

Dropping down the maximum item set: improving the stylometric authorship attribution algorithm in the text mining for authorship investigation

ABSTRACT

Problem statement: Stylometric authorship attribution is an approach concerned about analyzing texts in text mining, e.g., novels and plays that famous authors wrote, trying to measure the authors style, by choosing some attributes that shows the author style of writing, assuming that these writers have a special way of writing that no other writer has; thus, authorship attribution is the task of identifying the author of a given text. In this study, we propose an authorship attribution algorithm, improving the accuracy of Stylometric features of different professionals so it can be discriminated nearly as well as fingerprints of different persons using authorship attributes. Approach: The main target in this study is to build an algorithm supports a decision making systems enables users to predict and choose the right author for a specific anonymous author's novel under consideration, by using a learning procedure to teach the system the Stylometric map of the author and behave as an expert opinion. The Stylometric Authorship Attribution (AA) usually depends on the frequent word as the best attribute that could be used, many studies strived for other beneficiary attributes, still the frequent word is ahead of other attributes that gives better results in the researches and experiments and still the best parameter and technique that's been used till now is the counting of the bag-of-word with the maximum item set. Results: To improve the techniques of the AA, we need to use new pack of attributes with a new measurement tool, the first pack of attributes we are using in this study is the (frequent pair) which means a pair of words that always appear together, this attribute clearly is not a new one, but it wasn't a successive attribute compared with the frequent word, using the maximum item set counters. the words pair made some mistakes as we see in the experiment results, improving the winnow algorithm by combining it with the computational approach, achieved by using the CV statistical tool as a conditional threshold for attribute selecting; by doing so, the frequent pair result improved from 50% error to 0% in the improved frequent pair with a clear higher score result compared with the frequent word attribute. Conclusion/Recommendations: The new CV algorithm results improvement may lead to several new attributes usage that gave unsatisfying results before that might improve the direction for solving some hard cases couldn't be solved till now.

Keyword: Text mining; Stylometric attribution; Authorship attribution; Winnow algorithm; Computational stylistic