Estimating Required Recall for Successful Knowledge Acquisition from the Web *

Wolfgang Gatterbauer
Database and Artificial Intelligence Group
Vienna University of Technology, Austria
gatter@dbai.tuwien.ac.at

ABSTRACT

Information on the Web is not only abundant but also redundant. This redundancy of information has an important consequence on the relation between the recall of an information gathering system and its capacity to harvest the core information of a certain domain of knowledge. This paper provides a new idea for estimating the necessary Web coverage of a knowledge acquisition system in order to achieve a certain desired coverage of the contained core information.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Measurement, Theory

Keywords: Information extraction, Recall, Redundancy, Quantitative performance measures, Web metrics

1. INTRODUCTION

Recall is a well established measure in information retrieval (IR) and information extraction (IE) for evaluating the effectiveness of either retrieving relevant documents or extracting relevant statements with statements considered to be the smallest bits of information about entities [5]. For information gathering (IG) or knowledge acquisition from the Web [3, 4] – which can be modeled as a consecutive process of IR, IE and Information Integration (II) – the actual goal is not to locate and extract *all* appearances of relevant information, but to extract as much *unique* information as possible, disregarding all similar or redundant appearances. For domains with large portions of redundant information on the Web, such as digital consumer products or news, the dominant part of relevant information can be gathered even with low overall recall. As an example, a domain-specific recall of $r = .0001$ would probably be enough for learning the information that Shizuka Arakawa won the gold medal in figure skating at the Winter Olympic Games 2006[1].

Therefore, this paper argues to shift the attention from the redundant representation of information to the actual core information stripped off of all redundancy (Fig. 1).

*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.

[1]Approximately 10,000 hits for 'Shizuka Arakawa gold skating olympic games 2006' in Google on February 23, 2006

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

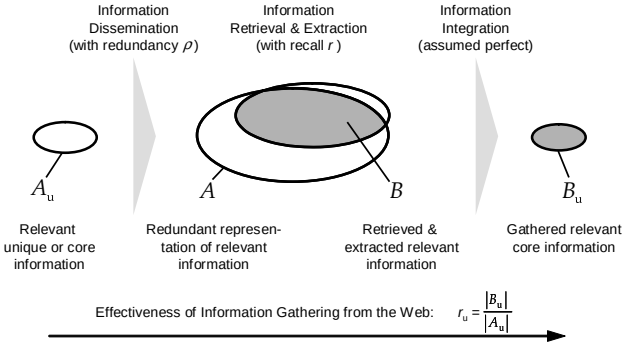


Figure 1: The actual target of IG is the core information A_u, not its redundant representation A.

2. UNIQUE RECALL

In what follows, we consider relevance to be a binary decision. Similarly, we disregard the issue of conflicting information and assume two statements to either express the same information, and hence be redundant, or not.

Definition 1. **(Redundancy)** Let a be the number of relevant statements and a_u be the number of unique statements among them, which is the smallest number of statements that contain all relevant information or the core relevant information. We define the term redundancy as

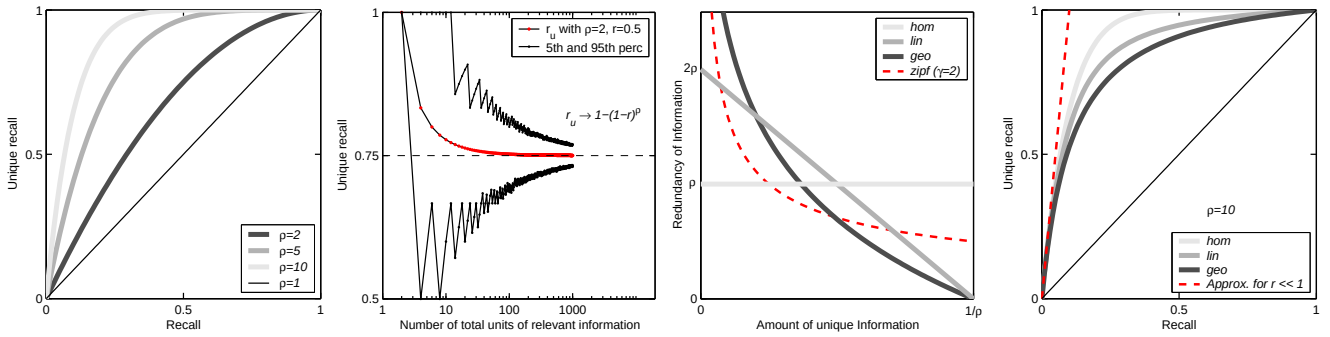
$$\rho = \frac{a}{a_u} \ . \tag{1}$$

Definition 2. **(Unique recall)** Let a_u be the number of unique statements within the total set of relevant statements, and let b_u be the number of gathered unique relevant statements, stripped off of all redundancy. We define the term unique recall as

$$r_u = \frac{b_u}{a_u} \ . \tag{2}$$

PROPOSITION 1. **(Unique recall formula)** *Assume redundancy (ρ) to be equal among all subsets of relevant statements. Further assume the probability of each occurrence of relevant information to be extracted by a given mechanism equal and expressed by the measure recall (r). Then the expected value of unique recall r_u can be approximated by*

$$\overline{r_u} = 1 - (1 - r)^{\rho} \ . \tag{3}$$



(a) $\overline{r_u}(r, \rho)$ (b) $\overline{r_u}(r = 0.5, \rho = 2, n)$ (c) Example canonical $f(\eta)$ (d) $\overline{r_u}$ for example $f(\eta)$

Figure 2: Relations between recall r, unique recall r_u and varying distributions $f(\eta)$ of redundancy ρ.

