How Informative are the Concept Associations Discovered by a Document Cluster?

Vinay Deolalikar, Deependra Singh
Groupon Data Mining
Palo Alto, CA 94040

Abstract. Over the past two decades, document clustering has matured into a fundamental data mining technique, with potential for use in diverse areas in industry. In this paper, we address an aspect of the structure of document clusters: namely, the information captured in concept associations that are discovered by document clustering. Can we *measure* this information in a principled and statistically sound manner? This problem is germane to several applications of data mining to industry, including eDiscovery and homeland security.

We approach this problem as follows. First, we construct representations for concepts in document clusters using frequent itemsets. We then use these representations to frame the notion of associations between concepts. These are the associations that the document cluster has "discovered." Next, we use information theory to measure the information contained in these concept associations. This requires us to measure what is the prior probability of such an association occurring, based on the statistics of the corpus. Information theory tells us that the more unlikely this association, the more informative it is.

Our approach results in two information theoretic scores for each document cluster. These scores are interpretable: the user can be informed why a cluster has a certain score. We also show how to implement our techniques with very little overhead by using structures that are already computed during document clustering. Finally, we demonstrate the flexibility of our framework by using it to perform a variety of tasks centered around concept associations and their information content on benchmark datasets for document clustering.