Value Space Flattening

by Daniel J. Sullivan

Pub Date: 08/24/11

Rev Date: 02/11/17

The value space is the set of all tokens which represent the basic components of natural language. A balance exists between token cardinality and semantics. This set is broken into cardinality sets:

Set 1: Finite Char or Symbol Set { the smallest set with the exception of ideogram based languages like CHINESE. }
Set 2: Functional Tokens { this set is slightly larger but still relatively finite, this would be the split tokens of all strings when you split on white space or null/empty space }
Set 3: Finite Token Sequences (length 1 is allowed) Which 'may' be interpreted as a Basic Type { this set represents one or more tokens from Set 1 which can be converted, via a rule, to some basic type (date/time, number, string, etc.) }
Set (N): All higher order sets or super sequences (even folding sequences) that can allow for expanded semantics.

All higher order sets (really semantically meaningful sequences of sequences as recursive elements of sets) obey the same basic rule - the growth of complex grammar increases at a decreasing rate, but the growth of combination of these grammar sequences is a "greater" set, that, in theory can be parsed out to a NEWER set of higher order sequences which can be combined...

Moving from Set 1 to Set 3, the extension of the language occurs based upon the rules of combinatorics, but these rules are applied in a meaningful universe which restricts higher order relations to narrow spaces of actualization. Token/Sequence Cardinality and Semantics exist in balance. The complete set of relations r(T1,T2) where T1 is equivalent to T2 and the product is Cartesian - but this complete Cartesian product is restricted by rules external to the system which constrain evolution and change. Randomness of data is limited and controlled by features of the information ecosystem into which they are born.

If this is true, then I believe we can state that the value space generally increases in accordance with f(x) = square_root(x) also known as the rational curve or the diminishing curve. This is the growth model for the principal tokens of natural language, and this rule is indifferent to whether you are talking about Set 1, 2 or 3. We will show that these observations have a direct impact on the development and design of a scalable data warehouse. We will also suggest that this view of data may offer possible optimizations for future IC design. Set (N) does represent the universal set of all higher order complex sequences (or folding sequences), as such it should not be seen as a single set, but a recursively expanding super set of ever more complex sequences (some of which may contain multiple folds).

Information is Modular and Recursive - the complexity manifests itself in the recursion.