An Introduction to Unsupervised Document Classification (Text Categorization)

Why is it important?

1. Determines information about corpus contents
2. Can save time in document retrieval when handling queries
3. Can be used to auto sort document
4. Text filtering applications, such as spam detection
5. Authorship attribution (the documents may cluster based on stop word usage?)
6. Text Cat for targeted advertisement, such as Google's Adsense

How is it currently done?

1. Centroid methods: sum, roochio, average, distance, k-means
2. SVMs: Single class, multiple class training
3. Neural Networks
4. Naive Bayesian

What is the Next Step for clustering?

1. Do queries with the name or similar information need cluster by the context of the query and possibly by the results they return. Therefore many queries may satisfy the same information need; therefore only need to execute it once--depending on the nature of the corpus
2. Automatically categorizing websites; possibly as offensive to certain audiences (i.e. spam versus not); possibly to help categorize search results to provide additional information through clusters of documents (although this is currently done).

Unsupervised document classification is the science of categorizing a document based on its contents. This tool for text categorization is very powerful for multiple applications. Text categorization provides important information about the contents of a corpus and can provide speedup to query processing. Documents can be sorted automatically based on their contents using this technology. Services such as Google's Adsense may use a method of text categorization to determine which ads appear on a website or email based on the contents of the document. The exact method Google uses may have more to do with key words and is proprietary information. However, text cat can be used to solve their problem; albeit possibly less efficiently.

Currently there are several primary methods of categorizing documents. An important method is centroid based, whereby documents that are similar in content and therefore category will drift towards each other and have the same centroid. The values for the centroid calculations vary from nearest neighbor to sum, average, etc. Support vector machines are also powerful in this task. Cannot forget Naive Bayesian methods. Also an important method involves the usage of neural networks.