Entity Compression for Incremental Entity Resolution

Christan Earl Grant
University of Florida
Dept. of Computer Science
Gainesville, Florida, USA
cgrant@cise.ufl.edu

Daisy Zhe Wang
University of Florida
Dept. of Computer Science
Gainesville, Florida, USA
daisyw@cise.ufl.edu

ABSTRACT

1. INTRODUCTION

The storage of user generated content within systems has introduced vast amounts of data. To clean the dump an important task is entity resolution. Entity Resolution (ER) is the problem of resolving the records in a data set that correspond the same real world entity.

Entity resolution is a notoriously computationally difficult problem. Several efforts in different domains have made outstanding progress []. The main issues still affecting runtimes of ER systems are twofold, first, the computation of large entities and second, excessive computation spent resolving unambiguous entities. Optimization that touches these difficult points is wholly understudied. Amdahl's argument suggests that compression and approximation techniques can efficiently decrease the runtimes of traditional ER systems.

Some recently, researchers have suggest methods of compressing entities. Wick et al Heirchical ... Singh et all efficient facttoring http://people.cs.umass.edu/ sameer/files/mcmcmc-emnlp12-ppt.pdf In this paper we propose aggressive compression methods to maximize resolution time. ...

We make the following contributions

In the paper we ...

2. BACKGROUND

Entity Resolution

Blocking

The distribution of entitiy sizes

3. RELATED WORK

Wick et al

Sameer Singh

Pay as you go ER

4. PROBLEM STATEMENT

We have two problems. First, Given an Entity $\mathcal{E}$ how can we create a a structure with minimal total size. Second, How can we incrementally add and remove items from the compressed structure. Third, how can we iterate over or sample from the compressed data structure to perform uncompression.

Distribution of Entity sizes in the data set

We can take advantage of the redundancy within large entities and compress the entities.

We look to perform add, remove and iterate operations over the compressed entity sets.

5. ALGORITHMS

6. IMPLEMENTATION

7. EVALUATION

8. CONCLUSION