Reading List for Entity Resolution

Rebecca C. Steorts

May 17, 2020

Reading List for Entity Resolution

Below, you will find a reading list of entity resolution papers.

Introductory Tutorials and Books

Tutorials and Books on Entity Resolution:

- Some of Entity Resolution Tutorial (https://github.com/cleanzr/record-linkage-tutorial), Steorts et al (2020).
- Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Christen (2012).

Seminal Papers and Extensions

Seminal papers on Record Linkage (Fellegi-Sunter method):

- Dunn (1946): https://www.datanetwork.org/wp-content/uploads/2017/02/HL-Dunn-Record-Linkages.pdf
- Newcombe et al. (1959): https://www.cs.umd.edu/class/spring2012/cmsc828L/Papers/Newcombe59.pdf
- Tepping (1968): https://www.tandfonline.com/doi/abs/10.1080/01621459.1968.10480930 https://books.google.com/books?hl=en&lr=&id=9yL5HMBUnFQC&oi=fnd&pg=PA39&ots=6Neg0lO3VI&sig=ugJFz4rY98myLrzfB9
- Fellegi and Sunter (1969): https://courses.cs.washington.edu/courses/cse590q/04au/papers/Felligi69.pdf

Follow up papers (just a few):

- Winkler (1988)
- Jaro (1989): https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1989.10478785#.XsGRBBNKg0o
- Winkler and Thibaudeau (1991): http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.39.2433&rep=rep1&type=pdf
- Thibaudeau (1993):
- Sadinle and Fienberg (2013): https://amstat.tandfonline.com/doi/abs/10.1080/01621459.2012.757231#.XsGRxRNKg0o
- Murray (2016): https://journalprivacyconfidentiality.org/index.php/jpc/article/view/643
- Enamorado, Fifield, and Imai (2019): https://imai.fas.harvard.edu/research/linkage.html

Blocking papers:

- Steorts, Ventura, Sadinle, Fienberg (2014), https://link.springer.com/chapter/10.1007/978-3-319-11257-2_20 https://arxiv.org/abs/1407.3191
- Mining Massive Datasets, http://mmds.org/, Chapter 3.

- Introduction to LSH, https://github.com/resteorts/data-mine/blob/master/lectures_2018/03-hash/03-lsh.pdf.
- Steorts and Shrivastava (2018): https://arxiv.org/abs/1810.05497
- Sadosky et al. (2015): https://arxiv.org/abs/1510.07714

String distance papers:

- Sadinle and Fienberg (2013), http://stat.cmu.edu/NCRN/PUBLIC/RLClassFiles/Lectures/Sadinle-Fienberg-JASA-2013.pdf
- Cohen, Ravikumar, and Fienberg (2003) http://www.cs.cmu.edu/~wcohen/postscript/ijcai-ws-2003.pdf

Readings on semi-supervised methods:

- *Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Ed.*, Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009). http://statweb.stanford.edu/~tibs/ElemStatLearn/ (Covers logistic regression, support vector machines)
- Ventura and Nugent (2014), https://link.springer.com/chapter/10.1007/978-3-319-11257-2_28
- Ventura, Nugent, and Fuchs (2015), https://www.sciencedirect.com/science/article/pii/S0048733314002406?via%3Dihub

Review papers on Record linkage:

- Winkler (1995): https://books.google.com/books?hl=en&lr=&id=suacGGQgkwcC&oi=fnd&pg=PA355&dq=Winkler+(
68c0buwln-0c7k7esl0#v=onepage&q=Winkler%20(1988)%20record%20linkage&f=false
- Christen (2019): https://hdsr.mitpress.mit.edu/pub/8fm8lo1e/release/2

Background for Bayesian Entity Resoultion papers

- Copas and Hilton (1990): https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2982975
- Fortini et al. (2001): https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2982975

Bayesian Fellegi-Sunter papers

- Gutman et al. (2013): https://amstat.tandfonline.com/doi/abs/10.1080/01621459.2012.726889#.XsGdVxNKiu4
- Sadinle (2014): https://arxiv.org/abs/1407.8219
- Sadinle (2016): https://arxiv.org/abs/1601.06630
- Sadinle (2018): https://arxiv.org/abs/1812.09590
- McVeigh et al. (2020): https://arxiv.org/abs/1905.05337

