Domain Classification

Jay Daley

Classification in a nutshell

- Start with a standard industry classification
    - e.g. SIC (US), NACE (Europe), ISIC (International) or ANZSIC (Australia/NZ)
    - All except SIC are very similar
- Classify domains by their website contents
- Two main methods
    - Manual – Humans visit a site and classify. Can achieve 500-1000 per day
    - Machine Learning – Crawler grabs text from site and uses trained neural net to classify
- Output normally a single primary classification
    - Sometimes includes multiple secondary classifications

Extract from ISIC

- "A","Agriculture, forestry and fishing”
  - "01","Crop and animal production, hunting and related service activities”
    - "011","Growing of non-perennial crops”
      - "0111","Growing of cereals (except rice), leguminous crops and oil seeds”
      - "0112","Growing of rice”
      - "0113","Growing of vegetables and melons, roots and tubers”
      - …
    - "012","Growing of perennial crops”
    - …
  - …
- "B","Mining and quarrying”
  - "05","Mining of coal and lignite”
    - "051","Mining of hard coal"

Benefits

- National statistics bodies use classification to size national industry
  - Number of companies / organisations
  - Number of employee
  - Turnover
- Economic value of domain name industry
  - Market penetration – by companies and turnover
  - Value of industry served by domain names
- Registrar level
  - Specialising in verticals
  - Directing sales / advertising

Three world leaders (maybe more)

- .nz
    - Attempting to classify every domain in register – mix of manual / machine
- CENTR
    - Working group of EUs largest registries and registrars
    - Created new classification standard specific for domain names
- Dataprovider
    - Commercial service that classifies domains among many other data points

.nz

- 700,000 domains in registry
- Large Hadoop cluster with customised distributed web crawler
- Manually classified 100,000+
- Multiple machine learning models tested and used to classify rest
- Accuracy varies by ANZSIC code
- Commercial product being rolled out
  - Combines classification with traffic measurement
  - Registrant can compare their traffic against others sites in same industry
  - Understand if investment delivers relative gains in traffic

CENTR

- The RRDG (registry-registrar data group)
    - Group of largest European ccTLDs and registrars
    - Work on producing frameworks, classifications and tools
    - Help the industry better understand the market of domain names
    - Informal with support and participation of CENTR
- RRDG output - Domain industry taxonomy (DIT)
    - This is a classification of industries and sub-industries matched up to European NACE codes.
    - More information at https://stats.centr.org/classifications#dit
    - Relative market penetrations in a country and cross country comparisons

Dataprovider

- Large scale global data crawling
  - Index and structure publicly available data from the web, 30 to 50 pages
  - Collect 150+ data attributes including industry classification and trust score
  - Re-indexed on a monthly basis to providing for insights into historical data
  - Data is across 50 countries to date, 29 languages analyzed.
  - Country data is defined from the content of websites: address, phone, TLD (if ccTLD), language etc. - WHOIS country data, less reliable
- Clients
  - D&B, PayPal, Symantec, GoDaddy, SIDN, CreditSafe
  - Brand/IP community and local enforcement authorities
- Use cases by companies to date
  - Insights into e-commerce companies
  - Insights into the digital footprint of websites
  - Marketing intelligence
  - Profiling registrants – classification and common ownership.

Questions

jay@daley.org.nz