



Generalized Inverse Document Frequency

Donald Metzler

Yahoo! Research

October 27, 2008

CIKM 2008



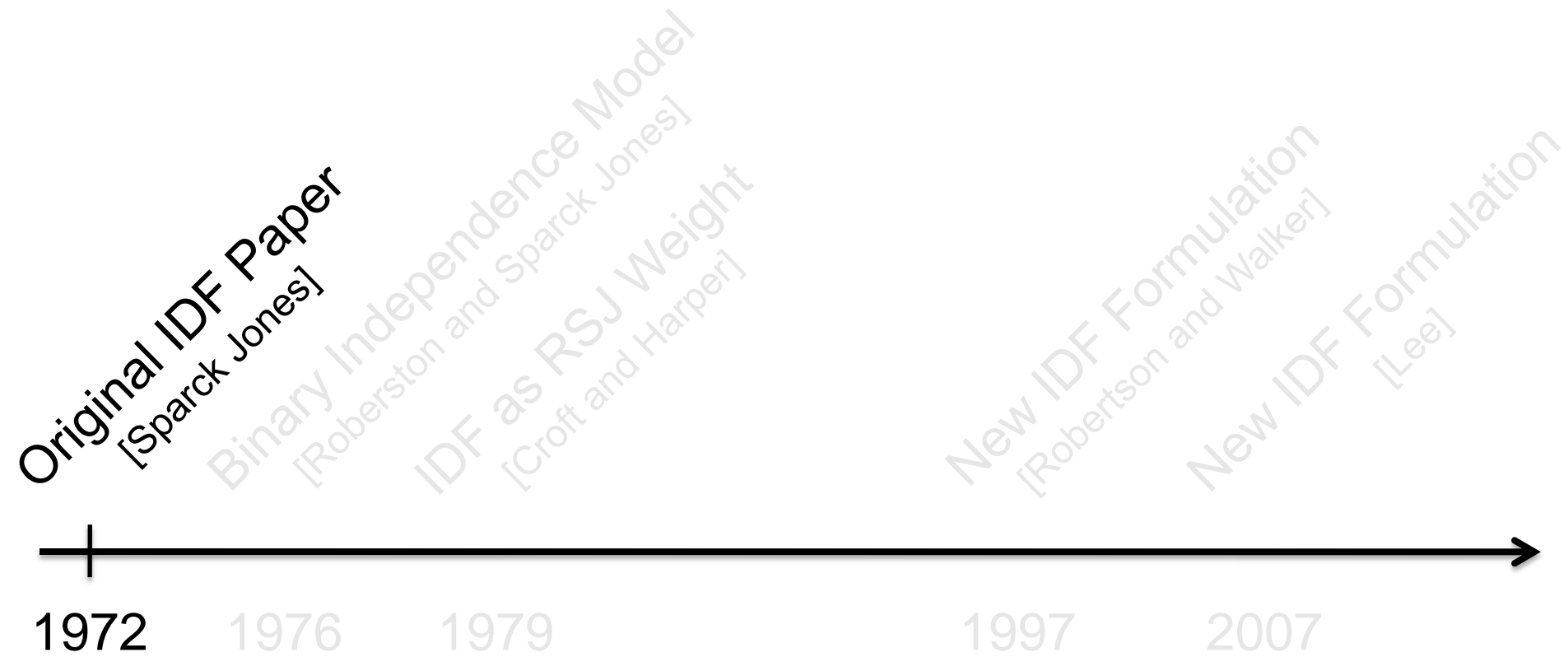
What is IDF?

- Global measure of the importance of an identifier (word, phrase, etc.)
- Used in a variety of tasks
 - Information retrieval
 - Text classification
- Classical formulation:

