Amirkabir University of Technology
Computer Engineering and Information Technology Department

Thesis Project Submitted in Partial
Fulfillment of The Requirements For The
Degree of Bachelor of Science in
Information Technology

The Design & Implementation of a New Webpage Classification Method Using WordNet

Advisor

Prof. Shahram Khadivi

by

Sayyedhassan Shavarani

Spring 2014

Surname: Shavarani Name: Sayyedhassan

Title: The Design & Implementation of a New Webpage Classification Method Using WordNet

Advisor: Prof. Shahram Khadivi

Degree: Bachelor of Science Field: Information Technology

Amirkabir University of Technology Computer Engineering & IT Department

Date: Spring 2014 Number of pages: 46

Keywords: Unsupervised Classifier, Semantic Similarity, WordNet, Single Document Analysis

Abstract

This Thesis Project is aimed to represent a new method for webpage classification using Wordnet and Lin's semantic similarity algorithm. Webpage classification and automatic determination of the label for each webpage is an important topic in design and implementation of search engines and web analyzers, specially in Farsi. Such applications and all other which need learning data for their classification and don't own much of it, will face performance problems as well as high costs of preparing a useful labeled dataset. This project is aimed to reduce such costs and automatically produce semi-labeled data for the aforementioned applications. The main fact used in design and implementation of this project, is the ability of the human mind to detect main keywords of the text and analyze its semantic structure while reading it for the first time. Meanwhile, with keeping in mind the need to auto-label the unlabeled data with no pre-learning phase, the method implemented in the project will be an unsupervised one. Furthermore, the data used for evaluation, has been taken for the Persica Corpus which contains about eleven thousand news parts from the Iranian Students' News Agency (ISNA). In our method each document is semantically analyzed with each of the pre-determined labels and the label with the highest semantic score is chosen as the webpage label. And, As implementation and optimization of parameters of the method is done, the accuracy ratio of 0.86 is reached which is the ratio of the accuracy of our method to the claimed accuracy of Persica, reported by its gatherers.