evaluation of several statistical techniques applied to the name popularity estimation task on a sizable dataset. The study provides guidance for choosing the most appropriate model depending on available data, task, and performance requirements.

Related work. Our study is close to personal name matching [6], a special case of *record linkage* – the task of matching records referring to the same real-world person in the presence of errors, spelling variants, omissions, abbreviation, etc. Most name matching method rely on pre-defined or machine-learned similarity measures for field values and tuples, see [7]. The main difference of our study is that we deal with *identical* names and no additional fields. Moreover, we do not adjust our methods to a particular database; we rather aim at modeling name popularity at a global scale. As such, name popularity models can deliver additional evidence for record linkage tasks applied to different databases and in case of scarce additional information. The advent and proliferation of online social networks had a powerful impact on quantitative research on names, as name is often the only available information about the user. There is a series of studies that derive ethnicity [4,16] and gender [2] from names in social network profiles. Perito et al. [17] and Liu et al. [15] introduce the problem of linking user profiles belonging to the same physical person between online social networks based solely on the uniqueness of usernames.

Smoothing techniques we employ in the study have been actively developed within statistical language modeling [12,5]. Khmaladze [14] introduced the notion of *LNRE* distributions and studied their statistical properties. Baayen [1] and Evert [8] elaborated the models for a better fitting of frequency distributions of words in large corpora, with special attention to *hapax legomena* (words with frequency 1). We use LNRE models for a more accurate choice of smoothing parameters in several evaluated methods.

2 Data

In our study we experiment with a dataset that originates from the *Russian registry of legal entities*[4]. There is a many-to-many relationship between persons and companies: each legal entity is associated with one or more persons – managers and/or founders; each real-word person can be associated with several companies. The registry contains about 32 million name mentions. Full names in Russian official documents are triples comprising of first, middle (patronymic), and last names, for example, *Alexander Sergeyevich Pushkin*.

A subset of records contains persons' taxpayer identification numbers (TINs) that can be used as a key. In the rest of the paper we focus on about 20.6 million

[4] `http://egrul.nalog.ru/`

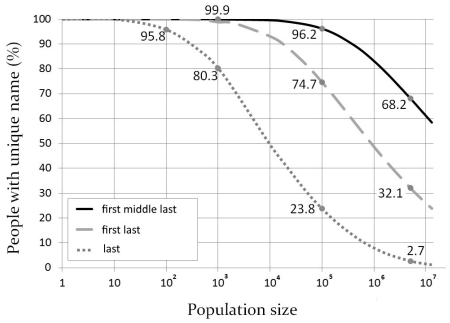


Fig. 1: Share of unique names depending on population size.

records containing both TIN and full name that refer to about 13.4 million real persons, which constitutes about one tenth of the entire Russian population.

First, middle, and last names taken separately or as full names are a good example of *LNRE* regime: the majority of names occur only once, while a small number of combinations are relatively common. Expectedly, last names tend to be more rare than first names and patronymics (the latter are derivatives from male first names). Figure 1 shows proportions of unique name combinations in random samples of different sizes. For example, in a random population of 100,000 a combination of first, middle and last name is an almost perfect identifier (about 96% people bear a unique name), while name pairs (first, last) reliably distinguish less then 75% of people in the same sample.

3 Methods

Name Popularity Prediction. In this section, we informally describe name popularity prediction models evaluated within the study. In what follows, $C(x)$ is the number of people with a name x in a training set S_{train}, where x can be either a full name or its constituents; f stands for first name, m and ℓ – for middle and last names, respectively; N_r is the number of names that occur exactly r times in S_{train} and N is the total number of persons in S_{train}.

We start with a naïve estimate assuming all people have unique names (model I). So, the number of people with the name x is equal to 1 in the population of any size. Then, we proceed with straightforward maximum likelihood estimates (MLE) for full names (II):

$$P_{MLE}(fm\ell) = \frac{C(fm\ell)}{N} \tag{1}$$

Model II assigns zero probabilities to names unseen in S_{train}. To partially mitigate the problem we can assume independence of name constituents and approximate the probability of a full name as follows, which defines model III:

$$P_{ind}(fm\ell) = P_{MLE}(f)P_{MLE}(m)P_{MLE}(\ell) = \frac{C(f)}{N} \cdot \frac{C(m)}{N} \cdot \frac{C(\ell)}{N} \qquad (2)$$

This model assigns a zero probability to a name if one of its components is new in the test set.

