Feature Selection Based on Term Frequency and T-Test for Text Categorization

Deqing Wang, Hui Zhang, Rui Liu, Weifeng Lv

(Submitted on 3 May 2013)

Much work has been done on feature selection. Existing methods are based on document frequency, such as Chi-Square Statistic, Information Gain etc. However, these methods have two shortcomings: one is that they are not reliable for low-frequency terms, and the other is that they only count whether one term occurs in a document and ignore the term frequency. Actually, high-frequency terms within a specific category are often regards as discriminators.
This paper focuses on how to construct the feature selection function based on term frequency, and proposes a new approach based on t-test, which is used to measure the diversity of the distributions of a term between the specific category and the entire corpus. Extensive comparative experiments on two text corpora using three classifiers show that our new approach is comparable to or or slightly better than the state-of-the-art feature selection methods (i.e., χ^2, and IG) in terms of macro-F_1 and micro-F_1.

Comments:	5pages 9 figures CIKM2012 paper
Subjects:	**Learning (cs.LG)**; Information Retrieval (cs.IR); Machine Learning (stat.ML)
Cite as:	**arXiv:1305.0638 [cs.LG]**
	(or **arXiv:1305.0638v1 [cs.LG]** for this version)

Submission history

Download:

- PDF
- PostScript
- Other formats

Current browse context:

cs.LG

< prev | next >

new | recent | 1305

Change to browse by:

cs
cs.IR
stat
stat.ML

References & Citations

- NASA ADS

Bookmark(what is this?)