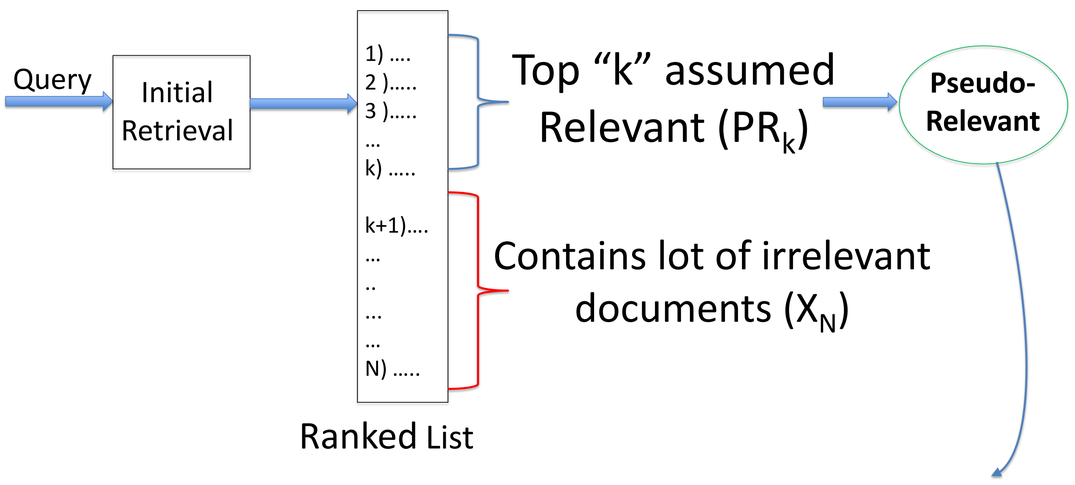
# Improving Pseudo-Relevance Feedback Using Pseudo-Irrelevant Documents

**Karthik Raman[1], Raghavendra Udupa[2] and Pushpak Bhattacharya[1]**
[1]Indian Institute of Technology Bombay
[2]Microsoft Research India

## PRF and Motivation behind Pseudo-Irrelevance




Can we use Pseudo-Relevant Documents to Identify some of the seemingly Irrelevant Documents?

## "Pseudo-Irrelevant Documents" & their Identification

High-Scoring Documents in $X_N$ (Ranked *k+1* to N) which are dissimilar to pseudo-relevant docs.

$Y_{PR}$: Documents similar to $PR_k$

$$Y_{PR} = \bigcup_{D} Q_D \text{ for } \forall D \in PR_K$$

where $Q_D$ : Top Documents retrieved if D was a query

PI: Pseudo – Irrelevant Documents

$$PI = X_N - (X_N \cap Y_{PR})$$

## Using Pseudo-Irrelevant Documents

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

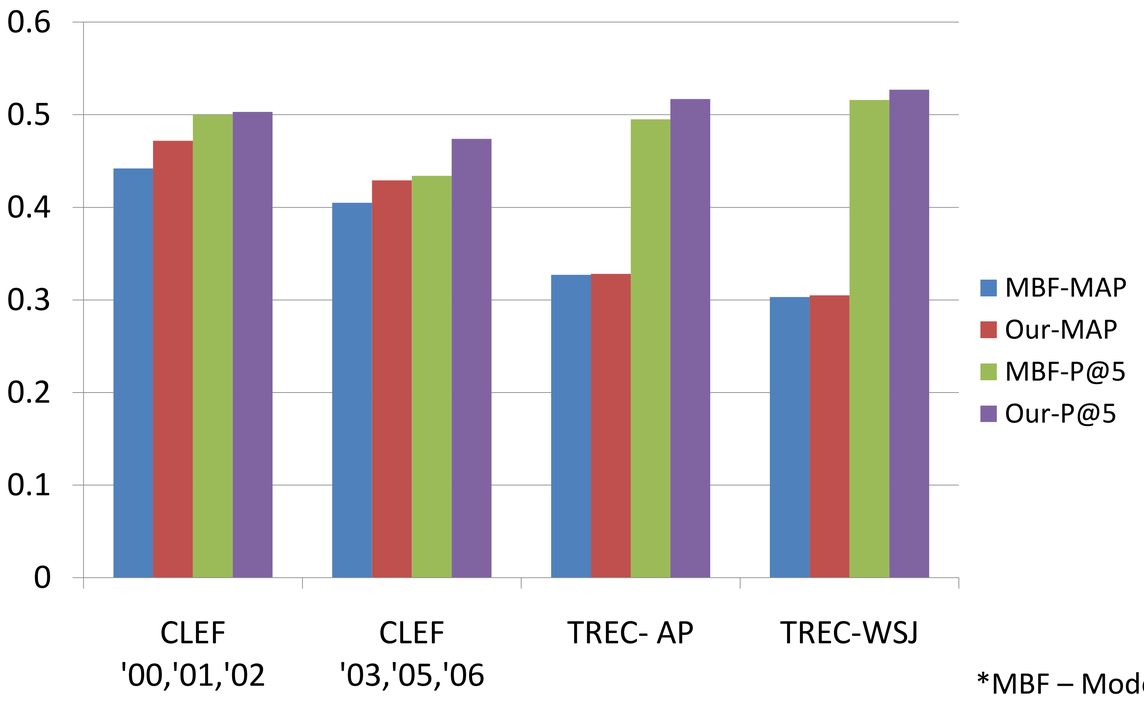
## Query Expansion Using "Discriminative Terms"

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?




*MBF – Model Based Feedback

## 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