Abstract

Due to the availability of required information in the web, as multiple documents, the need for summarizing these multiple documents and ordering of the sentences in the summary in an efficient way become a relevant task in data mining. We present a novel sentence ordering method based on maximum cost spanning tree algorithm to improve the readability and cohesion of the summary obtained by extraction method from related multiple documents. It is
based on extracting candidate sentences for the summary from multiple documents by ranking the sentences using cosine similarity measure and reducing the redundancy in the summary by Maximal Marginal Relevance (MMR) technique. Sentences in the summary are organized by constructing a graph where each sentence represents nodes of graph and edges are maintained between every pair of vertices which represents the similarity between the sentences. Most important task of our work is to find the first sentence to be placed in the ordered summary, by identifying the sentence which has minimum similarity with the sentences in the extracted summary. Ordering of remaining sentences in the summary is fixed one by one using Prim’s Maximum Cost Spanning tree algorithm. The proposed algorithm is tested with DUC 2002 data set and found that summary generated after ordering has better readability and cohesion than that generated without ordering. It is noted that results are more impressive as the summary size increases.

Reference


Index Terms

Computer Science
Algorithm

Key words

Cosine similarity
Maximal Marginal Relevance

Spanning Tree

Prim's algorithm