TECHNISCHE UNIVERSITÄT WIEN
VIENNA UNIVERSITY OF TECHNOLOGY

Estimating Required Recall for Successful Knowledge Acquisition from the Web

Wolfgang Gatterbauer
www.dbai.tuwien.ac.at/staff/gatter

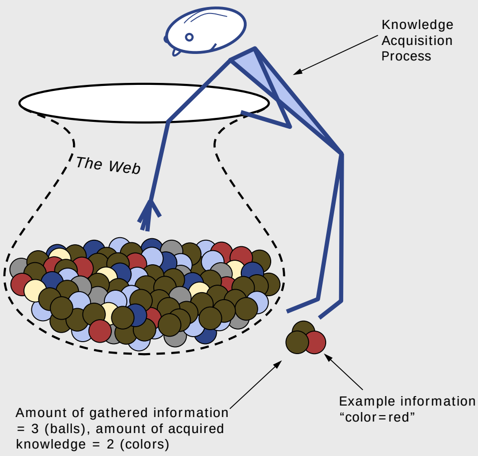
DATABASE AND ARTIFICIAL INTELLIGENCE GROUP

Motivation

Knowledge Acquisition from the Web

- The Web contains information, proportioned into statements
- Each statement is chosen to contain one single message
- We want to acquire knowledge by gathering and analyzing many different statements

How many statements do we have to process in order to learn a certain fraction of the available knowledge?



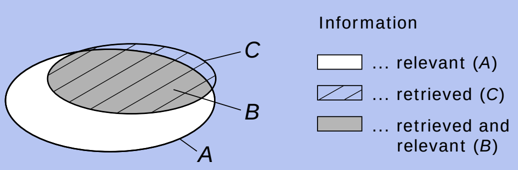
An Urn Model

- An Urn is filled with ***a*** balls
- Each ball has one out of $\boldsymbol{a_u}$ colors
- We want to learn about the diversity of colors

How many balls ***b*** do we have to draw in order to learn a certain number $\boldsymbol{b_u}$ or a certain fraction $\boldsymbol{r_u} = \boldsymbol{b_u} / \boldsymbol{a_u}$ of different colors?

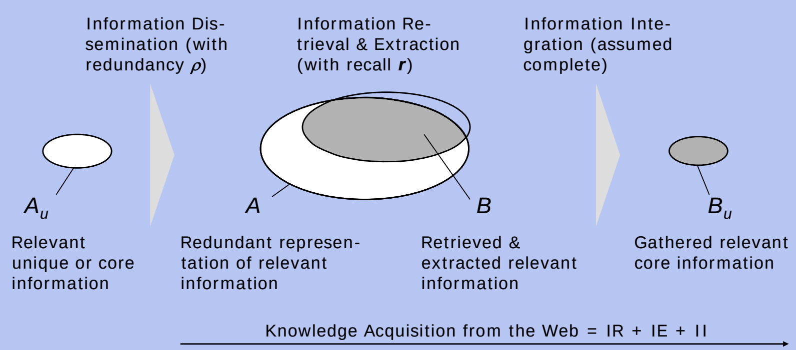
Model of Knowledge Acquisition from the Web

Information Retrieval (IR) Information Extraction (IE)



Focus of IR and IE: Recall $\boldsymbol{r} = |\boldsymbol{B}| \,/\, |\boldsymbol{A}|$

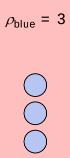
Process of Knowledge Dissemination and Acquisition by the Web



Focus of Knowledge Acquisition: Unique recall $\boldsymbol{r_u} = |\boldsymbol{B_u}| \,/\, |\boldsymbol{A_u}|$

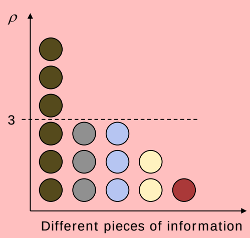
Useful Notions

Redundancy (ρ)



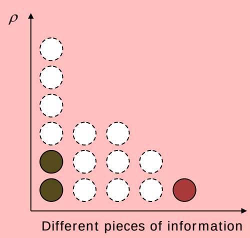
- 3 balls are blue
- Hence, redundancy ρ of information "color=blue" is 3

Redundancy Distribution



- Different colors have different numbers of occurrences
- Overall (or average) redundancy is 3

Unique Recall ($\boldsymbol{r_u}$)



- After drawing $\boldsymbol{b}$=3 balls, we saw $\boldsymbol{b_u}$=2 different colors
- Given $\boldsymbol{a_u}$=5 available colors, we learned a fraction of 2/5
- Hence, our $\boldsymbol{r_u}$ is 2/5

