Improving Pseudo-Relevance Feedback Using Pseudo-Irrelevant Documents

Karthik Raman¹, Raghavendra Udupa² and Pushpak Bhattacharya¹
¹Indian Institute of Technology Bombay
²Microsoft Research India

**PRF and Motivation behind Pseudo-Irrelevance**

- Top “k” assumed Relevant (PRₖ)
- Contains lot of irrelevant documents (Xₙ)
- Can we use Pseudo-Relevant Documents to Identify some of the seemingly Irrelevant Documents?

**Using Pseudo-Irrelevant Documents**

- Use Rocchio Algorithm with PI as –ve Feedback
- Zhang et. al ‘s Distribution Separation Model
- Other Negative Feedback Algorithms

**“Pseudo-Irrelevant Documents” & their Identification**

\[ Y_{PR} = \bigcup_{D} Q_{D} \quad \text{for} \quad \forall D \in PR_{k} \]

where \( Q_{D} : \) Top Documents retrieved if \( D \) was a query

\[ PI: \text{Pseudo – Irrelevant Documents} \]

\[ P = X_{n} - (X_{n} \cap Y_{PR}) \]

**Query Expansion Using “Discriminative Terms”**

Terms which distinguish Documents in \( PR_{k} \) from those in \( PI \)

- Run Logistic Classifier with:
  - \( PR_{k} : +ve \); \( PI : -ve \)
  - Term TF-IDF Values as Features

Expand Query with Most discriminative terms i.e. highest +ve feature weights,

**Preliminary Results?**

**What Next?**

- Can Use Pseudo-Irrelevant Documents to Identify Irrelevant Documents in \( PR_{k} \).
- Pseudo-Irrelevant Documents found to be closer to Irrelevant than Relevant Documents

**CLEF ’00,’01,’02**
**CLEF ’03,’05,’06**
**TREC-AP**
**TREC-WSJ**

*MBF – Model Based Feedback*