SourceRank: Trust and Relevance based Ranking of Web Databases for the Deep Web

Raju Balakrishnan, Subbarao Kambhampati
rajub@asu.edu  rao@asu.edu

Agreement Implies Trust & Relevance

Universal Set of Tuples (U)

Let \( t_1, t_2 \in R_T \), \( f_1, f_2 \in U - R_T \) and \( P_a \) be the agreement probability. Assuming independent selection,
\[
U \gg R_T \Rightarrow P_a(t_1, t_2) \gg P_a(f_1, f_2)
\]
Agreement between the sources is modeled as an Agreement Graph.

Source Selection in Deep Web

Computing agreement requires record linkage
- Value similarity: SoftTFIDF with Jaro-Winkler.
- Attribute importance is calculated based on mean inverse document frequency of the token values.
- Domain Independent.
- Predefined schema mapping is not assumed.
- Quadratic Time complexity on Sources.

SourceRank is calculated as the stationary visit probability of a weighted random walk on the database vertex in the agreement graph.

\\[ W(S_1 \rightarrow S_2) = \beta(1-\beta) \times \frac{A(R_1, R_2)}{|R_2|} \]

Sampling Databases
- Graph is computed based on samples from the databases.
- Non-Cooperative Query Based Sampling.

URL: http://factal.eas.asu.edu

Evaluated on Google base and online Sources. Comparison with Google Product Search

Trustworthiness is evaluated as the decrease in ranks of corrupted sources.