FAMER for Error Detection and Correction

Abstract

Data sources may not only be “unclean” in the sense that they contain duplicates they may also be unclean in the sense that they contain incorrect information. This research proposal aims to leverage the properties of FAMER and MSCD to identify and potentially correct errors in the source entities.

Problem

Many real world data sources contain user generated input, that has never been vetted for correctness. An online store such as ebay may contain listings for cameras from the manufacturer “Camera Co.” vs “Kamera Co.”. Mistakes such as these can be corrected if a complete “Golden Truth” exists but in the absence of such a golden truth source, finding and correcting outliers is a greater challenge.

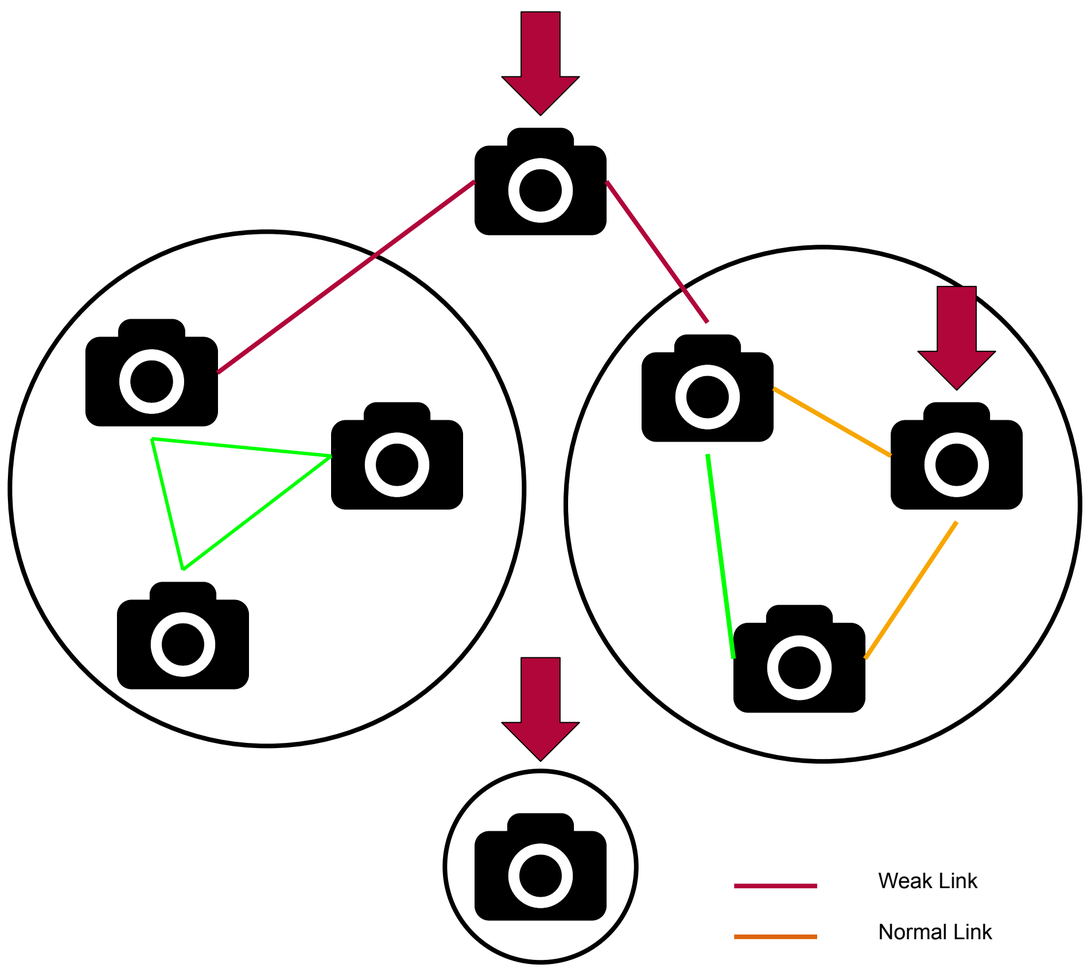
Solution

Detect “aberrant entities” in the clusters
Reintroduce aberrant entities with modified attributes to correct them.

