Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 10, October 2013, pg.131 – 137

RESEARCH ARTICLE

Text Document Clustering Using DPM with Concept and Feature Analysis

S Kajapriya[1], K.N Vimal Shankar[2]

PG Scholar[1], Asst. Professor[2]
Department of Computer Science & Engineering[1, 2]
V.S.B. Engineering College, Karur, India[1, 2]
Kpriya24san@gmail.com [1], yvsinformation@yahoo.in [2]

Abstract— Clustering is one of the most important techniques in machine learning and data mining tasks. Similar documents are grouped by performing clustering techniques. Similarity measuring is used to determine transaction relationships. Hierarchical clustering model produces tree structured results. Partitioned based clustering produces the outcome in grid format. Text documents are unstructured data values with high dimensional attributes. Document clustering group ups unlabeled text documents into meaningful clusters. Traditional clustering methods require cluster count (K) for the document grouping process. Clustering accuracy degrades drastically with reference to the unsuitable cluster count. Document features are automatically partitioned into two groups' discriminative words and nondiscriminative words. Only discriminative words are useful for grouping documents. The involvement of nondiscriminative words confuses the clustering process and leads to poor clustering solution in return. A variation inference algorithm is used to infer the document collection structure and partition of document words at the same time. Dirichlet Process Mixture (DPM) model is used to partition documents. DPM clustering model uses both the data likelihood and the clustering property of the Dirichlet Process (DP). Dirichlet Process Mixture Model for Feature Partition (DPMFP) is used to discover the latent cluster structure based on the DPM model. DPMFP clustering is performed without requiring the number of clusters as input. Discriminative word identification process is improved with the labeled document analysis mechanism. Concept relationships are analyzed with Ontology support. Semantic weight model is used for the document similarity analysis. The system improves the scalability with the support of labels and concept relations for dimensionality reduction process. The system development is planned with Java language and Oracle relational database.

Keywords— Database management; Dirichlet Process Mixture Model; Document Clustering; Feature Partition; Semi-Supervised; Text mining

---

Full Text: http://www.ijcsmc.com/docs/papers/October2013/V2I10201332.pdf