Discussion Section

APRIL 15

Syllabus (I)

- Search
  - Overview
  - Retrieval Models
    - Vector Space Models
    - Language Modeling
    - Axiomatic Approaches
  - Evaluation
  - Indexing + Query Processing
  - Feedback
  - Web Search
    - Basics
    - Link analysis
    - Machine learning for Web Search

Syllabus (II)

- Mining
  - Word association mining
  - Categorization
  - Clustering

Projects

- Comments due on April 13
- Many interesting ideas!!!
- Re-think about the scope of your project given the tighter timeline
- Schedule a meeting with me if you feel uncertain

Midterm

- Due on April 20
- Search related application
- Questions?

Inverted Index Compression

6

- Observations:
  - TF compression
    - Small numbers tend to occur more frequently.
  - Doc ID compression
    - "d-gap" (store differences): d1, d2-d1, d3-d2, ..
    - Feasible due to sequential access
- Implications
  - Exploit skewed frequency distribution and use variable-length encoding

Integer Compression Methods

7

- Binary: equal-length coding
  - 3=>00000011; 5=>00000101
- Unary: x≥1 is coded as x-1 one bits followed by 0
  - 3=> 110; 5=>11110
- γ-code: x=> unary code for $1+\lfloor \log x \rfloor$ followed by binary code for $x-2^{\lfloor \log x \rfloor}$ in $\lfloor \log x \rfloor$ bits
  - 3=>101, 5=>11001
- δ-code: same as γ-code, but replace the unary prefix with γ-code.
  - 3=>1001, 5=>10101

Examples

8

γ-code: x=> unary code for $1+\lfloor \log x \rfloor$ followed by binary code for $x-2^{\lfloor \log x \rfloor}$ in $\lfloor \log x \rfloor$ bits

- Compute the compression results using gamma compression for the following values
  - 2 unary code for 2 (10). binary code for 0 in 1 bit (0) → 10 0
  - 4 unary code for 3 (110). binary code for 0 in 2 bits (00) → 110 00
  - 14 unary code for 1+3(1110). Binary code for 6 in 3 bits (110) → 1110 110
  - 63 unary code for 6 (111110), binary code 31 in 5 bits (11111) → 111110 11111
  - 180

Examples

9

- Compute the compression results using gamma compression for the following values
  - 2 10 0
  - 4 110 00
  - 14 1110 110
  - 63 111110 11111
  - 180 11111110 01101000

Retrieval as Language Model Estimation

10

- Rank documents based on *query likelihood*

$$q = w_1, w_2, \ldots, w_n$$

$$p(q \mid d) = p(w_1 \mid d) \times p(w_2 \mid d) \times \ldots \times p(w_n \mid d).$$

$$\text{score}(q, d) = \log p(q \mid d) = \sum_{i=1}^{n} \log p(w_i \mid d) = \sum_{w \in V} c(w, q) \log p(w \mid d)$$

- Retrieval problem ≈ Estimation of $p(w|d)$

Dirichlet Prior Function

11

$$\sum_{w \in d, q} c(w, q) \log \left(\frac{p_{\text{seen}}(w \mid d)}{\alpha_d \cdot p(w \mid C)} \right) + |q| \log \alpha_d.$$

- **Dirichlet Prior Smoothing**

$$p(w \mid d) = \frac{c(w;d)+\mu\, p(w|C)}{|d|+\mu} = \frac{|d|}{|d|+\mu} p_{ml}(w \mid d) + \frac{\mu}{|d|+\mu} p(w \mid C)$$

$$\alpha_d = \frac{\mu}{|d|+\mu}:$$

$$\text{score}_{DIR}(q, d) = \sum_{w \in q, d} c(w, q) \log \left(1 + \frac{c(w, d)}{\mu \cdot p(w \mid C)} \right) + |q| \log \frac{\mu}{\mu + |d|}.$$

Dirichlet prior smoothing

Suppose we have a document collection with an extremely small vocabulary with only 6 words w1,…,w6. Let Q=w1w2 be a query.

1. Suppose we do not smooth the language model for d1 and d2. Compute the likelihood of the query for both d1 and d2, i.e., P(Q|d1) and P(Q|d2). (Do not compute the log-likelihood.) Show your calculations. Which document would be ranked higher?

   s(Q,d1) = P(Q|d1) = p(w1|d1) * p(w2|d1) = c(w,d1)/|D| * c(w,d2)/|D| = 2/10*3/10
   s(Q,d2) = 7/10 * 1/10 d2 will be ranked higher.

1. Suppose we now smooth the language model for d1 and d2 using Dirichlet prior smoothing method with **u=10**. Recompute the likelihood of the query for both d1 and d2, i.e., P(Q|d1) and P(Q|d2). Show your calculations. Which document would be ranked higher this time?

$$p(w \mid d) = \frac{c(w;d)+\mu\, p(w|C)}{|d|+\mu} = \frac{|d|}{|d|+\mu} p_{ml}(w \mid d) + \frac{\mu}{|d|+\mu} p(w \mid C)$$

S(Q,D1) = p_dir(w1|d1) * p_dir (w2|d1) = (10/(10+10) * 2/10 + 10/(10+10) *0.8) * … = 0.1
S(Q,D2) = 0.07

1. Intuitively, which document do you think show be ranked higher? D1 or d2? Why?

Word	P(w | REF)	C(w,d1)	C(w,d2
W1	0.8	2	7
W2	0.1	3	1
W3	0.025	2	1
W4	0.025	2	1
W5	0.025	1	0
W6	0.025	0	0
SUM	1.0	10	10