Some combinations of first, middle, and last names occur together more frequently than others. To capture these dependencies we use conditional probabilities. In the case of names triples we apply Markov assumption, in other words – we account only for dependencies between pairs of constituents leading to model IV:

$$P(fm\ell) = P(f)P(m|f)P(\ell|f,m) \approx P(f)P(m|f)P(\ell|m) \qquad (3)$$

Further, to mitigate the problem of zero probabilities of unseen name components, we use several smoothing techniques [5,12].

Laplace smoothing (models V and VI) is a simple additive smoothing method: pretend that every name x occurs $\alpha > 0$ times more than it has been observed in the training set. Thus, the number of people with previously unseen name is estimated to be α. If V is the set of unique names in S_{train}, then

$$P_L(x) = \frac{C(x) + \alpha}{N + \alpha|V|} \qquad (4)$$

In the case of LNRE distributions it is highly beneficial to have an estimate of unseen events for smoothing. LNRE models implemented in *zipfR* [10] allow us, starting with name the frequency distributions of S_{train}, to estimate the number of different names in a set of larger size and consequently the number of names not appearing in S_{train}. As Table 1 shows, the *Generalized Inverse Gauss-Poisson (GIGP)* model implemented in *zipfR* performs very well.

Table 1: Prediction of the number of unique names (the third column contains country-wide estimates for reference).

Name	*GIGP* estimates	Actual counts in S	Country-wide *GIGP* estimates
f	111,538	111,287	405,154
m	155,635	155,726	462,738
ℓ	461,343	463,613	729,218
$f\,\ell$	4,383,342	4,391,157	20,330,441
$f\,m\,\ell$	9,088,527	9,087,716	65,867,708

Good-Turing smoothing [11] is a more gentle smoothing approach widely employed in language modeling (VII). The general idea behind the approach is to estimate the probability of all unseen names roughly equal to the total probability of names that appear only once in S_{train}, i.e. $\frac{N_1}{N}$. The counts of all

other names are discounted as $C^*(x) = (C(x)+1)N_{C(x)+1}/N_{C(x)}$. The Good-Turing probability estimates are given by:

$$P_{GT}(x) = \begin{cases} \frac{C^*(x)}{N}, & \text{if } C(x) > 0 \\ \frac{N_1}{N} \cdot \frac{1}{E}, & \text{if } C(x) = 0 \end{cases}, \tag{5}$$

where E is a $GIGP$ estimate of hapaxes in S based on S_{train}. Note that it implies we know the size of the test set S beforehand.

One of the drawbacks of the Good-Turing smoothing is that it discounts probabilities uniformly in different frequency ranges. It leads often to severely distorted probabilities for high-frequency items. *Katz smoothing* [13] uses MLE for high-frequency names ($C(x) > 3$ in our experiment) and Good-Turing smoothing for low-frequency ones (model VIII).

Aiming at combining the simplicity of Laplace smoothing and the selectivity of Katz smoothing, we introduce *pseudo-Laplace smoothing* with a small $\alpha > 0$ (model IX):

$$P^*_{PL}(x) = \begin{cases} \frac{C(x)}{N+\alpha}, & \text{if } C(x) > 0 \\ \frac{\alpha}{N+\alpha}, & \text{if } C(x) = 0 \end{cases} \tag{6}$$

The idea is quite simple: names present in the training set obtain probability close to the MLE, while unseen names get reasonable non-zero probabilities. In a strict mathematical sense, these are not probabilities, since they do not sum up to unity (and that is why we denote it P^*). Such probability-like scores are widely used in many practical applications, see for example "stupid back-off" introduced in [3].

Name popularity estimation. The first experiment is estimation of name popularity, i.e. estimation of the number of people bearing each name. Evaluation of models on samples with a large number of unique events is not an easy task. Evaluation results may diverge significantly on different test samples and depend on the size of test sample, particularly in low frequencies ranges. For example, LNRE models are traditionally evaluated by looking at how well expected values generated by them fit empirical counts extracted from the same dataset used for parameter estimation [8,1]. In this experiment we follow extrapolation setting for evaluation described in [9]: the parameters of the model are estimated on a subset of the data used subsequently for testing. The whole data set S is a list of 13.4 million real-world persons represented by TINs and corresponding names. We randomly sampled a training set S_{train} of 6.7 million persons, which is 50% of S. We employ *root-mean-square error* (RMSE) between the estimates and actual counts averaged over all names as evaluation measure. RMSE of a model $\mathcal{M}$ on the test set of people S over the set of unique full names V is defined as follows:[5]

$$\sigma = \sqrt[2]{\frac{\sum_{x \in V}(|S| \cdot P_{\mathcal{M}}(x) - C(x))^2}{|V|}} \tag{7}$$

[5] Note, that in this case $C(x)$ corresponds to the number of persons bearing name x in S (not in S_{train} as in equations above).

