Optimizing Sample-based Entity Resolution over Streaming Documents

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.



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

Knowledge Base Acceleration

| Society for Industrial and Applied Mathematics From Wikipedia, the free encyclopedia Mot to be confused with Société de Mathématiques Appliquées et Industrielles. In article has multiple issues. Please help improve it or discuss these issues on the talk page. In its article relies largely or entirely upon a single source. (December 2012) In sarticle relies too much on references to primary sources. (December 2012) This article relies too much on references to primary sources. (December 2012) This article relies too much on references to primary sources. (December 2012) This article relies too much on references to primary sources. (December 2012) This meeting led to the organization whose members would meet periodically to exchange ideas about the uses of mathematics in industry. This meeting led to the organization of the Society for Industrial and Applied Mathematics. The membership of SIAM has grown from a few hundred in the early 1950s to more than 14,000 as of 2013. SIAM retains its North American influence, but it also has East Asian, Argentinian, Bulgarian, and UK is Iraliand sections. SIAM is one of the four parts of the Joint Policy Board for Mathematics. I Members 2 Focus 3 Activity groups 4.1 Journals 4 Pleade | [hide] Society for Industrial and Applied Mathematics Society for Industrial and Applied Mathematics Society for Industrial and Applied Mathematics SIAM logo Formation 1951 Headquarters Philadelphia, Pennsylvania, United States Membership >14,000 President Pamela Cook Website www.siam.org @ | proportion .000 0.005 0.010 0.015 | median = 356 days |
|--|--|--------------------------------------|---|
| SIAM Fellows [edit] In 2009 SIAM instituted a Fellows program to recognize certain members who have made outstanding | contributions to the fields SIAM serves ^[14] | | 5 10 20 50 100 500 2000 time lag (days) |
| 14. ^ <u>"Fellows Program"</u> & SIAM. Retrie | eved 2012-12-04. | The ave and its | erage time between an event appearance on Wikipedia is |

356 days.



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





Entities such as *Carnegie Mellon* are relatively unambiguous.

Test Set





