Abstract

Text Clustering is a problem of dividing text documents into groups, such that documents in one group are more similar than those in other groups. Although comparisons of the different algorithms have been done in an attempt to choose some over the others, such comparisons have been found to be either too limited or inadequate. In such comparisons, either the researchers (who are usually the authors of the algorithms being compared with others) did not apply a formal comparison methodology, or the comparisons were based on inadequate data, metrics and procedures. Also, the comparisons always focus on only the aspects where their algorithms are superior to the other algorithms. The few algorithms being compared with theirs obviously seem to be carefully selected such that they are the ones performing lesser than theirs on those aspects. Thus, there is still a large gap on the most suitable methodology for comparing the algorithms.

In this paper, a methodology for fairly comparing text clustering algorithms is proposed.
References


Index Terms

Computer Science Information Sciences

Keywords

Clustering, Text Clustering, Metrics.