Structural Entropic Difference

a tree distance metric

Richard Connor
Fabio Simeoni
Michael Iakovos
(Robert Moss)

Main Contributions

- new metric
  - over tree-structured data
  - strongly grounded in information theory
  - adaptable to any structured data
- bounded
  - gives results in [0,1]
- distance metric
  - symmetry, identity, triangle inequality
- directly and “efficiently” computable
  - not an approximation

talk outline

- description of the metric
- qualitative properties
- quantitative properties
- proofs

complexity and similarity

- intuition
  - complexity: think of *information content*
  - compare, for two objects:
    - the sum of their complexities
      - C(A) + C(B)
    - the complexity of their union
      - C(AB)

Consider a perfect *C*

- if A and B are *the same*:
  - $C(A) = C(B) = C(AB)$
- if A and B have *nothing in common*:
  - $C(AB) = C(A) + C(B)$
- if A and B have some commonality:
  - $C(A), C(B) < C(AB) < C(A) + C(B)$
    - varying *continuously* as A and B have more or less in common

given these properties:

$$\frac{C(AB)}{mean(C(A)+C(B))}$$

- ranges between
  - 1, if A is identical to B, and
  - 2, if A and B have nothing in common

our "perfect" C (!)

- information content of a data structure:
  - based on navigable paths within the data
  - enter structure at a random point
  - navigate, emitting navigation token
  - leave randomly and reenter
- intuition:

  information content of *data structure*

  =

  information content of *emitted event stream*

event emission….

<**family**>
<**surname**>Smith</surname>
<**person**>
<**name**>Tom</name>
<**age**>46</age>
</person>
<**person**>
<**name**>Dick</name>
<**age**>10</age>
<**shoeSize**>37</shoeSize>
</person>
</family>

Example event stream:

<**person**> <**name**> *break* <**shoeSize**> *break* <**family**> <**person**> <**shoeSize**> *break* <**name**> *break* <**person**> <**age**> *break etc…*

calculating information content

- information content of event stream
  - can be calculated using Shannon's entropy equation
  - can be calculated from structure of tree
- information content of object union
  - equated to the information content of the merged information streams
  - also calculated "statically" from tree structures

the real metric…

$$D(s,t) = \left(\frac{b^{H_b(s \cup t)}}{b^{mean(H_b(s)\ ,\ H_b(t))}} \right) - 1$$

Qualitative Properties

- it is bounded
  - range in [0,1]
- it is a distance metric
  - symmetry
  - (pseudo-) identity
  - triangle inequality
    - (see paper in proceedings…)
- continuity
  - small changes in input give small changes in result