TECHNISCHE UNIVERSITÄT WIEN
VIENNA UNIVERSITY OF TECHNOLOGY

Estimating Required Recall for Successful Knowledge Acquisition from the Web

Wolfgang Gatterbauer
www.dbai.tuwien.ac.at/staff/gatter

DATABASE AND ARTIFICIAL INTELLIGENCE GROUP

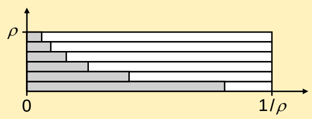
Solution for Homogeneous Redundancy Distribution

Normalized Homogeneous Redundancy Distribution

a = whole area

b = shaded area

r = ***b*** / ***a***

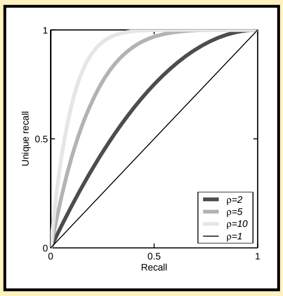


$\boldsymbol{r_u}$ = shaded fraction of the lowest layer of redundancy

Fundamental Unique Recall Formula

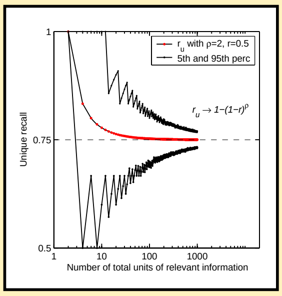
$$\mathbf{E}(r_u) = 1 - (1 - r)^{\rho}$$

Homogeneous Unique Recall Graphs



Relation between recall and unique recall for different redundancies

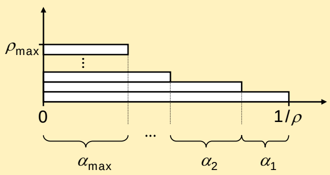
Unique Recall as Random Variable



Unique recall formula as asymptotic limit value of actual unique recall

Solution for General Redundancy Distribution

Normalized General Redundancy Distribution

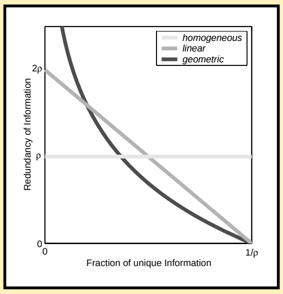


$$\sum_{i=1}^{\rho_{\max}} \alpha_i = 1 \qquad \sum_{i=1}^{\rho_{\max}} i\alpha_i = \rho$$

Generalized Unique Recall Formula

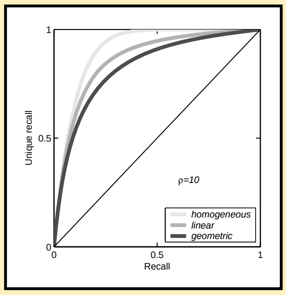
$$\mathbf{E}(r_u) = 1 - \sum_{i=1}^{\rho_{\max}} \alpha_i (1 - r)^i$$

Example Redundancy Distributions



3 example redundancy distributions in the continuous regime

Corresponding Unique Recall Graphs



Resulting characteristic unique recall graphs

Open Issue: Solution and Interpretation for Generalized Zipf

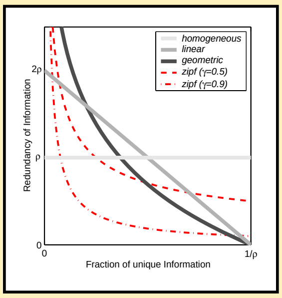
Zipf as Commonly Found Redundancy Distribution

Many findings suggest that redundancy of actual information follows a generalized Zipf function independent of the domain.

$$\rho_k = \rho_1 k^{-\gamma}$$

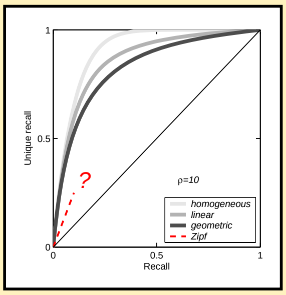
What are the implications for unique recall formula, and as such for knowledge acquisition from the Web in general?

Example Zipf Redundancy Distributions



Two example Zipf redundancy distributions

Yet Unknown Zipf Unique Recall Graphs



Unique recall with Zipf

Next Steps

- Solve generalized Zipf
- Understand implications of transitions from discrete to continuous regime
- Analyze errors due to generalizations of redundancy
- Simulate with real data to verify predictive power

Thanks

This research has been supported in part by the Austrian Academy of Sciences through a DOC scholarship, and by the Austrian Federal Ministry for Transport, Innovation and Technology under the FIT-IT contract FFG 809261.