Abstract

This paper presents a method for web-document retrieval by learning importance factors for tags which are used for document structuring. Web documents are usually written in Hypertext Markup Language (HTML). HTML consists of tags which make a document into a specific form and a homepage is designed using the tags according to its object. In this paper, we propose a method for improving the retrieval performance using the information of HTML tags. The importance factors for the tags are learned using a genetic algorithm. A tag is mapped into a gene and a set of tags represented as a chromosome. The results obtained by genetic learning are the weights for tag importance, and provided the retrieval engine as the weights of documents.

Experiments have been performed on an artificial dataset and a large collection of TREC (Text REtrieval Conference) documents. Our empirical results show that this algorithm learns the weights by tag importance factors, and can improve the retrieval performance on top-ranked documents.

Keywords: Information Retrieval, Web-Document, Tag Weight, Genetic Algorithm