I present a new similarity score, based on a statistical model, that is useful for clustering problems with high missing data rates and discrete data values. In settings that range from genomics to recommender systems, I demonstrate how this score can be used to develop fast algorithms for large-scale clustering. Together with collaborators at Lawrence Berkeley National Lab, UC Berkeley, and UC Santa Barbara, I developed the new similarity score by comparing the likelihood of observed data under an assumed clustering model, to the probability of observing the same data by chance. The advantage of our score over traditional similarity scores is its ability to leverage more data to make more accurate similarity comparisons, as long as a certain underlying clustering structure exists in the data. We applied our score to the recommender systems domain, where the challenges of high missing data rates and high dimensionality abound. We have shown that this new score is more effective at identifying similar users than traditional similarity scores, such as the Pearson correlation coefficient, in user-based collaborative filtering. We argue that our approach has significant potential to improve both accuracy and scalability in collaborative filtering. In ongoing work, we are building on the success of this similarity score in the sparse recommender systems setting to design new clustering algorithms for general discrete-valued data with high missing rates.