Detecting Aberrant Entities

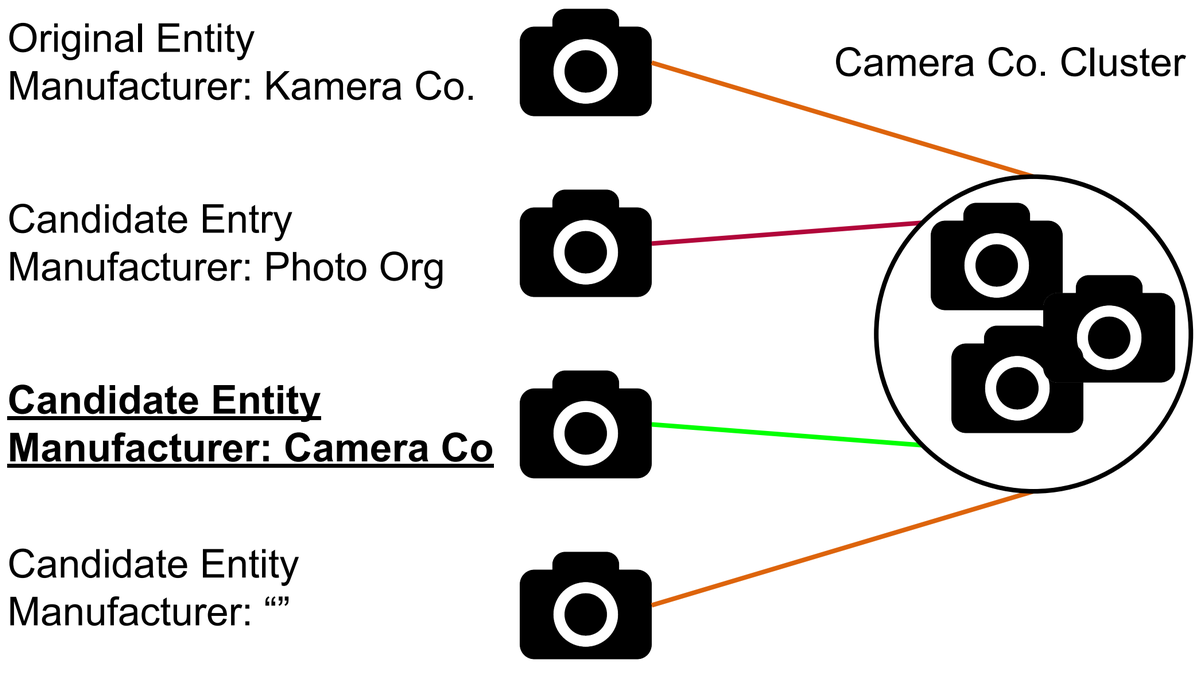
Investigate whether or not incorrect entities exhibit unusual clustering behaviours or metadata compared to their correct counterparts, for example:

1. Forming clusters of a single entity especially if a “clean” data source is present
2. Entities that are normally connected in a cluster that is otherwise strongly linked
3. Entities that switch clusters often during cluster repairs.



Correcting Entities

Once an aberrant entity has been identified, we can reintroduce different candidate versions of the entity to the graph, As FAMER allows entities to be added after the fact.
If one of the reintroduced candidate entities fits into the graph much better than before it might be a corrected version of the entity.
Candidate Entries are generated by replacing individual entity attributes with either values extracted from the cluster, a golden source or simply omitted.



Julius von Streit
Master Student
ITSE

E-Mail: julius.streit@student.hpi.de

Lecture Series on Database Research

Based on the Lecture by Prof E Rahm

Icons by Flaticon.com Freepik

Connection to lecture:
Leveraging the concepts introduced in the lecture by E. Rahm for Error detection and cleaning

HPI Hasso Plattner Institut