Rules of Thumb for Information Acquisition from Large and Redundant Data

Wolfgang Gatterbauer

33rd European Conference on Information Retrieval (ECIR'11)

Database group
University of Washington

http://UniqueRecall.com
Information Acquisition from Redundant Data

Pareto principle (80-20 rule)  
20% causes  
80% effect  
e.g. business  
clients  
sales  
e.g. software  
bugs  
errors  
e.g. health care  
patients  
HC resources

Information acquisition  
? 20% data  
? 80% information  
(e.g. web harvesting)  
web data  
web information  
e.g. words in a corpus  
all words  
different words  
e.g. used first names  
individual names  
different names

Motivating question:  
Can we learn 80% of the information, by looking at only 20% of the data?
Information Acquisition from Redundant Data

"Unique" information $A_u$ 

Available Data $A$ 

Retrieved and extracted data $B$ 

Acquired information $B_u$ 

Redundancy distribution $k$ 

Recall $r$ 

Expected sample distribution $\hat{k}$ 

Expected unique recall $\hat{r}_u$

Three assumptions
- no disambiguity in data
- random sampling w/o replacement
- very large data sets
Outline

• A horizontal sampling model
• The role of power-laws
• Real data & Discussion
A Simple Balls-and-Urn Sampling Model

Redundancy $k_i$

Data

(# balls: $a=15$)

Information $i$

(# colors: $a_u=5$)

Sampled Data

(# balls: $b=3$)

Sampled Information $i$

(# Colors: $b_u=2$)

Redundancy Distribution $k = (6, 3, 3, 2, 1)$

Recall $r = 3/15 = 0.2$

Unique recall $r_u = 2/5 = 0.4$

Sample Redundancy distribution $k = (2, 1)$
A model for sampling in the Limit of large data sets

\[
\mathbb{E}[r_u] = 1 - \frac{1}{a_u \left( \frac{a}{b} \right)} \sum_{i=1}^{a_u} \left( a - k_i \right)
\]

\[
\mathbb{E}\left[ r_u \right] = 1 - \sum_{k} \alpha_k (1 - r)^k
\]

\( k = (6, 3, 3, 2, 1) \quad k_{2-3} = 3 \)

\( \alpha_3 = 0.4 \)

vertical perspective

horizontal perspective
The Intuition for constant redundancy $k$

$$\lim_{a \to \infty} r = 0.5$$

$k = \text{const} = 3$

$$\Delta_2 = (\text{3 choose 2})0.5^2(1-0.5)^1 = 0.375$$

$$r_u = 1 - (1-0.5)^3 = 0.875$$

Unique recall

$$\mathbb{E}[r_u] = 1 - (1-r)^k$$

$\approx$ Indep. sampling with $p=r$

Expected sample distribution

$$\Delta_j = \binom{k}{j} r^j (1-r)^{k-j}$$

$\approx$ Binomial distribution
The Intuition for arbitrary redundancy distributions

Stratified sampling

\[ E[r_u] = 1 - \sum_k \alpha_k (1 - r)^k \]

\[ \omega_j = 1 - \sum_{y=0}^{j-1} \sum_{x=y}^{k_{\text{max}}} \alpha_x \binom{x}{y} r^y (1 - r)^{x-y} \]

\[ r_u = \alpha_6[1-(1-r)^6] + \alpha_3[1-(1-r)^3] + \alpha_3[1-(1-r)^2] + \alpha_1[1-(1-r)^1] \]
A horizontal Perspective for Sampling

Expected sample redundancy $\hat{k}_1 = 3$

Horizontal layer of redundancy $\omega_1 = 0.8$

$\hat{b}_u = 800$

$r = 0.5$

Fraction of Information $\eta$

Redundancy $k$

$a_u = 5$

$a_u = 20$

$a_u = 100$

$a_u = 1000$
Unique Recall for arbitrary redundancy distributions

![Graph showing unique recall rates for different redundancy distributions]

- Uniform
- Linear
- Logarithmic
Outline

• A horizontal sampling model
• The role of power-laws
• Real data & Discussion
Three formulations of Power law distributions

- Zipf-Mandelbrot
  \[ k_i \propto i^{-\delta} \]
  Zipf [1932]

- Power-law probability
  \[ \alpha_k \propto k^{-\beta} \]
  Stumpf et al. [PNAS’05]