Table 2: Name models performance for full name triples

Model	Description	σ_1	σ_{2-5}	σ_{6-20}	σ_{20-100}	$\sigma_{>100}$
I	Always 1	0.000	1.833	9.163	38.279	163.327
II	$P_{MLE}(fml)$	1.000	1.611	**3.061**	**5.949**	**12.627**
III	$P_{MLE}(f)P_{MLE}(m)P_{MLE}(\ell)$	0.940	1.842	4.633	14.573	56.297
IV	$P_{MLE}(f\|m)P_{MLE}(m\|\ell)P_{MLE}(\ell)$	0.897	**1.608**	3.165	6.639	16.925
V	$P_L(f\|m)P_L(m\|\ell)P_L(\ell) \quad \alpha = 1$	0.999	2.720	9.779	36.277	137.747
VI	$P_L(f\|m)P_L(m\|\ell)P_L(\ell) \quad \alpha = \frac{1}{\|S_{train}\|}$	0.897	**1.608**	3.165	6.639	16.925
VII	$P_{GT}(f\|m)P_{GT}(m\|\ell)P_{GT}(\ell)$	0.900	1.622	3.171	6.644	16.931
VIII	$P_K(f\|m)P_K(m\|\ell)P_K(\ell)$	0.901	1.614	3.165	6.639	16.925
IX	$P^*_{PL}(f\|m)P^*_{PL}(m\|\ell)P^*_{PL}(\ell) \quad \alpha = 1$	**0.885**	**1.608**	3.165	6.639	16.925

In order to have a better understanding of models' behavior and their applicability to different tasks and data volumes, we calculate σ for the following name frequency buckets: 1 (hapaxes), $2-5$, $6-20$, $21-100$, and >100.

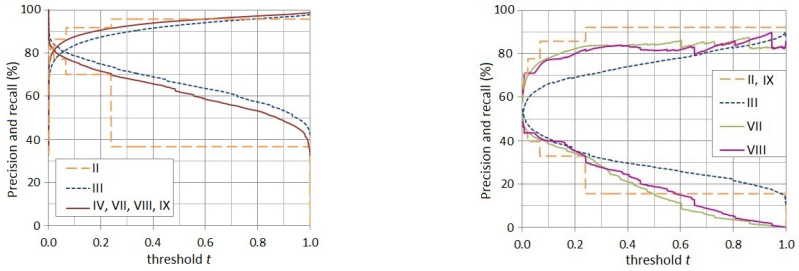
Record linkage. For the second task we calculate the probability that there is a single person with a given name x in the population of size $|S|$ using estimates by different models $\mathcal{M}$. If the probability surpasses the threshold t, we link records with identical names. Note that all identical names are linked at once, whereby q records with a given name trigger $\frac{q(q-1)}{2}$ linkages. The evaluation measure for the task are standard classification measures: *precision* – the fraction of linked records pairs that are correct, i.e. both refer to the same real-world person, and *recall* – the fraction of correct links identified. There are about 63.2 million pairs of identical names among 20.6 million occurrences, i.e. potential links between same-person records; 32% of them are correct according to TINs. Taking into account these figures, linking all possible pairs results in $precision = 32\%$ and $recall = 100\%$.

In contrast to the first experiment that presumably reflects a global distribution of names, the second experiment deals with a concrete database and its particular characteristics, e.g. the number of companies associated with a person.

4 Results

Table 2 summarizes evaluation results for nine name popularity prediction models.[6] The first model (I) is a naïve "always 1" baseline that assumes all names are unique. Obviously, the model performs ideally on hapaxes. MLE model for full name triples (II) demonstrates the best prediction results in higher frequency ranges. The product of individual probabilities for first, middle and last names (III) performs slightly better on hapaxes, but substantially underestimates the probability of more frequent names. We investigated different dependencies between full name constituents, and combination in the model IV performed best.

[6] We also performed an experiment with first-last name doubles that showed similar behavior of the models. We do not cite the results here due to limited space.