Bayesian Graphical Entity Resolution papers

- Tancredi and Liseo (2011): https://projecteuclid.org/download/pdfview_1/euclid.aoas/1310562733
- Steorts, Hall, Fienberg (2014): https://arxiv.org/abs/1403.0211
- Steorts, Hall, Fienberg (2016): https://arxiv.org/abs/1312.4645
- Steorts (2015): https://arxiv.org/abs/1409.0643
- Marchant, Steorts, Kaplan, Rubinstein, and Elzar (2020): https://arxiv.org/abs/1909.06039

Microclustering papers

- Betancourt et al. (2016): https://arxiv.org/abs/1610.09780
- Betancourt et al. (2020): https://arxiv.org/abs/2004.02008
- Johndrow et al. (2018): https://arxiv.org/abs/1703.04955

- Steorts et al. (2017): https://arxiv.org/abs/1703.02679

Two-stage regression/MSE and Entity resolution

- Lahiri and Larsen (2005): https://amstat.tandfonline.com/doi/abs/10.1198/016214504000001277#.XsHMYxNKj_Q
- Kim and Chambers (2012): https://www.sciencedirect.com/science/article/abs/pii/S0167947312001089
- Goldstein et al. (2012): https://jech.bmj.com/content/66/12/1198.2
- Hof and Zwinderman (2012): https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5498
- Sadinle (2018): https://arxiv.org/abs/1812.09590

One-stage regression/MSE and Entity resolution

- Gutman et al. (2013): https://amstat.tandfonline.com/doi/abs/10.1080/01621459.2012.726889#.XsGdVxNKiu4
- Hof et al. (2017): https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1311262
- Dalzell and Reiter (2018): https://www.tandfonline.com/doi/abs/10.1080/10618600.2018.1458624
- Steorts, Tancredi, Liseo (2018): https://arxiv.org/abs/1810.04808
- Tanredi, Steorts, Liseo (2020): https://projecteuclid.org/euclid.ba/1551949260
- Tang, Reiter, Steorts (2020): In preparation.

Canonicalization (Data fusion papers)

- Yan and Ozsu (1999): https://ieeexplore.ieee.org/abstract/document/792177
- Bleiholder and Naumann (2009): https://dl.acm.org/doi/pdf/10.1145/1456650.1456651
- Cohen and Sagiv (2005): https://dl.acm.org/doi/pdf/10.1145/1099554.1099674
- Culotta et al. (2007): https://dl.acm.org/doi/abs/10.1145/1281192.1281217
- Kaplan, Betancourt, Steorts (2020): https://arxiv.org/abs/1810.01538

Open Source Software

R implementations:

- RecordLinkage package: https://www.rdocumentation.org/packages/RecordLinkage/versions/0.4-12
- fastlink: https://github.com/kosukeimai/fastLink
- blink: Steorts (2015): https://github.com/cran/blink
- representr: Kaplan et al. (2020): https://github.com/cleanzr/representr

Spark (Java and scala implementations):

- dblink: Marchant et al. (2020): https://github.com/cleanzr/dblink

C++ and Python:

- fasthash: Chen, Shrivastava, and Steorts (2018): https://github.com/cleanzr/fasthash

Julia implementations:

- McVeigh et. al (2020): https://github.com/brendanstats/BayesianRecordLinkage.jl

Data Sets

Synthetic Data Sets

- RLdata500 and RLdata10000: These can be found in the Record Linkage package in R.

Real Data Sets

- SHIW: https://github.com/ngmarchant/shiw
- NLTCS: As described in Steorts, Hall, Fienberg (2016). This cannot be shared on any public spaces.
- ABSEmployee: As described in Marchant et al. (2020). This cannot be shared on any public spaces.
- NCVR: As described in Christen (2012). This cannot be shared on any public spaces.
- Syrian data set: As described in Chen et al. (2018). This data cannot be shared on public spaces. There are subsets that are public that were utilized in Tancredi, Steorts, and Liseo (2020).
- El Salvadorian data set: As described in Sadinle (2014). This data is public.
- U.S. Census Data: One must obtain special permission to access this data, and this is a long and rigourous process involving background checks.

Other data sets

- I'm interested in putting together some synthetic and real data sets this summer for broader use of the community and the research lab's use. Please let me know if this is of interest to you. One involves citation data, where we would build upon an existing data set (cora). One would involve twitter data.