Abstract

A domain specific crawler, as diverse from a general web search engine, focuses on a specific segment of web content. They are also called vertical or topical search engines. Common vertical search engines are meant for shopping, automotive industry, legal information, medical information, scholarly literature, and travel. Examples of vertical search engines are Trulia.com, Mocavo.com and Yelp. In contrast to general-purpose Web search engines, which attempt to index large portions of the World Wide Web using a web crawler, vertical search engines typically use a domain specific crawler that attempts to index only Web pages that are relevant to a pre-defined topic or set of topics. Vertical search offers several potential benefits over general search such as greater precision due to their limited scope, leverage domain knowledge including taxonomies and ontology and support of specific unique user tasks. This paper aims at analyzing the machine learning Techniques namely ANN, SVM and Hi-SVM being used for Web Page Classification and suggesting suitable improvements. Here a crawling framework has been designed and developed that allows flexible addition of new classifiers. This crawler has been used for classification of web content for few domains. The crawlers themselves are implemented as multithreaded objects that run concurrently. The results show that Hi-SVM is a better choice for guiding a topical crawler when compared to Support Vector Machine and Neural Network. The comparative analysis of the three classifier
techniques namely ANN, SVM and Hi-SVM showed that the performance of Hi-SVM is most efficient.

References

- Data Mining Algorithms In R-Classification-penalizedSVM - Wikibooks, open books for an open world. htm.
- Artificial Neural Networks Neural Network Basics - Wikibooks, open books for an open world. htm.

Index Terms

Computer Science

Web Services

Keywords

ANN SVM HiSVM VSM ROC REC POS WSD SOE.