Stemming algorithms

Abstract:

The amount of data on the web is still increasing, so there is a need to find some interesting knowledge on the web by means of web mining techniques. In [1], two-phase classification of web documents is performed to assign some of predefined class labels to web documents. [2] is focused on measuring web page similarity based on textual and visual properties. These topics are related to my PhD theses.

Stemming is very common requirement of Natural Language processing functions as well as a pre-processing step in Text Mining applications. The main purpose of stemming is to reduce different word forms like its noun, adjective, verb, adverb etc. to its root form. Stemming is used in two-phase classification and measuring similarity mentioned above.

In my presentation, I will discuss different methods of stemming, introduce classification of stemming algorithms and describe the most significant ones. I will explain Porters stemming algorithm (one of the most popular stemming algorithms proposed in 1980) in detail and I will also mention some interesting methods of stemming, especially Corpus Based Stemmer and Context Sensitive Stemmer. Finally, there will be a comparison between algorithms.

[1] Bartík V., Burget R.: Two-phase categorization of web documents. Faculty of Information Technology, Brno University of Technology.

[2] Bartík V.: Measuring Web Page Similarity Based on Textual and Visual Properties. Faculty of Information Technology, Brno University of Technology.

Faculty of Information Technology, Brno University of Technology
Modern Theoretical Computer Science
Petr Loukota
2014 / 2015