The LangGator System: Project Overview

Baden Hughes
The University of Melbourne
badenh@csse.unimelb.edu.au

Presentation Overview

- Introduction and Motivation
- Workflow
- Seed Data
- Core Processes
  - Query Permutation
  - Metasearch
  - Rank Aggregation
  - Content Acquisition
  - Data Storage
- Downstream Consumers

Introduction and Motivation

THE UNIVERSITY OF MELBOURNE

- Finding low density data on the web is at best a low precision activity
- Linguistic data is a low density data set in web terms
- Interesting linguistic data is from languages which tend to be minority languages
- Low density languages (small number of speakers, small number of web resources) hence suffer twice from the low precision problem
- Can we develop an automated approach to systematically leverage existing web search infrastructure, combined with intelligent treatment of linguistic data, that can reduce the cost of resource discovery and allow scaling up ?

LangGator

- A domain specific web crawler
- Features high precision identification and acquisition of linguistically interesting web content
- Next generation on from task specific linguistic content crawlers; general purpose

Software

- Under development since early 2004
- Written in Java (J2SDK 1.4.x, migrating to 1.5.x)
- Cross-platform components throughout
- Uses some components of Heritrix and Apache Lucene
- Open source (not yet GPL'd, but intended to be)

Basic Workflow

THE UNIVERSITY OF MELBOURNE

- Take intial query term
- Expand query systematically
- Execute query on range of search engines
- Aggregate results
- Acquire content from result URIs
- Create metadata for result URI content and bundle
- Store and set refresh interval

Seed Data

- Linguistic Seed Data
  - Language names
    - SIL Ethnologue
  - Lexical items
    - Rosetta word lists
    - Open source word lists (eg aspell/ispell)
  - Linguistic terms
    - OLAC linguistic fields
    - Ad-hoc linguistic descriptors
- Geographic Seed Data
  - Getty Thesaurus of Geographic Names (TGN)
  - GMI language/locational data (WLMS)
  - USGS / NGIS feature data

Query Permutation and Expansion

THE UNIVERSITY OF MELBOURNE

- For each primary language name in the Ethnologue
  - Linguistic Expansion
    - Generate query for each alternative using core name, alternative names plus descriptive terms from each of the linguistic seed data sets
  - Geographic Expansion
    - Generate query for each new term using core language name, country of origin and alternative names, plus geographic terms
- 1 language expanded to average 1800 separate queries

Metasearch

- Submit initial query set to selection of commercial search engines
  - Google, Yahoo, MSN
  - Programmatic access via Web Service interface and SOAP
  - Exemptions to usual API query quotas courtesy of Google, Yahoo, MSN
- Retrieve top 100 results from each engine for each query

Rank Aggregation

THE UNIVERSITY OF MELBOURNE

- Maximum 300 unique results for each query submitted
- Different search engines use different ranking algorithms
- Combine result sets per query using occurrence-based and rank-based similarity metrics
- Re-rank all URIs for a query for crawl priority

Content Acquisition

THE UNIVERSITY OF MELBOURNE

- Top N ranked URIs retrieved (user specified, default is all)
- Inbound links retrieved to depth 1
- Outbound links retrieved to depth 1
- URI content checksummed for change monitoring once per month
- URI metadata created
  - URI, URI checksum, URI content, URI content checksum, date identified, date retrieved, query provenance, ranking

Data Storage

- URI metadata and URI content bundled together
- Format compatible with Heritrix (Internet Archive crawler) – uses ARC format for storage
- Currently stored both locally on disk and remotely on Australian Partnership for Advanced Computing Mass Data Store (online and nearline storage)
- Refresh interval catalogue

Performance

- Current rate of execution is to exhaustively find, rank and acquire content based on 7K Ethnologue languages once every 3 months
  - Target 7K Ethnologue list once per month (better parallelisation)
  - Target 7K Ethnologue list once per week (more bandwidth)
- Extremely high precision (90%+) selection of “linguistically interesting content”
  - 1+ orders of magnitude better than Google, Yahoo, MSN based on result set comparison

Status March 2006

THE UNIVERSITY OF MELBOURNE

- Focus on languages ranked below 2000 in terms of number of speakers according to Ethnologue (ie the long tail)
- Crawler 1.0 acquiring content, 1.6 million URIs of “linguistic interest” identified and retrieved
- Total data on disk 1.8Tb
- Crawl running on 32 node cluster at University of Melbourne (CPU rather than memory intensive)
- Crawler 2.0 (distributed collaborative crawling) in development

Downstream Consumers

THE UNIVERSITY OF MELBOURNE

- A number of linguistically-oriented web applications consume the URI sets or data from LangGator
  - ODIN: online database of interlinear text
  - An Crudaban: thesaurus/dictionary/corpora building project
  - Linguist's Search Engine: URI content as collections for linguistic query
  - Rosetta Project: basic linguistic information about 3000 languages
- The Internet Archive also uses URI sets for selective archiving of web content
- National Library of Australia uses URI sets for selective archiving of Australian indigenous language content

Future Work

THE UNIVERSITY OF MELBOURNE

- Better parallelisation
- Enable collaborative crawling
- Utilisation of incoming and outgoing links on found pages
- UTF-16 support (within bounds allowed by search engines)
- Web Service interfaces
- Documentation and GPL release

Conclusion

- LangGator is a state of the art content acquisition engine for specifically linguistic content on the web
- Uses best of breed architectures and components from other crawler software
- Novel but proven method for expanding queries and rank aggregation
- Available for interested researchers to share both as software and as a service