An Efficient Concept-Based Mining Model for Enhancing Text Clustering

Abstract

Clustering, one of the traditional data mining techniques, is an unsupervised learning paradigm where clustering methods try to identify inherent groupings of the text documents, so that a set of clusters is produced in which clusters exhibit high intra cluster similarity and low inter cluster similarity. Methods used for text clustering include decision trees, conceptual clustering, clustering based on data summarization, statistical analysis, neural nets, inductive logic programming, and rule-based systems among others. In text clustering, it is important to note that selecting important features, which present the text data properly, has a critical effect on the output of the clustering algorithm. A novel concept-based mining model is proposed. The proposed model captures the semantic structure of each term within a sentence and document rather than the frequency of the term within a document only. In the proposed model, three measures for analyzing concepts on the sentence, document, and corpus levels are computed.

System Analysis

Existing System:

Most of the common techniques in text mining are based on the statistical analysis of a term, either word or phrase. Statistical analysis of a term frequency captures the importance of the term within a document only. However, two terms can have the same frequency in their documents, but one term contributes more to the meaning of its sentences than the other term. Thus, the underlying text mining model should indicate terms that capture the semantics of text.

Disadvantages:

- Matching Concepts on individual terms.
- Very Sensitive.
- Performance is low.

Proposed System:

A new concept-based mining model that analyzes terms on the sentence, document, and corpus levels is introduced. The concept-based mining model can effectively discriminate between non important terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. The proposed mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. The term which contributes to the sentence semantics is analyzed on the sentence, document, and corpus levels rather than the traditional analysis of the document only. The proposed model can efficiently find significant matching concepts between documents, according to the semantics of their sentences. The similarity between documents is calculated based on a new concept-based similarity measure. The proposed similarity measure takes full advantage of using the concept analysis measures on the sentence, document, and corpus levels in calculating the similarity between documents. Large sets of experiments using the proposed concept-based mining model on different data sets in text clustering are conducted.

Advantages:

- High Quality.
- Matching Concepts on sentences and document.
- Using Sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis.
- Very Robust and Accurate.

SYSTEM REQUIREMENTS

- **H/W System Configuration:-**

- Processor - Pentium –IV
- RAM - 4 GB (min)
- Hard Disk - 20 GB
- Key Board - Standard Windows Keyboard
- Mouse - Two or Three Button Mouse
- Monitor - SVGA

- **S/W System Configuration:-**
- Operating System : Windows 7 or 8 32 bit
- Application Server : Tomcat5.0/6.X
- Backend coding : Java
- Tool : Virtual Box
- Environment : Ubuntu
- Technology : Hadoop