$$IDF(w) = \log \frac{N}{N_w}$$

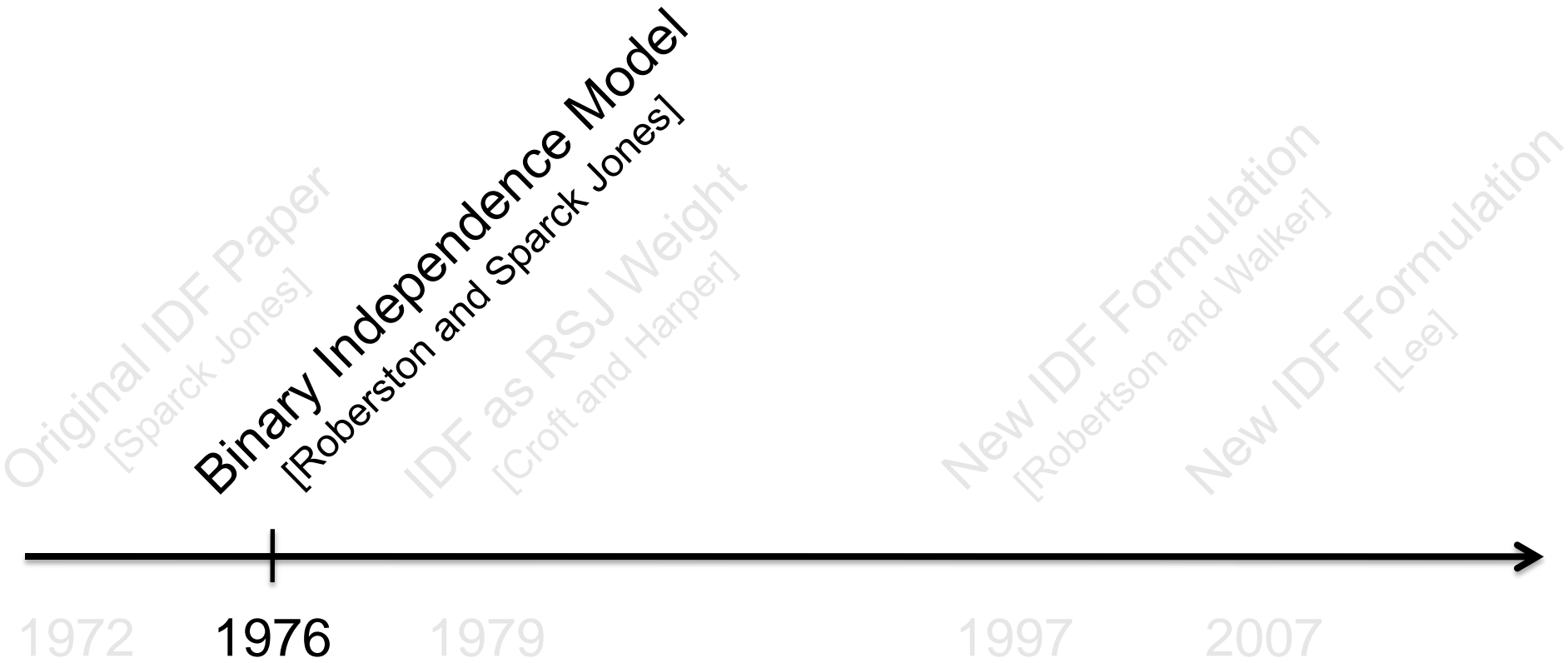


IDF Timeline



