

Project plan: Automated country-classification of web content using language features

Alexander Osborne
Supervisor: Paul Thomas

August 15, 2008

Background

National libraries are tasked with the duty of preserving their national cultural heritage. Several national libraries routinely attempt to capture as much as possible of their nation's footprint on the world wide web, by recursively following links, a process known as web crawling. Traditionally these crawls have been scoped by filtering for websites in their country domain or those that can be determined to be hosted on computers in their country through IP geolocation [1]. In the past this has proved to be quite an effective method, however in recent years third party hosting and user-contributed content have become much more popular. Due to low barriers of cost and speed, content written by an author in one country may often be hosted by a computer on the other side of the planet and therefore could be missed by a crawl scoped by hosting location.

Aim

This project will explore alternative methods for classifying web content in an automated fashion as authored by or about a given country. In particular it will focus on determining through analysis of language and to a limited extent page structure, whether a given web page is authored by an Australian or is primarily about Australia or Australians.

One potential method is to use a machine learning classifier, such as a naive Bayes classifier or support vector machine on the frequency of significant features, training against a set of known Australian pages and known non-Australian. Some significant features could include:

- locations (Sydney, Tasmania, Uluru, ...)
- well-known Australian people, fauna, flora and objects (Kevin Rudd, Kylie Minogue, The Chaser, Vegemite, kangaroo, gum tree, ...)
- dialect words, spelling and slang (g'day, colour, drongo, ...)
- meta-data (locale, time-zones, geotagging, CC Australia licensing, ...)

Performance will be evaluated on a set of known Australian data from the National Library's collection and a set of (mostly) non-Australian data from a general crawl perhaps manually pruned to remove any obviously Australian content. The primary goal will be to minimize the number of false negatives, with minimizing false positives as a secondary priority. This is because for preservation purposes it is far better to collect too much than too little. Emphasis will also be placed on classifiers that are computationally efficient with the intent that the classifier could be embedded within a web crawler.

Schedule

Week	Task
Aug 18	Obtain and preprocess data
Aug 25	Obtain and preprocess data
Sep 1	Trim data
Sep 8	Experiments
Sep 15	Experiments
Sep 22	Experiments
Sep 29	Experiments
Oct 6	Thesis
Oct 13	Thesis
Oct 20	Thesis
Oct 27	Thesis, presentation

Obtain and preprocess data

- Obtain and sample web data
- Strip tags and common words
- Break into terms
- Calculate term frequency statistics

Trim data

- Create reduced term sets using:
 - Robertson Selection Values [3, 4]
 - Inverse chi-square classification [2]
 - Manual selection (lists from Wikipedia etc)

Experiments

- Train a selection of classifiers on the data, likely using Weka [5]
- Evaluate their performance

Potential pitfalls

Hardware failure

Experiments will be performed (with permission) using the National Library's "petabox" cluster. The cluster provides redundant storage for data and additionally code and final results will be backed up to another PC.

Poor performance of tools

Performance may become an issue as the data set may potentially be very large. If this occurs the sample size can be reduced or alternative tools may be used.

Bibliography

- [1] Paul Koerbin. Web archiving at the national library of australia. *CDNLAO Newsletter*, March 2007.
- [2] Lung-Hao Lee and Cheng-Jye Luh. Generation of pornographic blacklist and its incremental update using an inverse chi-square based method. *Information Processing & Management*, 44(5):1698–1706, September 2008.
- [3] S. E. Robertson. On term selection for query expansion. *J. Doc.*, 46(4):359–364, 1990.
- [4] S E Robertson, S. Walker, S. Jones, M M Hancock-beaulieu, and M. Gatford. Okapi at trec-3. *Proceedings of TREC-3*, pages 109—126, 1995.
- [5] Ian H Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham. The waikato environment for knowledge analysis.