The Value of Connecting Things - Part III: Ten Benefits from Big Structure

by Mike Bergman - Thursday, September 04, 2014

http://www.mkbergman.com/1791/the-value-of-connecting-things-part-iii-ten-benefits-from-big-structure/

Benefits Can be Gained Incrementally, and Are Cumulative

In the earlier installments of this article series we first described how to estimate the value of connections amongst Big Data datasets, premised on the network effect. We then went into detail in the second part about the Viking algorithm (VKG) derived to capture the Value of Knowledge Graphs.

In this concluding part of the series we summarize the use and implications of the Viking algorithm on your Big Data planning. We do so by offering ten guidelines for how including Big Structure may be leveraged in the context of a knowledge graph. We then conclude the series with some caveats on the interpretation of these results and a discussion of possible future directions.

Ten Guidelines for Big Structure

As you think through your Big Data initiatives, we recommend you keep these ten guidelines relating to data structure in mind:

1. **More structure always provides benefits** -- adding structure always provides a multiplier effect in value
2. **Making connections is more valuable than adding more data** -- Big Data alone is practically worthless if not connected. Adding more records has an additive effect on the value of the datasets in comparison to the ***multiplier*** effects of structure
3. **Benefits of structure increases with increasing dataset sizes (scale)** -- the multiplier effect of more structure (connections) increases with scale. Big Data projects are thus perfect candidates for consciously "connecting the dots"
4. **Particular kinds of structure -- such as types or categorization -- have higher benefit than annotations** -- structural characteristics at the record level that enable cross-dataset selections and comparisons are inherently more valuable than record-specific annotations. Typing of records into entity types is a very powerful lever

5. **The potential value of a knowledge graph depends on the nature of the domain** -- knowing what kind of knowledge graph is in play is an important metric for being able to estimate potential value from connections. Further, by adding connections (correct and coherent) it may also be possible to move the entire structure to a lower average degree of separation (*D*), with further multiplier benefits
6. **Structure can be added incrementally, and is cumulative** -- because these additions to structure are based on the open world assumption (OWA), it is possible to add structure incrementally. OWA enables connection and structuring efforts to be accomplished as budgets allow. But, because these structural benefits are cumulative (and with multipliers), later contributions can have increasing benefits over earlier ones
7. **Data wrangling is justified as a means to increase the accuracy of fact assertions** -- data wrangling should not be viewed as an overall "cost" to the effort, but as a key means for achieving the multiplier benefits arising from structure and connections
8. **Adding structure at the time of data wrangling is a cost-effective approach** -- the corollary to standard data wrangling is the wisdom of explicitly including structuring and connection to the efforts. The multiplier benefits that accrue are a means to markedly lower the marginal costs of data wrangling in relation to realized benefits
9. **Ontologies provide inferred capabilities** -- though many kinds of structure can contribute to the Big Structure effort, ontologies, because of their logical structure, can be used to derive inferred "facts" in essence "for free." (And remember, more "facts" are the basis for the multiplier benefits.) Inference provides a powerful means to leverage existing connections without explicitly needing to assert new ones
10. **Ontologies are the most preferred means for Big Structure** -- besides inference, ontologies set a structural framework of relationships (schema) very useful to helping to guide the nature of connections made. Also, ontologies can provide the conceptual and descriptive richness useful for tagging and other structure-adding activities [1]. Because of these advantages and their testable nature based in logic, ontologies represent the pinnacle of structural forms to achieve these value benefits.

Of course, mere connections for structure's sake is silly. It is important that the structure added and connections made are correct, consistent and coherent. Even then, not all types of connections are created equal, with typing the most important, annotations the least.

The good thing is that Big Structure can be added as a slight increase over standard data wrangling efforts, and with much greater impact than standard wrangling. Further, the structures themselves, preferably guided by domain ontologies, are a means of testing these factors for subsequent structure additions. Not only does adding structure get easier with a foundation of existing structure, but it increases the value of the information by orders of magnitude.

Not the Last Word

Roughly twenty years ago Metcalfe's law triggered a gold rush in trying to achieve network effects at Internet scale. Though the algorithm proved too optimistic at larger scales, the idea of the benefit from connections was firmly established. Ten years ago it was clear that some form of diminishing returns needed to be applied to connections at scale. Zipf's law was not a bad guess, though we have subsequently learned that more graph-centric measures are more appropriate and accurate for estimating value. Now, the Viking algorithm has emerged as the best estimator of the value of connections within Big Data.

I suspect we will see further improvements to the Viking algorithm as: 1) we come to better understand graph structures (including the effects of clusters and cliques); and 2) we learn to distinguish the different value of different types of connections [2]. We can already see that typing and categorization have better structural effects above annotations. We further can see that the correctness of asserted "facts" is a key to realizing the multiplier benefits of connections and structures. Thus, we should see improved means for screening and testing assertions for their accuracy at scale.

At this stage, what the Viking algorithm gives us is a defensible means for assessing the value of adding structure (through connections) to our datasets. We see these multiplier effects to be huge, and to compound to even still further benefits with scale. We also see that the most developed forms of structure -- namely, ontologies -- bring still further benefits in inference and testable coherence. All Big Structure efforts should be aiming to express all of the structural insights for the organization and its datasets into these ontological forms [1].

While our current proxy for value -- namely, asserted "facts" -- is useful, it would also be helpful to be able to translate these "fact" assertions into a monetary value. As we move down this path we will discover, again, that not all "facts" are created equal, and some have more monetary value than others. Transitioning our estimates of value to a monetary basis will help set parameters for the cost-benefit analysis of data collection and structurizing that is the ultimate basis for planning Big Data and Big Structure initiatives.

In the end, many things need to be analyzed to understand the impacts of each connection and structure metric on the value of the resulting graph. But, what today's current understanding of the network effect and the Viking algorithm brings us is a better means to understand and quantify the benefits of connected information. By any measure, these value benefits are multiples of what we see for unconnected data, the multiples of which grow massively with the scale of the data and their connections.

Big Structure is fertile ground for bringing in the sheaves. Let the harvesting begin.

[1] Though not further discussed here, the ontologies also provide the means for tagging (providing structure) to unstructured documents, which also brings the multiplier benefits from structure. On the retrieval side, such structure also aids faceting and filtered "slicing and dicing" of underlying datasets, thereby improving retrieval efficacy.

[2] As one of the first approaches to capture these nuances, see Mischa Dohler, Thomas Watteyne, Fabrice Valois and Jia-Liang Lu, 2008. Kumar’s, Zipf’s and Other Laws: How to Structure a Large-Scale Wireless Network?, published in *Annales des Telecommunications - Annals of Telecommunications* 63, 5-6 pp. 239-251. See http://hal.archives-ouvertes.fr/docs/00/40/58/67/PDF/Large_Scale_Networks_journal_FINAL.pdf.