Evaluating Strategies for Finding Genealogical Information on the Web

Dallan Quass
Nathan Powell
Solveig Quass

Foundation for On-Line Genealogy

Introduction

- Lots of genealogical information on the Web
  - How much?
- Different strategies for finding it
    - General-purpose search engines
    - Genealogy-specific directories
  - How effective are they?
- Is there a better way?

General-purpose search engines

- Two approaches:
  - Enter ancestor's name
    - Often poor precision
  - Append "genealogical" words
    - Significant drop in recall
- Another issue: variant/ misspelled names

Query	Yahoo	Google
Smith Site:rootsweb.com	893K pages	4,200K pages*
Smith genealogy Site:rootsweb.com	228K pages	543K pages
% of "Smith" query	26%	13%
Smith "family history" -genealogy Site:rootsweb.com	16K pages	50K pages
Smith vital -"family history" -genealogy Site:rootsweb.com	N/A	93K pages

Genealogy-specific directories

- 100% precision
- Most “important” websites
- Recall is questionable
  - WeRelate.org has found genealogical content on over 58K hosts in a limited Web crawl

	Cyndis list.com	Link pendium .com	DMOZ.org
Number of links advertised as of 8 Feb 2006	251K	3,257K	5.2M but not all genealogy
Number of unique off-site URLs	115K	2,100K (est.)	19K genealogy
Number of unique hosts	25K	12K (est.)	5K

Web page classification

- Performance: 90% recall, 90% precision
  - five-fold cross-validation of the training data
- Repository of 3.5M web pages
  - 73% precision (400 page random sample)
    - due to poor repository admittance strategy
  - 58K unique hosts represented
- Precision of pages containing “genealogy”
  - 85% precision (400 page random sample)
- Another benefit:
  - Expand queries with variant & misspelled names

Estimating the total size of genealogical information on the Web

- Appear to be 25-100M pages containing “genealogy”
  - Google numbers fluctuate wildly
- 85% precision of “genealogy” combined with 26% recall yields 80-325M total pages
- Next: “Deep Web”

Date (2006)	Google genealogy	Yahoo genealogy
13 Feb	28.3M	94.4M
15 Feb	27.8M	95.2M
16 Feb	45.7M	95.8M
17 Feb	49.1M	95.3M
20 Feb	55.9M	96.5M
21 Feb	49.5M	96.6M
23 Feb	121M	96.8M
27 Feb	26.1M	109M
28 Feb	26M	110M

Deep Web

- Difficult to estimate
- Found 2 forms in a sample of 400 genealogy web pages
  - Would project over 400K+ forms, but that's high
- Two outliers:
  - Ancestry: 4B names
  - LDS Church: 1B names
- Deep Web dwarfed by off-line content
  - LDS Church microfilms: 3B images (est. 30% are registers)
  - Would yield 10's of billions of names

Conclusion

- **Lots** of genealogical information on the Web
- Difficult to find presently
- Web page classification coupled with targeted crawling shows promise
- On-line content dwarfed by off-line content