Abstract

This paper shows document clustering by applying fuzzy logic. The method involves cleaning up the text and stemming of words. Then, chose ‘m’ features which differ significantly in their
word frequencies (WF), normalized by document length, between documents belonging to these two clusters. The documents to be clustered are represented as a collection of ‘m’ normalized WF values. Then use Fuzzy c-means (FCM) algorithm to cluster these documents into two clusters. After the FCM execution finishes, the documents in the two clusters are analyzed for the values of their respective ‘m’ features. By using fuzzy logic, we not only get the cluster name, but also the degree to which a document belongs to a cluster.

References

- Vishal Gupta, Gurpreet S. Lehal; “A Survey of Text Mining Techniques and Applications”; Journal of Emerging Technologies in Web Intelligence, Vol.1, No.1, August 2009
- Sumit Goswami, Sudesha Sarkar, Mayur Rustagi: Stylometric Analysis of Bloggers' Age and Gender. ICWSM 2009
- Sumit Goswami, Mayank Singh Shishodia; “A fuzzy based approach to stylometric analysis of blogger?s age and gender”; HIS 2012: 47-51
- Fuzzy logic vs Probability (Good Math, Bad Math); http://scientopia.org/blogs/goodmath/2011/02/02/fuzzy-logic-vs-probability/, last checked on 28th July 2012.

Index Terms

Computer Science Fuzzy Systems

Keywords

Fuzzy c-means algorithm, fuzzy logic, Document clustering