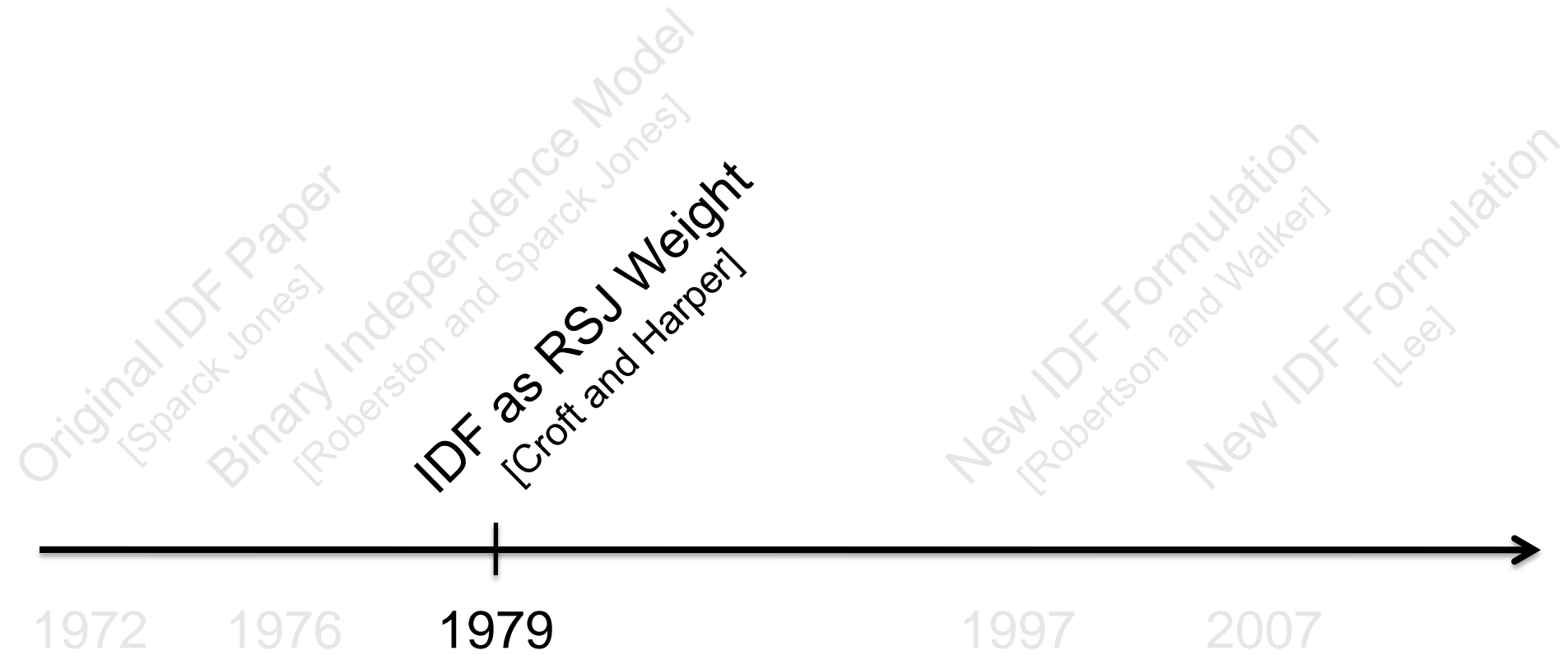


IDF Timeline



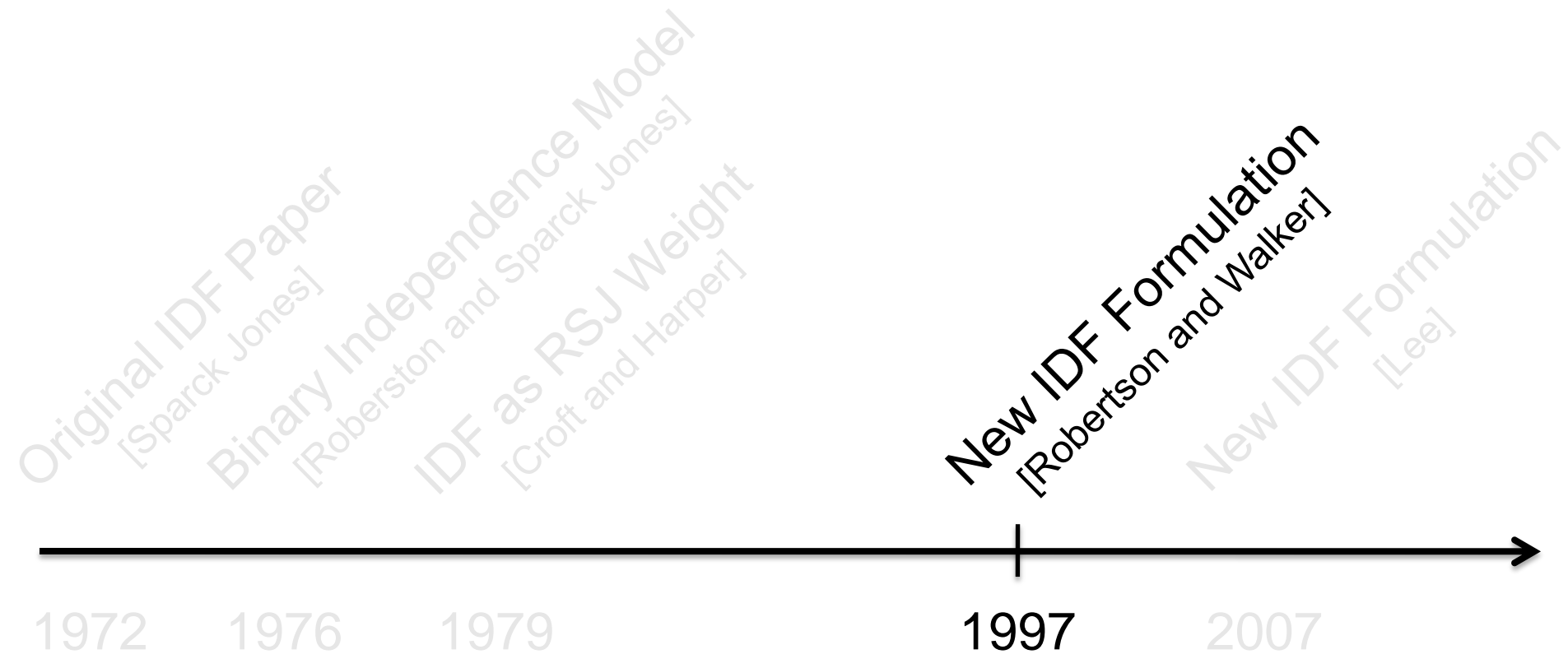


IDF Timeline



