PubMed Visaulization

Anirudh Negi, Junfeng Zhao, Yibo Zhou

November 8, 2021

Abstract

Currently there are millions of published articles and papers stored in a massive database for all articles published in medical journals since the 1960s. However, not only is it difficult to search through these articles, the sheer volume of data makes gathering insights next to impossible. We propose a hopefully permanent solution to this issue by creating visualizations of all the data that can be updated in a streaming context. The primary techniques to extract these visualizations will be clustering and ranking.

1 Goals

- Can we rank the "trustworthiness" of attributes like author, journal or article using PageRank?

- Can we visualize associations among attributes like author, journal or articles?

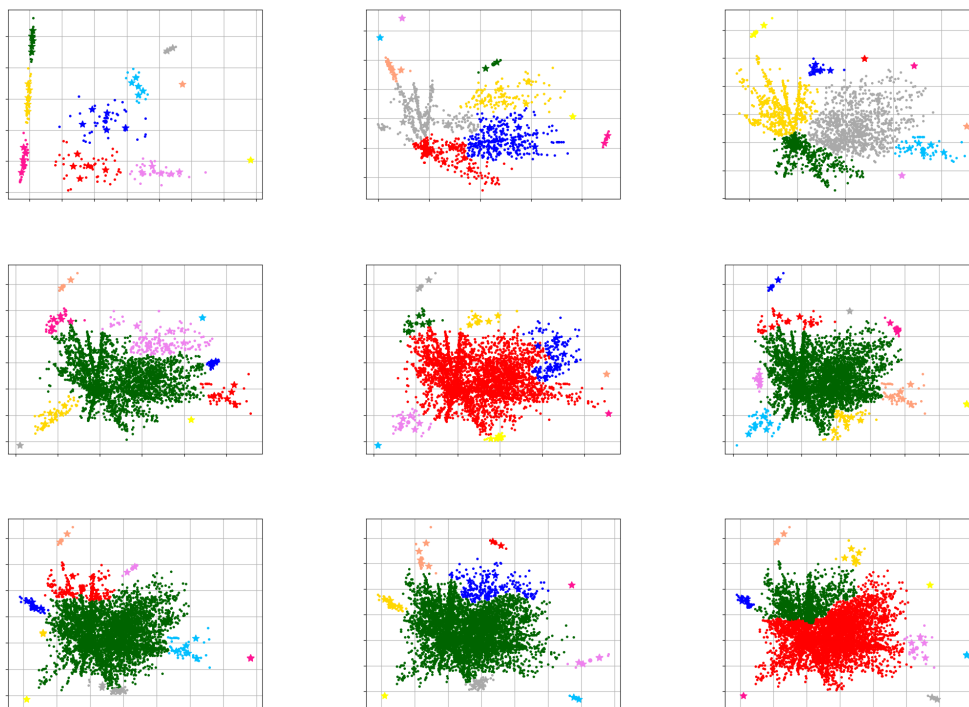- Can we process a massive dataset (30 million rows, > 10GB) efficiently using a streaming context? What will be the major challenges here?

2 Visuals

Figure 1: Processing of CURE