Figure 2(a) shows equation 3 for example values of ρ. Despite the apparent innocence of this formula, its derivation is not straightforward. The proof succeeds by applying a limit value consideration to a geometric model of a probabilistic lottery. Figure 2(b) illustrates this approximation to be suitable and to hold strongly for $n \to \infty$ with n being the number of unique occurrences of relevant information.

By building on Proposition 1, we can formulate the general equation for arbitrary redundancy distributions with $\rho_{\max}$ layers of partly redundant information as

$$\overline{r_u} = 1 - \sum_{i=1}^{\rho_{\max}} \alpha_i \left(1 - r\right)^i, \tag{4}$$

s.t. $\sum_{i=1}^{n} \alpha_i$ and $\sum_{i=1}^{n} i\alpha_i = \rho$, with α_i being the fractions of the total amount of unique information contained within a block with redundancy $\rho = i = \text{const}$, $i \in \{1, 2, ...\rho_{\max}\}$.

Calculating the first derivative and then taking the limit value for $r \to 0$, we further learn the approximation

$$\overline{r_u} \approx \rho r, \text{ for } r \to 0. \tag{5}$$

Figure 2(d) depicts equation 5 together with the closed solutions for the three canonical redundancy distributions *homogeneous*, *linear* and *geometric* from Fig. 2(c).

3. DISCUSSION AND FUTURE WORK

To the author's best knowledge, no prior attempt has been made to incorporate the effects of redundant information sources into a single effectiveness measure for the whole knowledge acquisition process, nor is the seemingly simple proposition 1 contained in major mathematical literature.

To demonstrate its practical relevance, assume we want to know the fraction r of available redundant information instances we have to retrieve and extract in order to learn a certain portion of the core information from a given domain of interest. Further assume the redundancy distribution of information within this domain to follow a known function $f(\eta)$ with $\eta \in [0, 1]$ and $\int_0^1 f(\eta)d\eta = \rho$. The inverse function ${r_u}^{-1}(r)$ of equation 4 then gives us the required joint recall r_{req} of the IR and IE steps of our IG system.

Deriving a particular analytic solution is not always simple, mainly due to the required transitions between discrete and continuous viewpoints. However, given our derived approximation of $r_u(r)$ for $r \to 0$ from equation 5, which holds even stronger for $r_u \to 0$ with $r_u \geq r$, we can state the following approximation, independent of the actual distribution $f(\eta)$:

$$r_{\text{req}} \approx \frac{r_u}{\rho}, \text{ for small } r_u. \tag{6}$$

Figure 2(d) shows that this relation holds well for small r_u, and since we defined ρ as the average redundancy, this result seems to contrast claims that the feature 'mean' has little practical value for skewed distributions [1].

The intriguing open and relevant problem is to derive the analytic solution for a generalized Zipf redundancy distribution, as this distribution is generally assumed to approximate well the frequency of appearance of individual pieces of information on the Web [6, 7]. A more playful formulation of the research question is as follows: Assume the Zipf distribution with exponential parameter $\gamma \approx 2$ of Fig. 2(c) adequately describes the redundancy distribution of information on the Web [2]. What is the generalization of the 20/80 rule [8] that holds for the Web?

Acknowledgement. The author would like to express his gratitude to Professor Georg Gottlob for overall motivation of this research, and to Konrad Richter for always having time to discuss some interesting mathematical problems.

4. REFERENCES

[1] Z. Bi, C. Faloutsos, and F. Korn. The "DGX" distribution for mining massive, skewed data. In *Proc. KDD*, pages 17–26. ACM, 2001.

[2] S. Bornholdt and H. Ebel. World Wide Web scaling exponent from Simon's 1955 model. *Physical Review E*, 64:035104, 2001.

[3] M. Craven, D. DiPasquo, D. Freitag, A. K. McCallum, T. M. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence*, 118(1/2):69–113, 2000.

[4] O. Etzioni, M. J. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Methods for domain-independent information extraction from the Web: An experimental comparison. In *Proc. AAAI*, pages 391–398. AAAI, 2004.

[5] W. Gatterbauer, B. Krüpl, W. Holzinger, and M. Herzog. Web information extraction using eupeptic data in Web tables. In *Proc. RAWS*, pages 41–48. VSB-TU Ostrava, 2005.

[6] P. G. Ipeirotis and L. Gravano. Distributed search over the hidden web: hierarchical database sampling and selection. In *Proc. VLDB*, pages 394–405. Morgan Kaufmann, 2002.

[7] D. A. Shamma, S. Owsley, K. J. Hammond, S. Bradshaw, and J. Budzik. Network arts: exposing cultural reality. In *Proc. WWW*. ACM, 2004.

[8] Wikipedia. Pareto principle, 2006. Available: http://en.wikipedia.org/wiki/Pareto_principle (February 2006).