Semantic similarity based web document classification using Artificial Bee Colony (ABC) algorithm

C.KAVITHA Dr.G.SUDHA SADASIVAM S.KIRUTHIKA
Department of Computer Science and Engineering
PSG College of technology
Peelamedu, Coimbatore, Tamil Nadu
INDIA
mail2kavithak@yahoo.com, sudhasadhasivam@yahoo.com, kiruthika.2728@gmail.com

Abstract:- Due to the exponential growth of information on the Internet and the emergent need to organize them, automated categorization of documents into predefined labels has received an ever-increased attention in the recent years for efficient information retrieval. Relevancy of information retrieved can also be improved by considering semantic relatedness between words which is a basic research area in fields like natural language processing, intelligent retrieval, document clustering and classification and word sense disambiguation. The web search engine based semantic relationship from huge web corpus can improve classification of documents. This paper proposes an approach for web document classification that exploits information, including both page count and snippets and also proposes the use of Artificial Bee Colony (ABC) algorithm as a new tool in the classification task. To identify the semantic relations between the query words, a lexical pattern extraction algorithm is applied on snippets. A sequential pattern clustering algorithm is used to form clusters of different documents. The page count based measures are combined with the clustered documents to define the features extracted from the documents. These features are used to train the ABC algorithm, in order to classify the web documents.

Keywords:- Artificial Bee Colony (ABC) algorithm, Document Classification, Term Document Frequency, Latent Semantic Indexing (LSI), Web Search Engine

1 Introduction

Classification is a form of data analysis that can be used to extract models describing important data classes. Such analysis can provide a better understanding of the data at large. Document classification can be applied as an information filtering tool and can be used to improve the retrieval results from a query process and to make good decisions. The documents to be classified may be texts, images, music etc. Each kind of document possesses its special classification problems. Documents may be classified according to their subjects or according to other attributes like document type, author and printing year. Mining useful information from a relatively unstructured source, such Hyper Text Markup Language (HTML), World Wide Web, news articles, digital libraries, online forums and other types of documents can be difficult. So extracting information from these resources and proper categorization and knowledge discovery is an important area for research.

Semantic similarity between terms changes over time and across domains. For example, *apple* is frequently associated with computers on the Web. This sense of apple is not listed in most general-purpose thesauri. A user, who searches for apple on the Web, may be interested in this sense of apple and not apple as a fruit. New words are constantly being created as well as new senses are assigned to existing words. Manually maintaining thesauri to capture these new words and senses is costly if not impossible. Each source of information provides a different viewpoint; a combination has the potential of having better knowledge than any single method.

Conventional document classification methods are directly performed in the entire document space. These conventional algorithms based on exhaustive searches of the document space become computationally infeasible. The self adaptability of population based evolutionary algorithms can be used to tackle the task of document classification. Artificial Bee Colony algorithm is considered new and widely used in searching for optimum solutions. This is due to its uniqueness in problem-solving method where the solution for a problem emerges