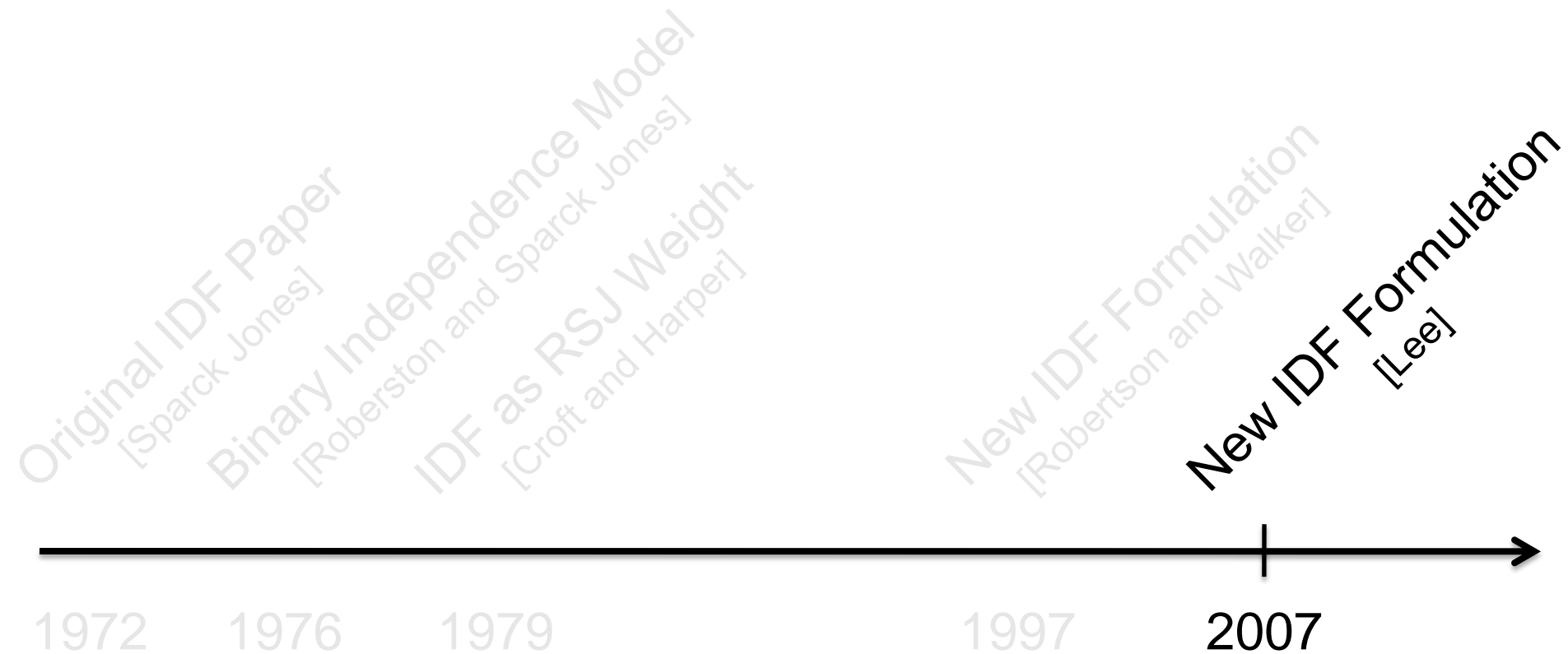


IDF Timeline





IDF Timeline





Why Study IDF?