(a) name triples (first, middle, last)

(b) name doubles (first, last)

Fig. 2: Record linkage evaluation results: precision (upper curves) and recall (lower curves) of various name count prediction methods depending on the threshold value t

As one can see, conditional probabilities considerably improve over model III that assumes independence of name constituents. The next five models incorporate smoothing. Add-1 smoothing (V) is too aggressive in case of LNRE distributions and model with independent name components (III) has too many zeros probabilities in case of one of name component is unseen. All other models perform slightly worse then MLE model, but comparably to each other models with smoothing. Our method (IX) performs best in the low-frequency range and equally well as models IV and VI in higher-frequency areas.

Record linkage Results of the record linkage experiment are presented in Figures 2a (name triples) and 2b (name doubles). The threshold t governs the linkage process: the higher the threshold the less name mentions are linked. One can imagine the process of gradual data linkage going from right to left, from higher to lower t values. Stepped curves of the MLE models are due to the fact that at some t values a large number of links is established at a time. In the case of full name triples (Figure 2a) all 'advanced' methods deliver almost identical results. The simplest MLE method for full names works well when we favor precision over recall. Threshold $t = 0.2$ delivers precision of about 90% and recall above 70%. In the case of first and last name doubles, the task of record linkage in such a sizable dataset based solely on name popularity estimates is much less effective (see Figure 2b).

5 Conclusion

In our experiments we make use of a large name dataset with unique identifiers that contains names of approximately one tenth of the Russian population. We conducted a series of experiments with different name popularity prediction models built upon the name dataset. We thoroughly evaluated several models, including well-known smoothing approaches and proposed a new simple yet effective method for adjusting probability estimates accounting for unseen events. Results show that the considered methods behave differently depending on the

frequency range of names to be estimated, the name structure (full name triples vs. first and last name doubles), and the population size for which the prediction is made. These experimental results can serve as guidelines for choosing the most suitable method for a specific task and available data.

We conducted a record linkage experiment in a database based solely on name popularity estimates. The outcomes suggest that name popularity estimates are a valuable signal for personal name matching. Results show that all methods using smoothing perform almost identically and the simplest method based on maximum likelihood estimates can be a good choice, when precision is more important than recall. However, these results reflect the peculiarities of a specific database and serve merely as an illustration of feasibility of the approach.

References

1. Baayen, H.: Word frequency distributions. Text, speech and language technology, Kluwer Academic Publishers (2001)
2. Bergsma, S., et al.: Broadly improving user classification via communication-based name and location clustering on Twitter. In: NAACL-HLT. pp. 1010–1019 (2013)
3. Brants, T., Popat, A.C., Xu, P., Och, F.J., Dean, J.: Large language models in machine translation. In: EMNLP-CoNLL. pp. 858–867 (2007)
4. Chang, J., Rosenn, I., Backstrom, L., Marlow, C.: ePluribus: Ethnicity on social networks. In: ICWSM. pp. 18–25 (2010)
5. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. Computer Speech & Language **13**(4), 359–393 (1999)
6. Christen, P.: A comparison of personal name matching: Techniques and practical issues. Tech. Rep. TR-CS-06-02, Australian National University (September 2006)
7. Christen, P.: Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer (2012)
8. Evert, S.: A simple LNRE model for random character sequences. In: JADT. pp. 411–422 (2004)
9. Evert, S., Baroni, M.: Testing the extrapolation quality of word frequency models. In: Corpus Linguistics Conference Series. vol. 1 (2005)
10. Evert, S., Baroni, M.: *zipfR*: Word frequency distributions in R. In: Proceedings of ACL. pp. 29–32 (2007)
11. Good, I.J.: The population frequencies of species and the estimation of population parameters. Biometrika **40(3 & 4)**, 237–264 (1953)
12. Goodman, J.T.: A bit of progress in language modeling. Computer Speech & Language **15**(4), 403–434 (2001)
13. Katz, S.M.: Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech, and Signal Processing **35**(3), 400–401 (1987)
14. Khmaladze, E.V.: The statistical analysis of a large number of rare events. Tech. Rep. MS-R8804, CWI (1988)
15. Liu, J., et al.: What's in a name? An unsupervised approach to link users across communities. In: WSDM. pp. 495–504 (2013)
16. Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.: Understanding the demographics of Twitter users. In: ICWSM (2011)
17. Perito, D., Castelluccia, C., Kaafar, M., Manils, P.: How unique and traceable are usernames? In: PETS, pp. 1–17 (2011)