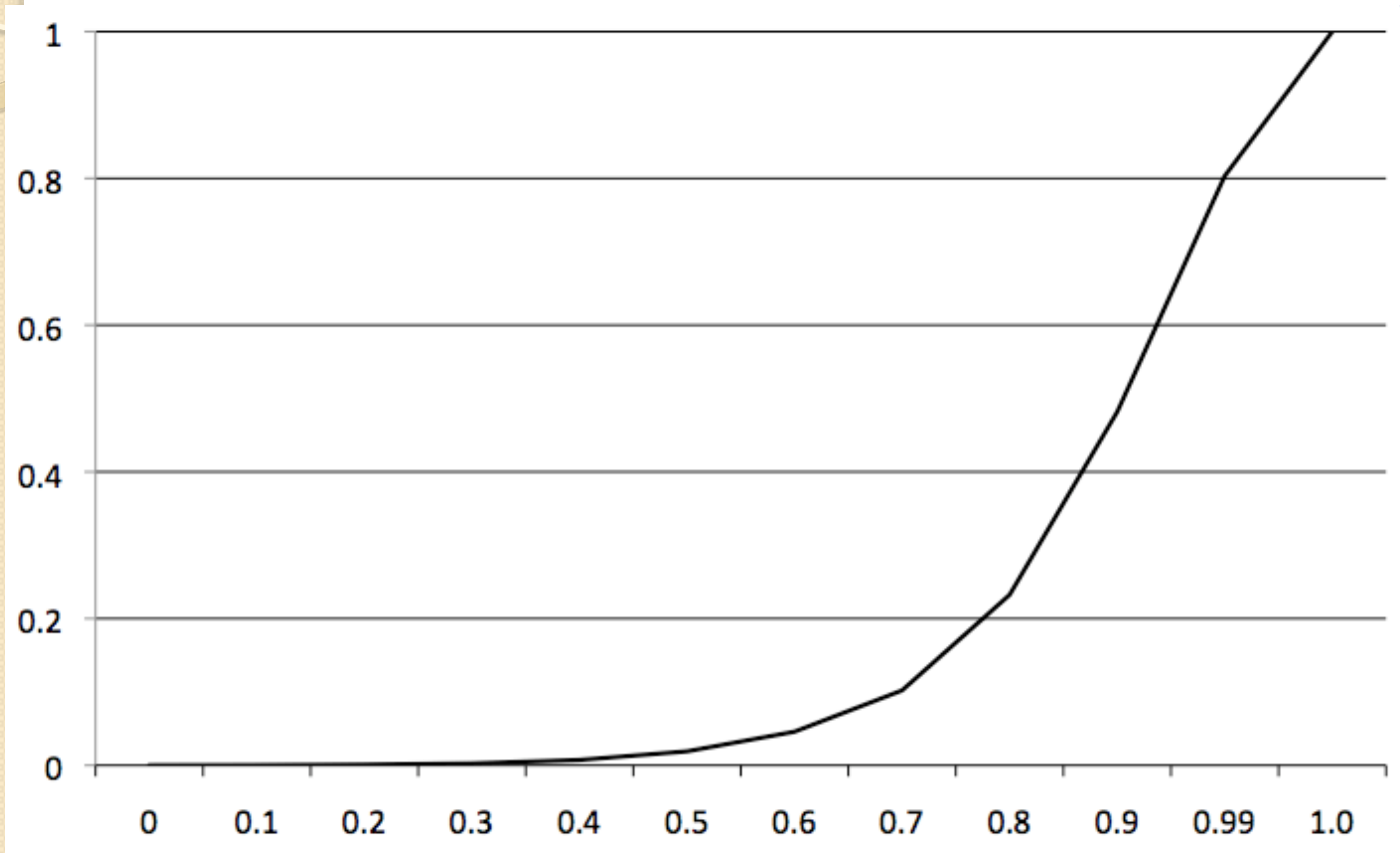
SISAP code plagiarism data set

```
S(T(),Z(Y(), d(Y(), d(Y(),W(), c(),W()), c())))
P(0.4000000059604645)
d(Y(), z(T(), b(2)), H())
p(b(-5), Q(a(),b(1)))
p(d(Y(), c(), d(Y())), K(Y(), c()))
%(U(U(U(M(),X()),X()),X()),^(b(1),b(-1)))
d(Y(),V(d(Y()),p(b(0), Q(a(),b(1)))),W())
K(Y(), Q(d(Y()),b(284)))
p(R(K(Y(),W()),P(16.0)), P(1.0))
p(d(Y(), c()), Q(a(),b(-1)))
S(T(),Z(Y(), K(Y(), d(Y(), c(), c())), b(0)))
J(I(8363.0),S(T(),V(H(),S(T(),A(W(),b(12))))))
J(K(Y(), R(K(Y(), d(Y())),W())),P(3.141592653589793))
A(V(U(M(),X()),W()),b(2))
d(Y(), S(T(),d(Y(), Z(Y(), d(Y())))))
```

Results on SISAP code examples:

Rank	Distance	Line no	Source
1	0.00	735	d(Y(), Z(Y(), W()), W(), d(Y(), W(), W(), Z(Y(), W()), W(), d(Y(), U(M(),X()), W())))
2	0.06	172608	d(Y(), W(), W(), Z(Y(), W()), W(), d(Y(), U(M(),X()), W()))
3	0.07	94351	d(Y(), d(Y(), U(M(),X()), Z(Y(), W())), d(Y(), Z(Y(), W())))
4	0.08	128977	d(Y(), d(Y(), d(Y(), U(M(),X()), W()), W()))
5	0.11	94338	d(Y(), Z(Y(), W()), W(), d(Y(), U(W(),X()), d(Y(), Z(Y(), W()), W())))
6	0.11	99666	d(Y(), W(), d(Y(), d(Y(), U(M(),X()), d(Y(), W()))))
7	0.11	21014	d(Y(), W(), W(), W(), d(Y(), U(W(),X()), d(Y(), Z(Y(), W()), W())))
8	0.11	176190	d(Y(), d(Y(), U(M(),X()), d(Y(), W())))
9	0.11	129005	d(Y(), W(), W(), d(Y(), U(W(),X()), d(Y(), Z(Y(), W()), W())))
10	0.11	106485	d(Y(), d(Y(), d(Y(), U(M(),X())), W()))

Sensitivity



Efficiency

- naïve implementation calculates 225,000 tree comparisons in around a minute
  - 0.25 msec per comparison (from "cold")
- major cost is constructing fingerprint
  - fast indexing (eg AESA) requires
    - *n* fingerprint constructions
    - ½(n x n) comparisons
  - fingerprint comparison is 1-2 orders of magnitude faster than fingerprint construction

proofs

- Paper gives proofs of distance metric properties
  - symmetry
  - (pseudo)-identity
    - equivalence function = bisimilarity
  - triangle inequality
- inherent tension between boundedness and (non-trivial) triangle inequality
- we are not aware of any other tree functions which are bounded distance metrics

Conclusions

- new distance metric
- bounded, metric properties proved
- grounded in information theory
- efficient to calculate