- Term frequency and document length normalization are focus of many studies
- Inverse document frequency is often overlooked and not very well understood
- Momentum building for improved understanding and modeling of IDF



Common IDF Formulations

- Robertson-Sparck Jones IDF:

$$wt_i = \log \frac{|C| - c(d_i, C) + 0.5}{c(d_i, C) + 0.5}$$

- Robertson and Walker IDF:

$$wt_i = \log \frac{|C| + 0.5}{c(d_i, C) + 0.5}$$

- Both can be derived as a RSJ weight from the BIR model



Generalized IDF

$$\begin{aligned}RSV(q, d, J) &= \frac{P(r, |d, q, J)}{P(\bar{r}, |d, q, J)} \\&= \log \prod_{i=1}^{|\mathcal{V}|} \frac{P(d_i|q, r, J)P(q, r, J)}{P(d_i|q, \bar{r}, J)P(q, \bar{r}, J)} \\&\stackrel{rank}{=} \log \prod_{i=1}^{|\mathcal{V}|} \frac{P(d_i|q, r, J)}{P(d_i|q, \bar{r}, J)} \\&\stackrel{rank}{=} \sum_{i:d_i=1} \log \frac{P(d_i|q, r, J)P(\bar{d}_i|q, \bar{r}, J)}{P(\bar{d}_i|q, r, J)P(d_i|q, \bar{r}, J)}\end{aligned}$$

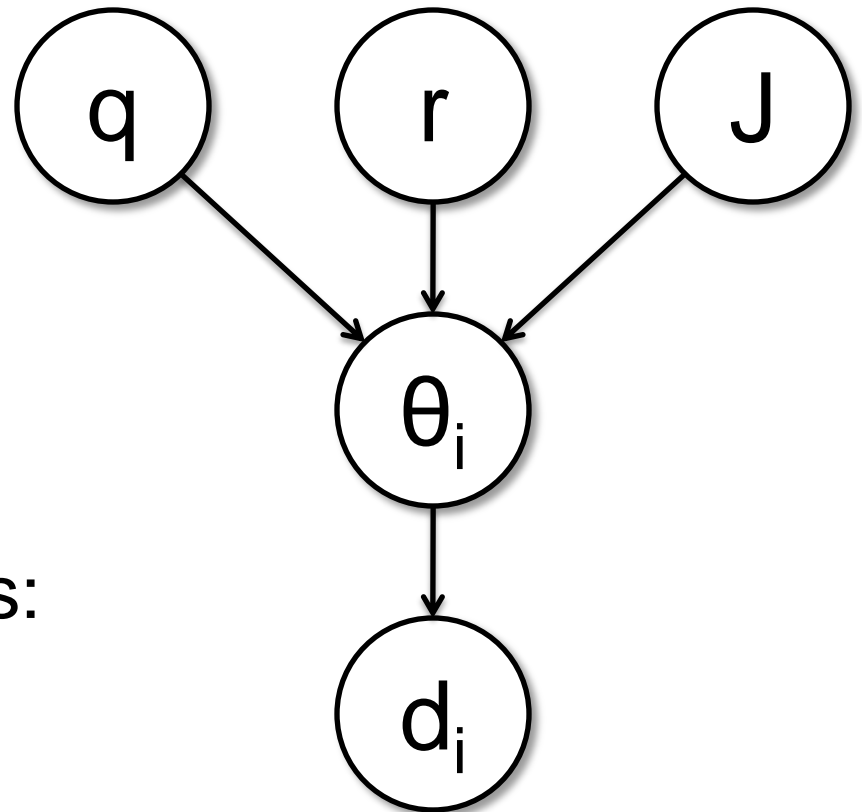


Document Generation Model

- A model (θ) is sampled given a query (q), relevance class (r), and judgments (J)
- A term (d_i) event is sampled from the model
- Distributional assumptions:

$$\theta_i \sim \text{Beta}(\alpha_i, \beta_i)$$

$$d_i \sim \text{Bernoulli}(\theta_i)$$





Generalized IDF

- Estimates

$$\begin{aligned} P(d_i|q, r, J_r) &= \int_{\theta} P(d_i|\theta)P(\theta|q, r, J)d\theta \\ &= \mathbb{E}[\theta|q, r, J] \\ &= \frac{\alpha_i(q, r, J)}{\alpha_i(q, r, J) + \beta_i(q, r, J)} \end{aligned}$$

