Published on *xpriori.com* (http://xpriori.com)

Dealing effectively with Big Data...automatic document classification and effective culling

Submitted by tdix on Fri, 2013-05-10 09:08

One of my colleagues forwarded this link this morning. It is a thoughtful discussion of Big Data: http://qz.com/81661/most-data-isnt-big-and-businesses-are-wasting-money-... [1]

The author notes that people tend to deal effectively with data once it is clustered with similar data into relatively smaller batches. This is especially true when dealing with large amounts of unstructured information. We are now working with a series of advanced technologies that enables automatic clustering of data to a very granular level and provides tools to look at and manage this clustered information deploying -- simplified coding for extraction; mass error correction and redaction among others and things that you accomplish can be persistent throughout the dataset. It really helps to cut many of the issues of big data down to size.

We find that applying many analytical tools across large amounts of unclassified unstructured data can produce a lot of blind alleys. These tools work better on things that are similar contextually. This is particularly true when working with semi-structured information such as business forms or spreadsheets where the position of the information on the page or in the form provides the context for the datum. Without this type of classification, the author notes: "The “bigger” your data, the more false positives will turn up in it, when you’re looking for correlations." The same is true for conceptual associations.

The problem in big data is not the cost of storage, it is providing meaningful access and finding associations that require less time to sort through. We are finding great success with this new tool. The author concluded his article with the following: "Remember: Gregor Mendel uncovered the secrets of genetic inheritance with just enough data to fill a notebook. The important thing is gathering the right data, not gathering some arbitrary quantity of it."

.

Xpriori, LLC ? 2864 South Circle Drive, Suite 401 ? Colorado Springs, CO 80906 ? 719-425-9840 ? Fax 719-203-6496?© 2014-2018

Source URL: http://xpriori.com/dealing-effectively-big-data-automatic-document-classification-and-effective-culling

Links:

[1] http://qz.com/81661/most-data-isnt-big-and-businesses-are-wasting-money-pretending-it-is/