Relevance Models

Berlin Chen 2005

References:

1. W. B. Croft and J. Lafferty (Editors). *Language Modeling for Information Retrieval*. Chapter 2, July 2003
2. V. Lavrenko and W.B. Croft, "Relevance-Based Language Models" ACM SIGIR 2001

Introduction

- Probabilistic Models (e.g., Robertson and Sparck Jones, 1976)
  - Ranking docs by the odds of their being observed in the relevant class

$$SIM(Q,D) \approx \frac{P(D|R_Q)}{P(D|\overline{R}_Q)}$$

  - Explicitly model the class/concept of relevance

- Language Modeling Approaches
  - View documents themselves as models and queries as strings of text (observations) randomly sampled from these models

$$SIM(Q,D) \approx P(Q|D) \approx \prod_{q_n \in Q} P(q_n|D)$$ bag-of-words assumption

    - E.g., as the HMM, TMM, PLSA, mentioned previously

  - Quite several probability estimation techniques were proposed

Relevance Models (RM)

- A convergence of ideas from classical probabilistic approaches and language modeling approaches
  - Explicit model the class R of relevant documents
  - Describe unigram generative models for a set of text samples(in the relevant class" $P(w|R)$
    - Documents are randomly drawn from such a class

RM: Ranking Approaches

- Probability Ratio $\left(\text{given that } D = d_1 \cdots d_n\right)$

$$\frac{P(D|R)}{P(D|N)} = \frac{P(d_1 \cdots d_n|R)}{P(d_1 \cdots d_n|N)} \approx \frac{\prod_{i=1}^{n} P(d_i|R)}{\prod_{i=1}^{n} P(d_i|N)} \approx \frac{\prod_{i=1}^{n} P(d_i|R)}{\prod_{i=1}^{n} P(d_i|C)}$$

  - Higher score is better (more relevant)

- R : the relevant class
- N : the non-relevant class
- C : the document collection or corpus

$\left(|R| \text{ is small} \Rightarrow |N| \approx |C|, \text{ all measured in terms of word number}\right)$

- $P(d_i|R)$ or $P(w|R)$ can be estimated from a set of training examples or from the query alone (a focus of much active research)

(Inspired by the probability ranking principle)

RM: Ranking Approaches (cont.)

- Cross-Entropy

$$P(R \| D) = -\sum_{w \in V} P(w|R) \log P(w|D)$$

vocabulary

  - Lower score (≥ 0) is better (more relevant)

  - The negation of this ranking approach is equivalent to "query-likelihood" language-modeling approach (a special case of the cross-entropy approach itself ?)
    - Suppose $P(w|R)$ is the relative frequency of w in the query Q

$$-P(R \| D) \approx \sum_{w \in V} P(w|Q) \log P(w|D) = \sum_{w \in V} \frac{C(w,Q)}{|Q|} \log P(w|D)$$

$$= \frac{1}{|Q|} \sum_{w \in V} \log P(w|D)^{C(w,Q)} = \frac{1}{|Q|} \log \prod_{w \in V} P(w|D)^{C(w,Q)}$$

$$= \frac{1}{|Q|} \log \boxed{P(Q|D)}$$

RM: Estimation from a Set of Examples

- Given that we have perfect knowledge of the entire relevant class R

$$
\begin{aligned}
P(w|R) &= \sum_{D \in Collection} P(w, D|R) \\
&= \sum_{D \in Collection} \underline{P(w|D,R)}P(D|R) \\
&= \sum_{D \in Collection} \underline{P(w|D)}P(D|R) \\
&= \sum_{D \in Collection} P_{ML}(w|D)P(D|R)
\end{aligned}
$$

conditional independence assumption

– Where

$$
P_{ML}(w|D) = \frac{C(w,D)}{|D|} \qquad P(D|R) = \begin{cases} 1/|R| & \text{if } D \in R \\ 0 & \text{otherwise} \end{cases}
$$

no. of documents in R

RM: Estimation from a Set of Examples (cont.)

- Probability smoothing by simple interpolation

$$P_{smooth}(w|D) = \lambda_D P_{ML}(w|D) + (1-\lambda_D)P(w)$$

where

no. of documents in the collection

$$P(w) = \sum_{D \in C} P_{ML}(w|D) \cdot \frac{1}{|C|}$$

  – Which is also equivalent to adjust document posterior probability

$$P_{smooth}(D|R) = \begin{cases} \frac{(1-\lambda_D)}{|C|} + \frac{\lambda_D}{|R|} & \text{if } D \in R \\ \frac{(1-\lambda_D)}{|C|} & \text{otherwise} \end{cases}$$