- Generalized IDF

$$wt_i = \underbrace{\log \frac{\alpha_i(q, r, J)}{\beta_i(q, r, J)}}_{IDF_r} + \underbrace{\log \frac{\beta_i(q, \bar{r}, J)}{\alpha_i(q, \bar{r}, J)}}_{IDF_{\bar{r}}}$$



Generalized IDF

- Different settings of hyperparameters give different IDF formulations
- Various ways to estimate
 - Collection statistics
 - (Pseudo-)Relevance feedback
 - Click data
- We propose and evaluate several simple estimates as “proof of concept”



Relevance Distribution: Assumption Set 1

Hyperparameters

$$\frac{\alpha_i(q, r)}{\beta_i(q, r)} = \frac{\gamma}{1 - \gamma}$$

Estimate

$$P(d_i | q, r) = \gamma$$

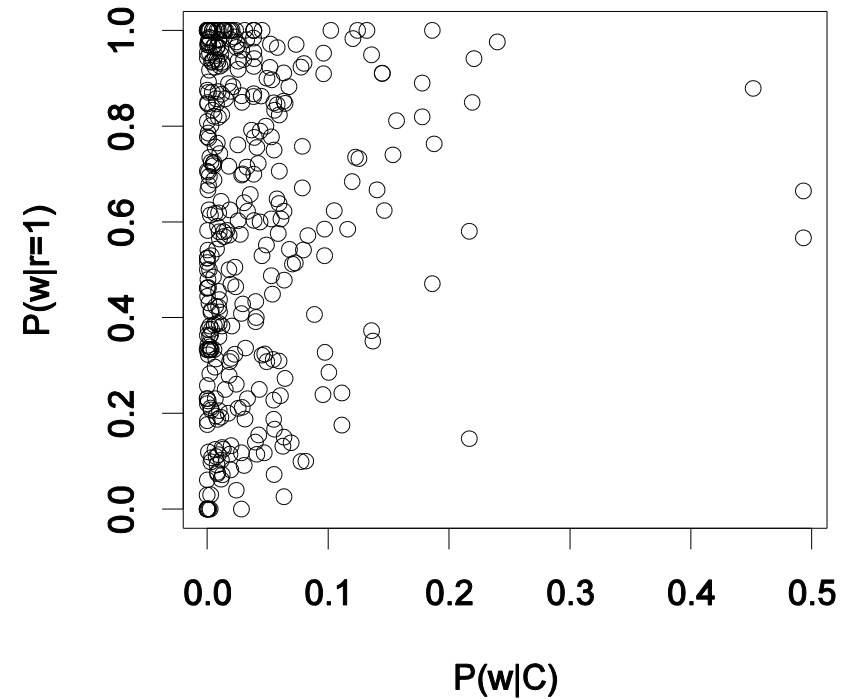
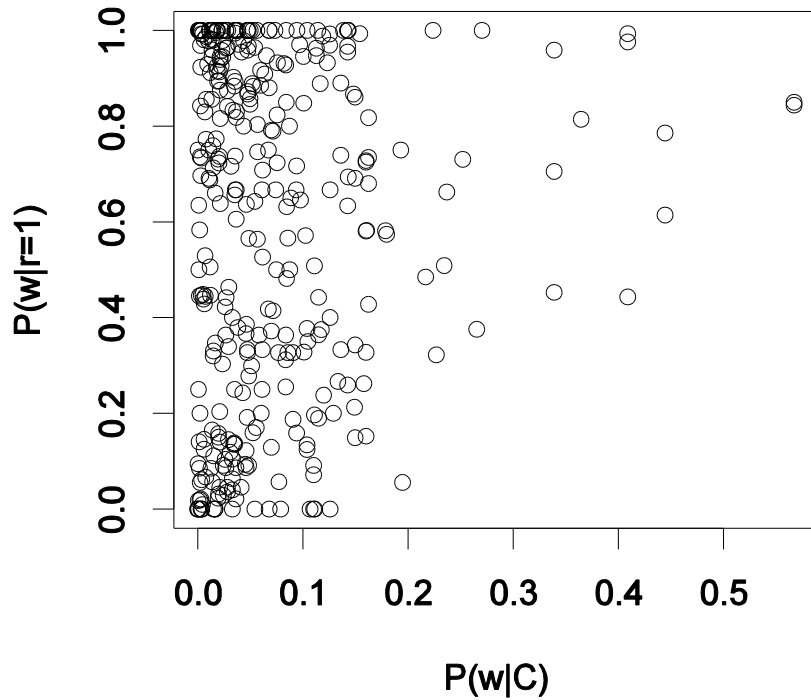
IDF

