International Journal of Innovative
Computing, Information and Control
Volume **7**, Number **5(B)**, May **2011**

ICIC International ©2011 ISSN 1349-4198
pp. **2689–2698**

ENHANCING TEXT REPRESENTATION FOR CLASSIFICATION TASKS WITH SEMANTIC GRAPH STRUCTURES

Jiangning Wu, Zhaoguo Xuan and Donghua Pan

Institute of Systems Engineering
Dalian University of Technology
No. 2, Linggong Road, Ganjingzi District, Dalian 116023, P. R. China
{ jnwu; gyise }@dlut.edu.cn; xzg@dl.cn

Received January 2010; revised June 2010

Abstract. *To represent the textual knowledge more expressively, a kind of semantic-based graph structure is proposed, in which more semantic and ordering information among terms as well as the structural information of the text are incorporated. Such model can be constructed by extracting representative terms from texts and their mutually semantic relationships. Afterward, it is represented as a graph, whose nodes are the selected terms and whose edges are the corresponding relationships respectively. Moreover, the weight is assigned to each edge so that the strength of relationship between two terms can be measured. Furthermore, for this weighted directed graph structure, a novel graph similarity algorithm is developed by extracting the maximum common subgraph between two concerned graphs, which can therefore be used to measure the distance between two graph structures, i.e., two texts, and further be applied to classification tasks. Finally, some experiments have been conducted with the Chinese benchmark corpus for validation. The experimental results have proved the better performance of the proposed textual knowledge representation model in terms of its precision and recall.*
Keywords: Text representation, Graph structure, Maximum common subgraph, Classification

1. **Introduction.** Text representation is the essential step for the tasks of text mining, such as text clustering, text classification and so on. It nowadays has become one of the popular research topics in text mining since text is the most common form of information storage. One of the underlying problems with the textual representation is the expressivity of semantic information in the texts. Typical model like the vector space model (VSM) [1] is simple and only allows the application of traditional methods that deal with numerical feature vectors in a Euclidean feature space. However, the traditional paradigm in these kinds of models has discarded the important semantic and structural information when the original text is converted to a vector of numerical values. Considering that graphs are the strong mathematical constructs and can model relationships and structural information effectively, they are accordingly adopted in our study.

There are various forms of graphs, such as trees, networks and so forth, where the network-like structure can better reflect both contextual and semantic information of the text, and ad hoc syntactic information (e.g., phrase structure, word order, proximity information). The idea of graph representation for web content was originally presented in [2] and has been applied to such fields as symbolic images [13], document retrieval [14], etc. In [2], they viewed terms of HTML documents as nodes of the graph and relationships between terms as edges, thus a graph structure representation model for web content can be built. Motivated by the previous work [2-4, 13-16], the paper presents a novel weighted graph-structure model, in which more semantic and ordering information