SDM-BSA Workshop

2015 SIAM International Conference on Data Mining Workshop: Big Data and Stream Analytic

Reviews For Paper

Paper ID 3

Title Optimizing Sampling-based Entity Resolution over Streaming Documents

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Novelty (5 is the highest)	3
Correctness (5 is the best)	3
Readability (5 is the best)	1
	There is a very serious issue with the presentation of this paper. I copy-paste several sentences which have language issues during the course I read the paper. There is even a whole paragraph in which none of its sentences is grammatically correct. There are many concepts that are not well defined and aticulated. The logical flow is broken all over the place. I don't think this is something that is expected from a publishable manuscript, and I cannot expect technical rigor from such a manuscript. Sentences with language issues: 1. To this end, the National Institute of Standards hosted a three-year introduced a track to accelerate the extraction of information and construction of knowledge bases from streaming web resources 2. In sampling-based ER, using Markov Chain Monte Carlo (MCMC) techniques, trades some raw performance for a flexible representation and guaranteed convergence 3. In this paper, we argue that compression and approximation techniques can effeciently decrease the runtime of traditional ER systems thus making them usable for streaming environment. 4. There is no one size fits all for these samplingalgorithms 5. Early stopping is not always precise and adding an extra conditions in the sampling metropolis hastings looping structure may slow down computation. 6. Comparisons of the mentions in clusters is a factor of the number of pairwise links between mentions.

Detailed Comments	7. Additionally, cluster wide features are also computed and weighted. 8. First, is a method for approximating the score of factor counts and second is a compression method. 9. To this end, we look at three works that state-of-art techniques, respectively, for (1) early stopping methods, (2) entity resolution compression, (3) streaming entity resolution. 10. In this work, they note that sampling is usually inexpensive except in the case of many variable, or if scoring a sample is low. To this end they introduce Monte Carlo MCMC, this is a method of performing uniform and condent-based sampling over the computation of factors to avoid the full computation of features. This work shows signicant speed up in a large-scale experiments. 11. The hierarchical structure adds some book keeping to the sampling process but because large entity clusters are compress there are less comparison for each sampling iteration. This works displays orders of magnitude speedup when compared to a pairwise method. 12. The use the doubling algorithm [3] to grow sets of clusters and a clever disk arrangement strategy [8] to retrieve growing entities on demand. 13. We use a large real-world corpus to a motivating example. 14. First, given a source and destination entity (es; ed) and the mention, how can we score the proposal in the least amount of time. 15. There are many compression techniques, one being we only keep the mentions an entity that have a unique representation. 16. These number suggest that at times we we can take advantage of the redundancy within large entities and compress the entities. 17. This suggest that there are several unique representations of entities that during entity resolution entities sizes can expect to grow by an order of magnitude in size and the number of smaller entities will decrease. 18. The condence-base scoring method performs uniform samples of the mentions from the source and destination entities clusters.

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
Novelty (5 is the highest)	3

Correctness (5 is the best)	4
Readability (5 is the best)	3
Detailed Comments	Strong aspects of the work and summary of contributions: This paper addresses the run time challenges of entity resolution systems by considering the practical limitations of such systems when they are exposed to a streaming environment. The authors pinpoint the burning time of the MCMC and the quadratic run time for scoring entities as the main reasons behind the run time challenges. They promote the use of compression techniques. Also in order to understand how the computational efforts for an early stopping and compressing method scales with varying size of entities and mentions, some experiments have been conducted. The addressed problem is very relevant to the industrial streaming challenges. Technical Week aspects and recommendations: The idea of pre-filtering on the data in the form of a compression technique, in order to help MCMC samplers avoid resolving clear entities is a great one. However I am not sure if this idea has been implemented in this paper. Instead some experiments have been conducted to understand the running time behaviours for early stopping versus a baseline method as well as the compression time for different entity sizes. Please make sure to explain the main contribution of the paper more clear. In the factor graph representation, it would be better to be consistent in notations for example stick to one of the notations: xi and vi. Also as you never use the formulation it may be better to skip the mathematical representation and refer the reviewers to the suitable references. Readability and the grammatical issues: This paper requires substantial proofreading. As few example: Abstract: “waisted” : wasted “..creation a proposal..”: creation of a proposal... In the introduction section: there are two verbs in “hosted a three-year introduced a track to”... ” In sampling based ER, using Markov Chain Monte Carlo (MCMC)techniques, trades some raw”: there is no subject in the sentence. “ First, the computation of large entities and second, excessive computation spent resolving unambiguous entities”: This is not a sentence (there is no verb), maybe you need a conjunction with the previous sentence.

	I have found numerous mistakes of these sorts. I strongly invite the authors to proofread their paper before any possible publications please.

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Novelty (5 is the highest)	4
Correctness (5 is the best)	4
Readability (5 is the best)	3
Detailed Comments	The paper presents an algorithm for optimizing sampled based entity resolution over document streams. The optimization is based on two ideas: (1) a simpler sample approach based on confidence-based scoring; (2) run-length encoding to compress entity representation. Initial implementation and evaluation on wikilink corpus shows that the optimization is promising. The techniques proposed by the paper are valid and sound. My major concern is over the presentation of the paper: there are lots of misspellings and grammatical errors, making the paper very hard to follow and understand; also, the overall structure of the paper and writing could be further improved.

Masked Reviewer ID: Assigned_Reviewer_4

Review:

Question	
Novelty (5 is the highest)	3
Correctness (5 is the best)	4
Readability (5 is the best)	4
Detailed Comments	Optimizing Sampling-based Entity Resolution over Streaming Documents This paper describes an initial approach for optimizing sampling for the entity resolution process, with experimentations in large wiki document dataset. The authors propose an optimizer that attacks two major limitations, the size of the entities and the redundant computation. This paper motivated the need for the optimizer and examined the feasibility of its treation. Hence it's a relevant paper with good preliminary results and suitable for our Workshop in Stream Analytics.