Abstract

The dynamic web has increased exponentially over the past few years with more than thousands of documents related to a subject available to the user now. Most of the web documents are unstructured and not in an organized manner and hence user facing more difficult to find relevant documents. A more useful and efficient mechanism is combining clustering with ranking, where clustering can group the similar documents in one place and
ranking can be applied to each cluster for viewing the top documents at the beginning. Besides the particular clustering algorithm, the different term weighting functions applied to the selected features to represent web document is a main aspect in clustering task. Keeping this approach in mind, here we proposed a new mechanism called Tf-Idf based Apriori for clustering the web documents. We then rank the documents in each cluster using Tf-Idf and similarity factor of documents based on the user query. This approach will helps the user to get all his relevant documents in one place and can restrict his search to some top documents of his choice. For experimental purpose, we have taken the Classic3 and Classic4 datasets of Cornell University having more than 10,000 documents and use gensim toolkit to carry out our work. We have compared our approach with traditional apriori algorithm and found that our approach is giving better results for higher minimum support. Our ranking mechanism is also giving a good F-measure of 78%.

References

- http://tartarus.org/martin/PorterStemmer/def.txt
- Jaroslav Pokorhy, Jozef Smizansky. Page Content Rank, An approach to the web content mining.
- http://en.wikipedia.org/wiki/F1_score

**Index Terms**

Computer Science  Web Engineering

**Keywords**

Apriori  Clustering  Gensim  Ranking  Vector Space Model