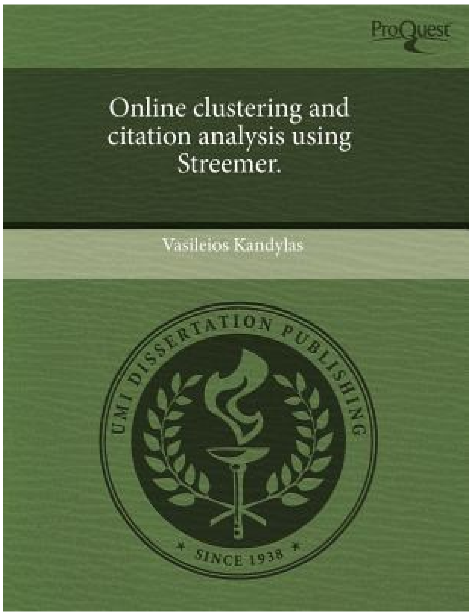
ProQuest
Online clustering and citation analysis using Streemer.
Vasileios Kandylas
UMI DISSERTATION PUBLISHING
SINCE 1938

Online Clustering and Citation Analysis Using Streemer

Vasileios Kandylas

Clustering algorithms can be viewed as following an algorithmic or a probabilistic approach. Algorithmic methods such as k-means or streaming clustering are fast and simple but tend to be ad hoc and hence hard to customize to particular problems, whereas the probabilistic methods are more flexible, but slower. In this work we propose online algorithms which combine the advantages of the two classes of approaches giving fast, scalable clustering, while allowing more flexible models of the data, such as foreground clusters interspersed within a diffuse background. These clusters are shown to be useful in modeling scientific citations. We start the thesis by giving a non-probabilistic, few-pass algorithm, called Streemer. Streemer uses thresholds on similarities between points to find a large number of clusters on the first pass over the data. It then merges them to find larger and more cohesive clusters. In a final pass it assigns points to the clusters or to a diffuse background. Streemer avoids the standard k-means assumptions that clusters are of similar sizes. We also discuss the nature of the objective function that Streemer optimizes through its several steps and heuristics. At a cursory glance, Streemer appears to be an ad hoc algorithm, but in a subsequent chapter we develop a principled algorithm that emulates Streemers steps and we make the connection between Streemer and online Dirichlet Process Mixture Models. We use Streemer to cluster documents based on the documents they cite and find knowledge communities of authors that build on each others work. The evolution over time of these clusters gives us insight into their growth or shrinkage. We also build predictive models with features based on the citation structure, the vocabulary of the papers, and the affiliations and prestige of the authors and use these models to study the drivers of community growth and the predictors of how widely a paper will be cited. The analysis shows that scientific knowledge communities tend to grow more rapidly if their publications build on diverse information and use narrow vocabulary and that papers that lie on the periphery of a community have the highest impact, while those not in any community have the lowest impact. We also present a probabilistic mixture model with a Dirichlet Process prior and Gaussian component distributions. This model allows for variable cluster numbers and sizes. We show how to use this model for clustering in an online fashion and also propose a two-pass algorithm, where the first pass clusters points in many clusters and the second pass clusters the output of the first pass. With the exception of foreground/background clustering, the model with the two-pass algorithm corresponds closely to Streemer. Finally, we present an EM-based clustering method that can simultaneously cluster two or more variables using one or more tables of co-occurrence data. One application of this multi-way clustering algorithm is for constructing or augmenting ontologies. We test our algorithm by simultaneously clustering verbs and nouns using both verb-noun and noun-noun co-occurrence pairs. This strategy provides greater coverage of words than using either set of pairs alone, since not all words appear in both datasets. We demonstrate it on data extracted from Medline and evaluate the results using MeSH and Wordnet.