Abstract

As well actual clustering algorithms have to deal with explosive growth of documents of various sizes and terms of various frequencies, an appropriate term-weighting scheme has a crucial impact on the overall performance of such systems. Term-weighting is one of the critical process for document retrieval and ranking in most search result clustering systems. In this paper we introduce a new technique for clustering algorithms that solve the problem of indexing the terms of big datasets and their characteristics which exist in most of current clustering approaches. The paper focus on term frequency normalization step of clustering algorithms. A new factor has been applied to basic term-weighting schemes for using in clustering process. The evaluated results confirm the impact of this factor to increase the performance of clustering techniques. The experiments were carried out on the standard algorithms and ODP-239 datasets which validated by statistical tests.

Index Terms

Computer Science

Information Sciences
Keywords
Weighted clustering  Term importance degree  Term frequency normalization