Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup.

Text mining is a field of change. What is clear is no matter what assignment facing the text miners today their will a fair amount of disagreement when it comes to the manner in which they procead. The KDD challenge cup was a “golden scenario” for text miners today. A clean set of easily readable HTML with a realistic and fair extraction assignment. Gather information on genes in full text papers and designate if they warrant curation.

Really two notable findings arose from the KDD challenge cup. One: that as much as programmers hate it “natural language” extraction was the way to go when it comes to text mining. Also notably second that removing the “extra noise” or other unneeded data was key mining well.

In our discussion we talked a lot about the standardization of the data. For a while we thought that clean data was the excuse used by the data miners to sluff off the real world responsibility of dealing with hard realities of data conversions. But they we thought maybe the assumption of the KDD challenge cup providers that test data should be mined from more golden data was not really that extreme a stance. In truth the best data set is the cleanest. As database creator’s aggregators and curators it is some part the job of informational professional to help create the best data. To start with the XML revolution gives us a better platform with better Meta tags. It also gives a better ability to do well in the KDD challenge cups second finding “removing the noise.” XML makes it much easier to dump the display-coding present in HTML publishing and especially PDF. But even if we have a standard coding language we still have whole human luggage factor to overcome. “No matter what standards are enacted catalogers and programs will inevitably not follow them,” said a number of people.

In the end it was hard to come a consensus about what the future would hold for text mining. But we thought that answer lay in a two-pronged attack that started with a standardization of the mined data and ended in better extraction that used more natural language forms.