$$IDF_r = \frac{\gamma}{1 - \gamma}$$

$P(d | q, r)$ is constant for all terms.



Empirical Relevance Distributions





Relevance Distribution: Assumption Set 2

Hyperparameters

$$\alpha_i(q, r) = (1 - \lambda) \cdot \mathbb{E}[\theta|r] + \lambda P(d_i|C)$$

$$\beta_i(q, r) = 1 - \alpha_i(q, r)$$

Estimate

$$P(d_i|q, r) = (1 - \lambda) \cdot \mathbb{E}[\theta|r] + \lambda P(d_i|C)$$

IDF

$$IDF_r = \log \frac{(1 - \lambda) \cdot \mathbb{E}[\theta|r] \cdot |C| + \lambda c(d_i, C)}{|C| - \lambda c(d_i, C) - (1 - \lambda) \cdot \mathbb{E}[\theta|r] \cdot |C|}$$

$P(d | q, r)$ is linearly smoothed between empirical mean and collection model.



Non-relevance Distribution: Assumption Set 1

Hyperparameters

$$\frac{\alpha_i(q, \bar{r})}{\beta_i(q, \bar{r})} = \frac{\gamma}{1 - \gamma}$$

Estimate

$$P(d_i | q, \bar{r}) = \gamma$$

IDF

$$IDF_{\bar{r}} = \frac{\gamma}{1 - \gamma}$$

$P(d | q, \bar{r})$ is constant for all terms.



Non-relevance Distribution: Assumption Set 2

Hyperparameters

$$\alpha_i(q, \bar{r}) = c(d_i, C) + \gamma$$

$$\beta_i(q, \bar{r}) = |C| - c(d_i, C) + \gamma$$

Estimate

$$P(d_i | q, \bar{r}) = \frac{c(d_i, C) + \gamma}{|C| + 2\gamma}$$

IDF

$$IDF_{\bar{r}} = \log \frac{|C| - c(d_i, C) + \gamma}{c(d_i, C) + \gamma}$$

$P(d | q, \bar{r})$ is increasing with $P(w | C)$



Non-relevance Distribution: Assumption Set 3

Hyperparameters

$$\alpha_i(q, \bar{r}) = c(d_i, C) + \gamma$$

$$\beta_i(q, \bar{r}) = |C| + \gamma$$

Estimate

$$P(d_i | q, \bar{r}) = \frac{c(d_i, C) + \gamma}{|C| + c(d_i, C) + 2\gamma}$$

IDF

$$IDF_{\bar{r}} = \log \frac{|C| + \gamma}{c(d_i, C) + \gamma}$$

$P(d | q, \bar{r})$ is increasing with $P(w | C)$



Non-relevance Distribution: Assumption Set 4

Hyperparameters

$$\alpha_i(q, \bar{r}) = ((1 - \lambda) \cdot \mathbb{E}[\theta|\bar{r}] + \lambda P(d_i|C))$$

$$\beta_i(q, \bar{r}) = 1 - \alpha_i(q, \bar{r})$$

Estimate

$$P(d_i|q, \bar{r}) = (1 - \lambda) \cdot \mathbb{E}[\theta|\bar{r}] + \lambda P(d_i|C)$$

IDF

$$IDF_{\bar{r}} = \log \frac{|C| - \lambda c(d_i, C) - (1 - \lambda) \cdot \mathbb{E}[\theta|\bar{r}] \cdot |C|}{(1 - \lambda) \cdot \mathbb{E}[\theta|\bar{r}] \cdot |C| + \lambda c(d_i, C)}$$

$P(d | q, -r)$ is linearly smoothed between empirical mean and collection model.



