Chapter 10
Implementation and Testing Details of Document Classification

ABSTRACT

It is trivial to achieve a recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough, but one needs to measure the number of non-relevant, for example by computing the precision. The analysis was performed for 30 documents to ensure the stability of precision and recall values. It is observed that the precision of large documents is less than a moderate length document, in the sense that some unimportant keywords get extracted. The reason for this may be attributed to the frequent occurrence and its unimportant role in the sentence.

SYSTEM TESTING

Reuters Data Set

Researchers have used benchmark data, such as the Reuters- 21578 corpus of newswire test collection (Sholom M. W., Indurkhya, N., Zhang, T. and Damerau, F. 2010), to measure advances in automated text classification. We performed testing of our system using a sample of the same.

DOI: 10.4018/978-1-7998-3772-5.ch010

Modules of Execution

1. Document Entry
2. Stop Word removal
3. Stemming
4. Keyword generation
5. Document Classification

Document Entry

Doc_id : DOC1

Doc_content :

Table 1. Words after tokenization

hard
problem
text
classification
aspects
potential
solution
keyword
extraction
maximal
frequent
item
set
used
attributes
mining
association
rules
basis
measuring
similarity
new
documents
existing
association rules
issue
keyword
extraction
text
collection
emerging
research
filed
promotes
maximal
frequent
item
set
generation