**JUSTICE NII-AYITEY**

**STAT517 PROJECT #1**

**Recent Developments in Document Clustering**

In the context of machine learning, classification is supervised learning and clustering is unsupervised learning. Classification has prior knowledge of classes and classify new sample into known classes. It uses Decision Trees, Bayesian classifiers, etc. as its algorithms and has labeled samples from a set of classes. On the other hand, clustering has no prior knowledge of classes and suggest groups based on patterns in data and uses K-means, Expectation Maximization, etc. as its algorithms with unlabeled samples. Clustering are better when collections are organized into groups such that each group has similar documents and comparatively different to other groups. Document clustering also known as text clustering is the application of cluster analysis to textual documents. It has applications in automatic documents, topic extraction and fast information retrieval or filtering of which this report focuses on its recent developments.

In a clustering problem, the first challenge is to determine which features of a document are to be considered discriminatory which simply means the document model. Most existing clustering approaches choose to represent each document as a vector, therefore reducing a document to a representation suitable for traditional data clustering approaches. Clustering algorithms can be divided into discriminative and generative types of which discriminative algorithms allow to classify points without providing a model of how the points are generated while generative algorithms make structure assumptions on the model. The generative algorithms include Gaussian model, Expectation maximization, von Mises-Fisher model and Model-based kmeans. In general, generative model need to model more than the discriminative and hence are sometimes not as effective. However, generative models often outperform discriminative models on smaller datasets because their generative assumptions place more structure on the model to prevent overfitting.

There are quite several ways to evaluate clustering algorithms but also, there are few uncertainties as in to which is the best way. The evaluation methods are based on the fact on which the research was being conducted or performed. Two ways to check for the performances (accuracy) of good, bad, and ugly clustering is by using their precision and recall. In information retrieval, precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances, hence both precision and recall are therefore based on an understanding and measure of relevance. Given R as recall and P as precision, the generalized F-measure is defined as;

Fα=

Although vector model is usually developed for automatic indexing but also with it, a collection of ‘p’ documents with ‘q’ unique terms is represented as an ‘p x q’ term-document matrix (where each document is a vector of m dimensions). Vector space representation in one way or another used mostly clustering algorithms. At this point, the vector model is sometimes called the bag of words or dictionary model because no information about the word order is encoded. Also, there are two vital properties under the vector space model. First, in a collection of heterogeneous topics, the number of unique terms will be quite large which results in document vector of high dimensionality and second, a matrix resulting from a typical corpus under the vector model will be highly sparse. Although high dimensionality could be reduced by PCA, Nonnegative matrix factorization, Soft spectral coclustering, and Lingo but the properties listed above need several preprocessing steps that take as input a plain text document and output a set of tokens to be included in the vector model. These steps typically consist of but not limited to filtering, tokenization, stemming, stopword removal, and pruning.

Historically, hierarchical and partitional algorithms have been the dominant clustering methods as many recent developments have basis in theses approaches. In hierarchical clustering, each document is initially its own cluster while in the partitional methods (of which the classical examples is kmeans) starts by choosing k initial documents as clusters, and iteratively assign documents to clusters while updating the centroids of these clusters. The extensions to kmeans include Online spherical kmeans and Kernel kmeans,. As part of the recent developments, Spectural clustering such as Divide & merge clustering and Fuzzy coclustering find cuts in graphs that produce good clusters. Also, Phrase-based models can naturally describe clusters by phrases, and it generally agreed that these are more descriptive of the cluster contents which includes Suffic tree clustering and Document index graph. Comparative analysis could also be performed using Query clustering and Collection clustering.

In conclusion, although most algorithm described above can produce good clustering according to typical measures, the most “natural” groupings are not always produced.