Harmony Assumptions: Extending Probability Theory for Information Retrieval

Thomas Roelleke

Queen Mary University of London

In many applications, independence of event occurrences is assumed, even if there is evidence for dependence. Capturing dependence leads to complex models, and even if the complex models were superior, they fail to beat the simplicity and scalability of the independence assumption. Therefore, many models assume independence and apply heuristics to improve results. Theoretical explanations of the heuristics are seldom given or generalisable.

[1] reports that some of these heuristics can be explained as encoding dependence in an exponent based on the generalised harmonic sum. Unlike independence, where the probability of subsequent occurrences of an event is the product of the single event probability, harmony is based on a product with decaying exponent.

For independence, the sequence probability is $p^{1+1+\dots+1} = p^n$. For harmony, the probability is $p^{1+1/2+\dots+1/n} \approx p^{1+\log(n)}$. The generalised harmonic sum is the exponent of p, and this leads to a spectrum of *harmony assumptions*. We will discuss that settings of the term frequency (TF) in IR correspond to harmony assumptions. We will focus on four settings of the TF:

$$\mathrm{TF}(t,d) := \begin{cases} \mathrm{tf}_d & \text{total TF: corresponds to assuming independence} \\ \sqrt{\mathrm{tf}_d + 1} - 1 & \text{sqrt TF: middle between total TF and log-TF} \\ \log(\mathrm{tf}_d + 1) & \text{log-TF: assumes a form for harmony} \\ \mathrm{tf}_d/(\mathrm{tf}_d + K_d) & \text{BM25 TF: assumes a strong form of harmony} \end{cases}$$

[1] shows series-based explanations of the TF settings, and these lead to new insights regarding the relationships between IR and probability theory. From an IR point of view exciting is the finding that the BM25-TF is the harmonic sum of Gaussian sums.

$$\frac{\mathrm{tf}_d}{\mathrm{tf}_d + 1} = \frac{1}{2} \cdot \left[1 + \frac{1}{1+2} + \dots + \frac{1}{1+2+\dots+\mathrm{tf}_d}\right]$$

This finding provides a probabilistic interpretation of the BM25-TF quantification.

An experimental study for IR and social media investigates assumptions that explain the dependence between term occurrences. Interestingly, the assumption sqrt-harmony, i.e. the middle between the total-TF and log-TF, is on average a better assumption than independence or the strong harmony assumptions corresponding to log-TF and BM25-TF. The potential impact of harmony assumptions lies beyond IR, since many scientific disciplines and applications rely on probability theory and apply heuristics to compensate the independence assumption. Given the concept of harmony assumptions, the dependence between multiple occurrences of an event can be reflected in an intuitive and effective way.

References

1. Thomas Roelleke, Andreas Kaltenbrunner, and Ricardo A. Baeza-Yates. Harmony assumptions in information retrieval and social networks. *Comput. J.*, 58(11):2982–2999, 2015.