Making Interval-Based Clustering Rank-Aware

Julia Stoyanovich, UPenn

In online applications such as online dating, users often query and rank large collections of structured items. Top results tend to be homogeneous, which hinders data exploration. For example, a dating website user who is looking for a partner between 20 and 40 years old, and who sorts the matches by income from higher to lower, will see a large number of matches in their late 30s who hold an MBA degree and work in the financial industry, before seeing any matches in different age groups and walks of life. An alternative to presenting results in a ranked list is to find clusters in the result space, identified by a combination of attributes that correlate with rank. Such clusters may describe matches between 35 and 40 with an MBA, matches between 25 and 30 who work in the software industry, etc., allowing for data exploration of ranked results.

We refer to the problem of finding such clusters as rank-aware interval-based clustering and argue that it is not addressed by standard clustering algorithms. We formally define the problem and, to solve it, propose a novel measure of locality, together with a family of clustering quality measures appropriate for this application scenario. These ingredients may be used by a variety of clustering algorithms, and we present BARAC, a particular subspace clustering algorithm that enables rank-aware interval-based clustering in domains with heterogeneous attributes. We validate the effectiveness of our approach with a large-scale user study, and perform an extensive experimental evaluation of efficiency, demonstrating that our methods are practical on the large scale. Our evaluation is performed on large datasets from Yahoo! Personals, a leading online dating site, and on restaurant data from Yahoo! Local.