Computing the Odds of a Persona Record Match

P(M|Ei) = P(M)P(Ei|M)/P(Ei) where Ei is the ith Evidence (e.g. same birth date, same mother’s birth surname). We assume mutual independence of the evidence for record matching. (Birth and death event dates and places of randomly selected persons are independent as are marriage event dates and places of randomly selected couples. For randomly selected persons, names of related persons—father, mother, spouse, children—are independent.)

Considering all the evidence, we would like to compute:

P(M|E1, E2, …, En) = P(M)πin P(Ei|M)/P(Ei)

Determining the probabilities on the rhs for a book (or collection of records) is hard (requires human-created ground truth) and thus untenable for our application. So we instead compute the odds in favor of P(M|E1, E2, …, En), which is obtained as follows (see [White02]). (The method here is actually a variation of [White02] in which we ignore finding the odds of a mismatch because we reject a possible match based on temp-merge red flags. Thus, we only consider the evidence in favor and ask if it is sufficient to declare a match.)

log(P(M|E1, E2, …, En)) = log(P(M)πin P(Ei|M)/P(Ei)) = log(P(M)) + ∑in(log(P(Ei|M)/P(Ei))

In this formula log(P(Ei|M)/P(Ei)) is the odds in favor of a match given evidence Ei for each i. Intuitively this makes sense if we reason as follows:

* Since we are summing the odds of each piece of evidence to obtain the overall odds of a match, we can ignore the constant term for it only serves to offset the threshold for deciding whether there is sufficient evidence to declare a match.
* Given that two persona records are records for the same individual, P(Ei|M) = 1 for each i. (Although this must be true when we consider the actual facts—e.g. a person has only one birth date, has only one mother who has only one given birth name—the records themselves do not always reflect these truths. There may be recording errors, OCR errors, alternate declarations of these truths. We take this into account later by tempering each odds value by the probability of sameness, but for now we can greatly simplify our odds formula.)
* Given the above two simplifications, we are left with only having to estimate ∑in(log(P(1/P(Ei)). (Furthermore, as it turns out, we can obtain the estimates for each piece of evidence on the fly from the book we are processing.)
* Considering the odds estimates intuitively, note that when the denominator is small, the odds formula value will be large. (When the likelihood that the evidence could occur randomly is small, the confidence in the evidence becomes greater.)
* As mentioned, since we only have what’s written in a persona record and don’t know if what’s written is accurately comparable, we temper the strength of the evidence by the probability of a match. (Examples: “W.W. TEEGARDEN” ?= “WM. WALTER TEEGARDEN”, “Nov. 4, 1891” ?= “1891” as stated death dates, “Waddington” ?= “Clitheroe” as stated birth places. We thus temper the odds for a piece of evidence by an estimated probability that the attributes match. These match probabilities depend only on the type of evidence and thus can be set in advance.)

Since P(Ei|M) = 1 and log(P(M)) can be ignored, the only value we need to compute is P(Ei). We use GreenQQ as an entity recognizer and we use inference to augment the information, standardization to regularize the information for matching, and canonicalization to allow for fuzzy probabilistic matching. Based on extracted and curated information, we can quite accurately obtain the actual counts needed to estimate the probability values we need for a book.

When obtaining probability estimates, we use exact matches over names, dates, and places. We consider all available evidence extracted by GreenQQ and curated by inference, canonicalization, and standardization. We categorize the evidence as follows:

* Names: (1) given names, (2) birth surname, (3) married surnames of females.
* Dates: dates to whatever precision is given.
* Places: locations (standardized as city/town/…-county/shire/…-state/province/…-country) to whatever precision is given.

An estimate of P(Ei), the probability of a match for the ith piece of evidence, can be obtained from the information extracted and curated as follows.

1. For each piece of evidence Ei, obtain the occurrence count Nij in the book for each categorical value j extracted as part of a persona record by GreenQQ and curated by our enhancement code in 7.osmx-enhanced.
2. Let the sum over j for Ei be Ni.
3. Let ki be the number of distinct j-values for Ei.
4. The estimated probability P(Ei) for a match given two person records both with a value for Ei is ∑j=1kiNij(Nij - 1)/Ni(Ni - 1) (selection without replacement) or we could more simply just use ∑j=1ki(Nji/Ni)2. (Example: P(two randomly chosen persona records both have the same mother given name) ≈ the estimated probability that two selected records both have mother given names “Mary” + the estimated probability that the two selected records both have mother given names “Abigail Huntington” + … “Maria” + … “Anna Margaretta” + …)

If both persona records have values for the ith attribute (e.g. mother’s birth surname) and the values match, the contribution to the overall probability of this evidence is 1/P(Ei). Intuitively, this makes since because the smaller the probability P(Ei) the greater the weight of the evidence. The log of 1/P(Ei) is the weight of the ith piece of evidence, which can be summed across all the evidence to yield a quantitative odds value indicating the confidence of the match (the larger the value, the greater the confidence). Because attribute instance values may be the same even when they are not identical, we temper the weight by the probability of sameness when summing over the evidence.

Given the overall evidence (the summation of the tempered odds values for each piece of evidence), we still have the problem of determining how much evidence is sufficient (what the threshold value should be to say that we are confident in the match). This value will be different for every book. We compute the threshold for each book using a variation of the “five fingerprints” rule of thumb. We find the average odds weight, multiply it by 5, and use it as the threshold value. Most likely we will need to experiment some with the idea in mind of using a different number than 5 and perhaps a mode or median or weighted average instead of the average.