Dynamic networks for currency estimation

Jono Sonck [1], Antoon Bronselaer [1], Guy De Tré [1]

[1] Dept. of Telecommunications and Information Processing, Ghent University, Belgium
{Jono.Sonck,Antoon.Bronselaer,Guy.DeTre}@UGent.be

Currency refers to the probability that data reflects the real-world values at a certain point in time. It is an important aspect of data veracity, as out-of-date values can lead to incorrect conclusions. Estimating the currency of a data value in an accurate manner can thus allow us to change the effect a value has in a decision-making process [4, 5].

However, research on currency estimation is divided. Some currency estimators use so-called *shelf-life models* that derive a probability from the tail of for example an exponential distribution [3, 4, 5]; others take a *logic constraints-based approach* where the changes in one variable should imply changes in other variables [1, 2, 6]. These approaches oppose different views on currency estimation, but could benefit from some form of unification.

We thus propose a modular framework that includes all aspects of currency estimation and offers the flexibility of combining components depending on the case at hand. To do this, we implement a dynamic Bayesian belief network, where the nodes represent probabilities of attribute age that can be estimated using various techniques, and the edges impose conditions and constraints on the (transfer of) values in the network.

At its core, each node in the network uses distributions to reflect the probable age and thus the currency of an attribute, partially dependent on the previous time period. When new data becomes available, the state variables of the node are updated (so-called *update* step). These new state variables allow us to recalculate the underlying probability distribution with the addition of a new time period (so-called *inference* step). The modularity of our network allows us to adapt the inner workings of our nodes to accommodate the characteristics of the attribute under consideration.

The relations between nodes imply logical dependencies between attributes through conditions or constraints on the update step of child nodes, from values and distributions imposed by parent nodes. The network thus forms a hierarchy in the feature set, which is curated by someone with expert knowledge on the given subject/dataset.

Further research will explore all possibilities in our framework for probability distribution calculation methods, constraints on relations, and automated network construction through machine learning techniques.

References

[1] Malte Barth et al. *Detecting Stale Data in Wikipedia Infoboxes*. Version 1. 2023. DOI: 10.48786/EDBT.2023.36. URL: https://openproceedings.org/2023/conf/edbt/3-paper-33.pdf.

[2] Tobias Bleifuß et al. "Exploring change: a new dimension of data analytics". In: *Proc. VLDB Endow.* 12.2 (Oct. 1, 2018), pp. 85–98. ISSN: 2150-8097. DOI: 10.14778/3282495.3282496. URL: https://dl.acm.org/doi/10.14778/3282495.3282496.

[3] Junghoo Cho and Hector Garcia-Molina. "Estimating frequency of change". In: *ACM Trans. Internet Technol.* 3.3 (Aug. 1, 2003), pp. 256–290. ISSN: 1533-5399. DOI: 10.1145/857166.857170. URL: https://dl.acm.org/doi/10.1145/857166.857170.

[4] Bernd Heinrich and Mathias Klier. "Assessing data currency — a probabilistic approach". In: *Journal of Information Science* 37.1 (Feb. 1, 2011). Publisher: SAGE Publications Ltd, pp. 86–100. ISSN: 0165-5515. DOI: 10.1177/0165551510392653. URL: https://doi.org/10.1177/0165551510392653.

[5] Bernd Heinrich and Mathias Klier. "Metric-based data quality assessment — Developing and evaluating a probability-based currency metric". In: *Decision Support Systems* 72 (Apr. 1, 2015), pp. 82–96. ISSN: 0167-9236. DOI: 10.1016/j.dss.2015.02.009. URL: https://www.sciencedirect.com/science/article/pii/S0167923615000299.

[6] Mostafa Milani, Zheng Zheng, and Fei Chiang. "CurrentClean: Spatio-Temporal Cleaning of Stale Data". In: *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. 2019 IEEE 35th International Conference on Data Engineering (ICDE). ISSN: 2375-026X. Apr. 2019, pp. 172–183. DOI: 10.1109/ICDE.2019.00024. URL: https://ieeexplore.ieee.org/document/8731522.