Automatic Classification of Formal and Informal Texts

Yinghui Xia

Follow this and additional works at: http://openscholarship.wustl.edu/vol8_iss1

Recommended Citation
http://openscholarship.wustl.edu/vol8_iss1/168

This publication is brought to you for free and open access by the Office of Undergraduate Research through Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu
Document classification is a topic in information science. In this study, I show an algorithm for automatically classifying text documents into formal or informal style. To get enough information of formality, people use large training corpora in many algorithms. This work proposes a method for extracting formality markers from a small, classified training corpus for assessing the formality of emails content using markers. The research compares the effectiveness of the classifiers in the built algorithm and two other classification methods. As a test case, I use Enron Email Database, a relevant email database of Enron Corporation, extract formality features from it and evaluate the performance of the markers in each algorithm. The results of this study show that several hundred classifiers are able to predict the classes of new texts with high accuracy. The training corpus for text classification does not have to be large to provide enough information for document classification. Thus, it greatly improves the speed of current classification algorithms.