International Journal of Population Data Science

Journal Website: www.ijpds.org

A suggestion on how to include calendar dates in Bloom Filters

Stevens, Antony[1*]

[1]Ministry of Health, Brazil

Objective

Bloom Filters have been used in a number of studies conducted for the Ministry of Health. They are usually recommended because of the possibility that they may participate in secure protocols for the exchange of data. In our case the speed of the program, once the filters have been prepared, is so high that that itself is sufficient motive for their adoption.
Nevertheless if two calendar dates differ by one character this may merit more attention than a similar difference in personal names. This became evident in a large linkage between mortality records and hospital separations where the patient had died. Higher scores were obtained when the date fields differed by only one character, but when that character represented a year there would no reason to notice the pair. When the character difference was compatible with a difference of a few days this would be more interesting because in studies like the one just cited it would be reasonable to admit differences of a few days or even, perhaps, weeks between the events (recording of the death of the patient).

Approach

How then to represent the difference between dates in a Bloom Filter? A date can be represented as a Boolean vector where the day (or week) is set to '1'. It may be represented by several contiguous '1's to admit admissible uncertainty in comparisons. The similarity between two dates can then just be the Dice Coefficient of the corresponding vectors.

Results

But a vector representing a date may then be very large. It could be as much as 365 bits per year, far more than is usually used for the other fields. The number of logical word comparisons would go up and the program would become slower. Knowing that the admissible range is presented by contiguous '1's means that we can obtain the effect of constructing the Bloom Filter and calculating the Dice Coefficient more directly. Starting with the two dates we can obtain the number of bits that are shared, which will depend on the admissible range. The Dice Coefficient can then be calculated directly without the need to construct the Filter.

Conclusion

We are then left with the decision on how to add the result to the value obtained from the other variables, and this will depend on what importance it is felt the date should have.

*Corresponding Author:
Email Address: antony.stevens@gmail.com (A. Stevens)

http://dx.doi.org/10.23889/ijpds.v1i1.33