Iterative Term Weighting for Short Text Data

**ABSTRACT**

With the development of social media applications, short text mining is becoming more and more important. Due to the sparseness of short text data, both the feature correlation information (word co-occurrence) and data contiguity information (context information) are less reliable, thus most existing text mining methods which are designed to address regular text data are less efficient in short text mining tasks. According to our observation from analysis of discriminative term distribution in short text data, we found that discriminative terms distribute in a non-uniform way among different domains, while background words have a tendency to distribute uniformly. This observation can be measured by a suitably defined functional of a term’s probability distribution over different domains. In this topic, we adopt this distribution as the weight of terms to address the sparseness problem of short text data. We evaluate our method on two datasets, and experimental results show that our method outperforms previous approaches which require information infusion, and a number of state-of-the-art clustering algorithms. Furthermore, our method can obtain a more coherent clustering result.

This paper will be presented at the IEEE International Conference on Systems, Man, and Cybernetics (SMC2015) held at City University of Hong Kong, October 9-12, 2015.

Supervisor: Dr WONG Hau San Raymond
Research Interests: Short Text Clustering; Manifold Learning

All are welcome!

In case of questions, please contact Dr WONG Hau San Raymond at Tel: 3442 8624, E-mail: cshswong@cityu.edu.hk, or visit the CS Departmental Seminar Web at [http://www.cs.cityu.edu.hk/news/seminars/seminars.html](http://www.cs.cityu.edu.hk/news/seminars/seminars.html).