- Pareto
  \[ \eta_k \propto k^{-\gamma} \]
  Mitzenmacher [IM’04]

Commonly assumed to be different representations of the same distribution

Adamic [TR’00]
Mitzenmacher [Internet Math.’04]
Unique Recall with Power laws

For $\gamma=1$

$$r_u = \frac{r}{\sqrt{1-r}} \text{artanh}(\sqrt{1-r})$$

$$r_u = 1 - \frac{\text{Li}_\beta(1-r)}{\zeta(\beta)}$$

$$r_u = -\frac{r \ln r}{1-r}$$

log-log plot

Redundancy $k$

Fraction of information $\eta_k$

Tail

Core

Root

Unique recall $r_u$

Recall $r$
**Rule of Thumb 1:** When sampling 20% of data from a Power-law distribution, we expect to learn less than 40% of the information.
Invariants under Sampling

Given our model: Which redundancy distribution remains invariant under sampling?

\[ \alpha_k = \frac{\tau(1-\tau) \cdots (k-1-\tau)}{k!} = \binom{\tau}{k}, \text{ with } 0 < \tau \leq 1 \]
Invariant is a power-law! Hence, the tail of all power laws remains invariant under sampling.
Also, the power law tail breaks in

**Rule of Thumb 2:** When sampling data from a Power-law then the core of the sample distribution follows \( r_{uk} \sim r^\gamma \)
Outline

• A horizontal sampling model
• The role of power-laws
  • Real data & Discussion
Sampling from Real Data

**Rule of Thumb 1:**
- not good!

**Rule of Thumb 2:**
- works, but only for small area

**Tag distribution on delicious.com**

**Rule of Thumb 1:**
- perfect!

**Rule of Thumb 2:**
- works!
Theory on Sampling from Power-Laws

Stumpf et al. [PNAS’05]
"Here, we show that random subnets sampled from scale-free networks are not themselves scale-free."

This paper
"Here, we show that there is one power-law family that is invariant under sampling, and the core of other power-law function remains invariant under sampling too.

\[
\begin{align*}
\text{Invariant } \tau = 0.5, r = 0.5, m = 500 \\
\rho \propto \eta^{-1} \text{ (Zipf distribution), } r = 0.5, m = 500
\end{align*}
\]
Some other related work

Population sampling

- goal: estimate size of population
- e.g. mark-recapture animal sampling
- sampling of small fraction w/ replacement

- urn model to estimate the probability that extracted information is correct.
- random sampling with replacement

Downey et al. [IJCAI’05]

Ipeirotis et al. [Sigmod’06]

- decision framework to search or to crawl
- random sampling without replacement with known population sizes

Bar-Yossef, Gurevich [WWW’06]

- biased methods to sample from search engine's index to estimate index size

Stumpf et al. [PNAS’05]

- show that random subnets sampled from scale-free networks are not scale-free
Summary 1/2

- A simple **model of information acquisition** from redundant data
  - no disambiguity
  - random sampling w/o replacement

- **Full analytic solution**
  - large data

---

Wolfgang Gatterbauer  
**Rules of Thumb for Information Acquisition from Redundant Data**  
http://uniqueRecall.com
Summary 2/2

- Rule of thumb 1:
  - **80/20 → 40/20**
  - sensitive to exact power-law root

- Rule of thumb 2:
  - power-law core remains invariant

\[ r_{uk} \approx r^\gamma \]

\[ \alpha_k = \left( \frac{\tau}{k} \right), \text{ with } 0 < \tau \leq 1 \]

http://uniqueRecall.com
backup
Geometric interpretation of $\Delta_k(\rho, k, r)$

$$\Delta_k(\rho, k, r) = \binom{\rho}{k} r^k (1 - r)^{\rho-k}$$
Information Acquisition from Redundant Data

3 pieces of data, containing 2 pieces of ("unique") information*

Data (instances) / Information (concepts)

e.g. used first names in a group
individual names / different names

e.g. words of a corpus:
word appearances / vocabulary

e.g. web harvesting:
web data / web information

Motivating question:
Can we learn 80% of the information, by looking at only 20% of the data?

*data interpreted as redundant representation of information

Capurro, Hjørland [ARIST'03]

Wolfgang Gatterbauer  Rules of Thumb for Information Acquisition from Redundant Data  http://uniqueRecall.com