Optimizing Sample-based Entity Resolution over Streaming Documents

Christan Grant and *Daisy Zhe Wang*
{cgrant, daisyw}@cise.ufl.edu

Abstract

Increasingly, organizations have employed methods to understand unstructured text across the web. Entity resolution is used to identify mentions in large, streaming text corpora. Sampling-based entity resolution using Markov Chain Monte Carlo (MCMC) techniques guarantees convergence to a stationary distribution and can jump out of a local optimum.

When performing entity resolution over streams of incoming data, the growing quantity of data amplifies two central issues. First, because the sampling process is random, many iterations are wasted attempting to resolve unambiguous entities. Second, the quadratic runtime for scoring entities becomes prohibitive for largest entities. Frequent streaming updates from the web exacerbate these difficulties. In this paper, we discuss the creation of a proposal optimizer, in the spirit of database optimizers.

This optimizer observes the proposal updates to the entity resolution model then makes recommendations to improve the processing and storage of the model. We motivate the use of compression techniques to reduce the amount of processing when scoring MCMC updates proposal. We also discuss statistical early-stopping techniques for scoring entities. We describe our initial progress over a large entity resolution data set and how an optimizer can improve performance when processing entity resolution streams.

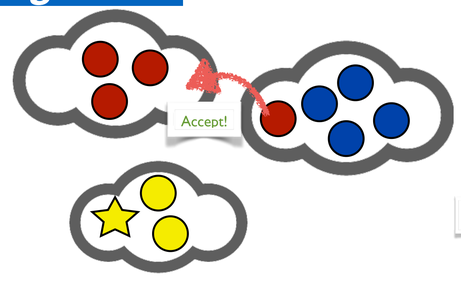
Entity Resolution

Entity resolution is the process of identifying and clustering different manifestations (e.g., mentions, noun phrases, named entities) of the same real world object.

- Difficult because of ambiguity
  Same Name, Different Person
  Different Name, Same Person

Entity Resolution Algorithm

1.Select a source mention at *random*.
2.Select a destination mention at *random*.
3.Propose a merge.
4.Accept when it improves the state.

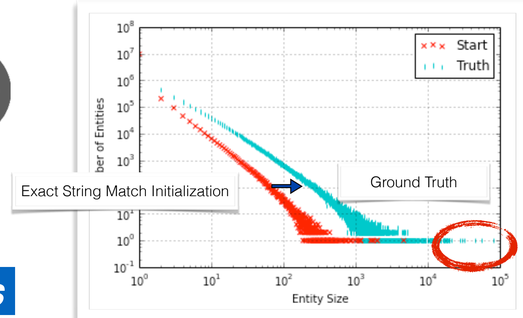


Test Set

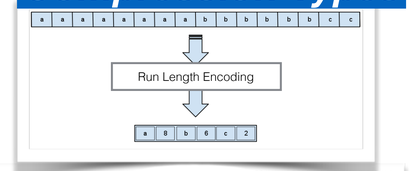
- **Wikilink Data Set** *(Singh, Subramaniya, Pereira, McCallum, 2011)*
  - Largest fully-labeled data set
  - 40 Million Mentions
  - 180 GBs of data

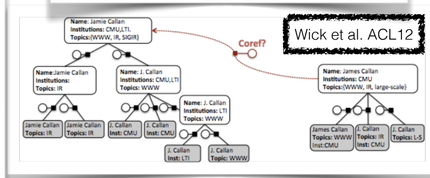
Figure 1: Links to Wikipedia as Entity Labels

When to compress?

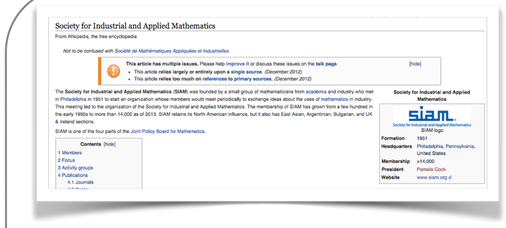


Compression Types

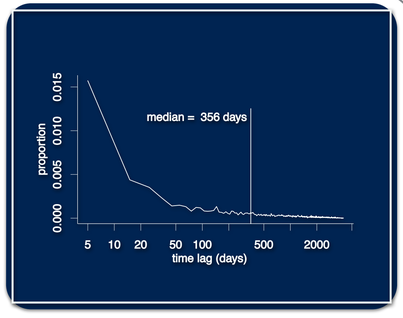




Knowledge Base Acceleration



14. ^ "Fellows Program". SIAM. Retrieved 2012-12-04.



The average time between an event and its appearance on Wikipedia is **356** days.

Entity Resolution Inefficiencies

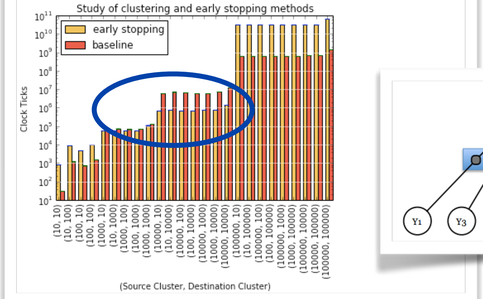
1. Large clusters are the slowest.

   Pairwise comparisons are expensive. $\Theta(n^2)$

2. Excessive computation on unambiguous entities

   Entities such as *Carnegie Mellon* are relatively unambiguous.

When to Approximate?



Early Stopping

Singh et al. EMNLP'12