

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., Indurkha, 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

9 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/implementation-and-testing-details-of-document-classification/268469

Related Content

Fusion Cubes: Towards Self-Service Business Intelligence

Alberto Abelló, Jérôme Darmont, Lorena Etcheverry, Matteo Golfarelli, Jose-Norberto Mazón, Felix Naumann, Torben Pedersen, Stefano Bach Rizzi, Juan Trujillo, Panos Vassiliadis and Gottfried Vossen (2013). *International Journal of Data Warehousing and Mining* (pp. 66-88).

www.irma-international.org/article/fusion-cubes-towards-self-service/78287

Enhancing the Process of Knowledge Discovery in Geographic Databases Using Geo-Ontologies

Vania Bogorny, Paulo Martins Engeland Luis Otavio Alavares (2008). *Data Mining with Ontologies: Implementations, Findings, and Frameworks* (pp. 160-181).

www.irma-international.org/chapter/enhancing-process-knowledge-discovery-geographic/7577

When Should We Ignore Examples with Missing Values?

Wei-Chao Lin, Shih-Wen Ke and Chih-Fong Tsai (2017). *International Journal of Data Warehousing and Mining* (pp. 53-63).

www.irma-international.org/article/when-should-we-ignore-examples-with-missing-values/188490

Combining Machine Learning and Natural Language Processing for Language-Specific, Multi-Lingual, and Cross-Lingual Text Summarization: A Wide-Ranging Overview

Luca Cagliero, Paolo Garza and Moreno La Quatra (2020). *Trends and Applications of Text Summarization Techniques* (pp. 1-31).

www.irma-international.org/chapter/combining-machine-learning-and-natural-language-processing-for-language-specific-multi-lingual-and-cross-lingual-text-summarization/235739

Optimal Features for Metamorphic Malware Detection

P. Vinod, Jikku Kuriakose, T. K. Ansari and Sonal Ayyappan (2014). *Data Mining and Analysis in the Engineering Field* (pp. 1-32).

www.irma-international.org/chapter/optimal-features-for-metamorphic-malware-detection/109973