Record Linkage Project: Brainstorming Ideas

Project Presentations Oct 9th, 14th

Projects will be generated on some aspect of record linkage. Undergraduate students will work in pairs, master's students on their own (hereafter called a "group"). Each group will present its project (pdf or ppt) on October 9th or 14th. Presentations should be 15 min. A report describing the project and its results is due by XXXXXXXX. Slides should also be turned in at that time.

Brainstorming Project Ideas:

- Explore different text similarity metrics; how do they perform? What are their advantages/disadvantages?
- Use SoundEx to find similarity?? (exists in Record Linkage library)
- Different classifiers? feature selection methods for the classifiers?
- What if we modeled the dependency between fields in the Fellegi-Sunter?
- Using an E-M to fit the Fellegi-Sunter; logistic regression?
- Compare different blocking schemes; sequential blocking schemes?
- Clustering algorithms to assign records to unique entities; how do they behave? advantages/disadvantages?
- Given a set of labels, how sensitive are your methods? can we "optimize"?
- What else?

Data Sets:

- data sets available in Record Linkage library
- USPTO patent data (from Sam)
- UCI Machine Learning Repository - Record Linkage Comparison Data Set
- Kaggle/Microsoft/KDD Challenge - "Author Disambiguation Challenge" `https://www.kaggle.com/c/kdd-cup-2013-author-disambiguation`
- ICPSR (includes data on the National Long Term Care Survey)
- FEBRL; generation/simulation (Peter Christen)