Abstract

Suffix Tree Clustering (STC) uses the suffix tree structure to find a set of snippets that share a common phrase and uses this information to propose clusters. As a result, STC is a fast incremental algorithm for automatic clustering and labeling but it cannot cluster semantically similar snippets. However, the meaning of the words is indeed an important property that relates them to other words, although there may not be a match of text strings per se. In this paper, we propose a new semantic search results clustering algorithm, called semantic suffix net clustering (SSNC). It is based on semantic suffix net structure (SSN). The proposed algorithm uses the net pruning technique to merge the related suffixes through their suffix links for finding base clusters. This logic causes both string matching and meaning of the words to be used as conditions for the purpose of clustering. Experimental results show that the proposed algorithm has time complexity lower than CFWMS, SSTC and STC+GSSN which are current semantic search results clustering methods. Moreover, the F-measure of the proposed algorithm is similar to that of the original STC, CFWMS, STC+GSSN, and higher than that of MSRC and SSTC.

Index Terms

Computer Science Algorithms
Keywords

search results clustering  semantic suffix net  net pruning techniques  semantic suffix net clustering

semantic clustering