Automated Text Classification in the DMOZ Hierarchy - Project Plan

Lachlan Henderson

August 18, 2009

1 Overview

1.1 Objective

The goal of this project is to build a text classifier[2, 4, 7] for the DMOZ hierarchy of classification labels. Based on previous successes[3, 5, 6], this project will focus on non-parametric methods such as nearest-neighbour algorithms exploring different feature representations. Various approaches will be implemented and evaluated in the initial stage, focusing on the use of the classification hierarchy to improve performance and development of language-independent classification techniques. Of the most successful approach(es), work in the latter stage will work on making the classifier as efficient as possible.

1.2 Motivation

The growth in the availability of on-line digital text documents has spurred considerable interest in Information Retrieval and Text Classification. The Internet particularly represents a considerable opportunity for many corporations and individuals to exchange ideas, access products and services. Automation of the management of this wealth of Internet hypertext is becoming an increasingly important endeavor as the rate of new material continues to grow at its substantial rate.

The DMOZ open directory project[1] is an on-line service which provides a searchable and browsable hierarchically organised directory to facilitate access to the Internets resources. DMOZ is a collaborative effort of over 56,000 volunteers who contribute and renew Internet content for a growing list of over 718 thousand categories. This represents a considerable convenience for users of the Internet and also a valuable resource for Data Mining applications.

2 Project Plan

2.1 Project Variables

1. Training/Test data selection: DMOZ description, Actual page data.

2. Sampling of DMOZ: Tree depth limit, Random sampling over all categories, Summarise categories as total accumulated category and subcategories, Summarise categories as set of accumulated subcategories.

3. Generation of features: n-gram bag of words, n-gram character vector ..

4. Feature Representation: TF, TF-IDF, Binary.

5. Algorithms: Centroid, KNN, SVM, Naïve Bayes, Decision Trees.

2.2 Project Artifacts

1. Train and test data organised in Weka sparse format

2. Command line interfaces to run train and test evaluation with options for for nearest neighbour, other classifiers (Weka).

3. Basic documentation for usage of corpus and classifiers.

2.3 Implementation Summary

In meeting the objectives the following list of tasks will be undertaken.

1. Implement a Nearest Neighbour classifier utilizing successive refining methodology. Trialing a number of document sampling, representation, and feature selection techniques.

2. Utilise an open source machine learning system such as Weka[8] to compare classifier performance over the same train/test data against a collection of standard classification algorithms.

3. Evaluating the results obtained.

4. Preparing report and final presentation.

2.4 Schedule

Week	Date	Activity	Milestone
5	16/08/2009 - 22/08/09	Write/utilise software to parse and generate files for use by Weka classifiers, collation of data files.	Initial raw document feature sets.
6	23/08/2009 - 29/08/09	Implementation of Nearest Neighbour classifier.	
7	30/08/2009 - 05/09/09	""	Document corpus. Operational Nearest Neighbour classifier
8	06/09/2009 - 12/09/09	Implement Interface to Weka and initial testing of selected algorithms.	
9	13/09/09 - 19/09/09	""	Set of algorithms and results for comparison.
10	20/09/09 - 26/09/09	Analysis of results.	
	27/09/09 - 03/10/09	Draft report and Final Presentation	
	04/10/09 - 10/10/09	""	Draft Report and Final Presentation complete.
11	11/10/09 - 17/10/09	Review of Report and Final Presentation	
12	18/10/09 - 24/10/09	""	
13	25/10/09 - 31/10/09		Report Submission
14	01/11/09 - 07/11/09		Final Presentation

References

[1] The open directory project (http://www.dmoz.org/).

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, October 2007.

[3] M. Ceci and D. Malerba. Classifying web documents in a hierarchy of categories: a comprehensive study. *Journal of Intelligent Information Systems*, 28(1):37–78, 2007.

[4] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. 1973.

[5] Marko Grobelnik and Dunja Mladenic. Simple classification into large topic ontology of web documents. *CIT*, 13(4):279–285, 2005.

[6] Eui-Hong Sam Han and George Karypis. Centroid-based document classification: Analysis experimental results. pages 424–431, 2000.

[7] Fabrizio Sebastiani and Consiglio Nazionale Delle Ricerche. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002.

[8] I.H. Witten and E. Frank. Data mining: Practical machine learning tools and techniques, 2005.