Experimental Methodology

- Data
 - Three TREC news collections (AP, WSJ, ROBUST 2004)
 - One TREC web collection (WT10G)
- Mean average precision (MAP) used for evaluation
- Parameter Estimation
 - Tuned to maximize MAP on training set
 - Evaluated on test set

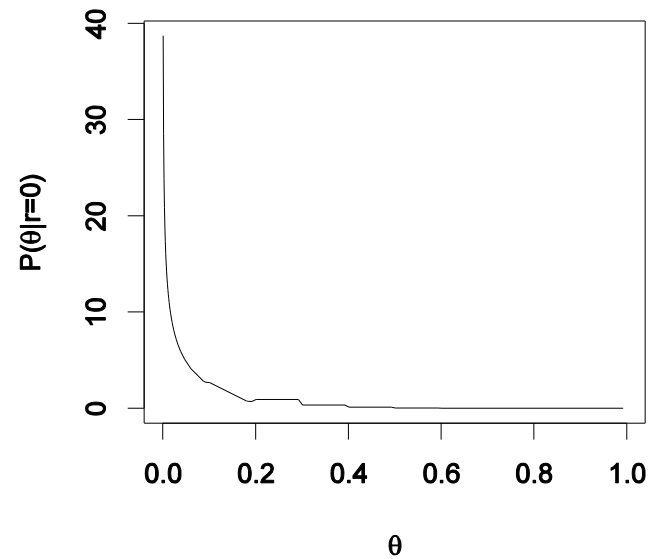
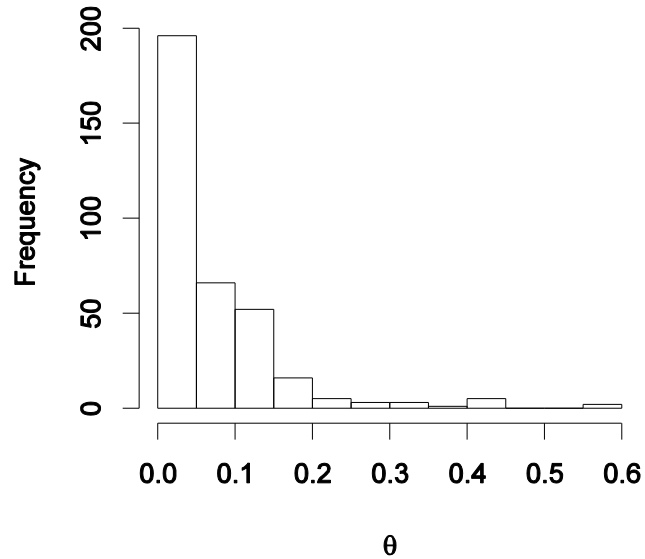
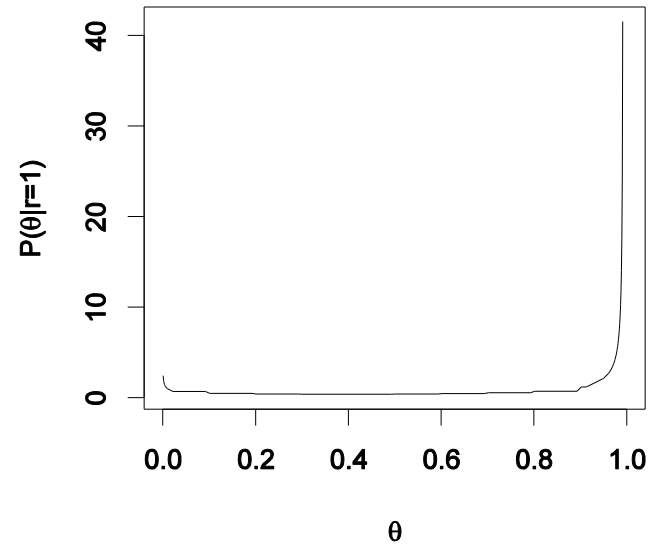
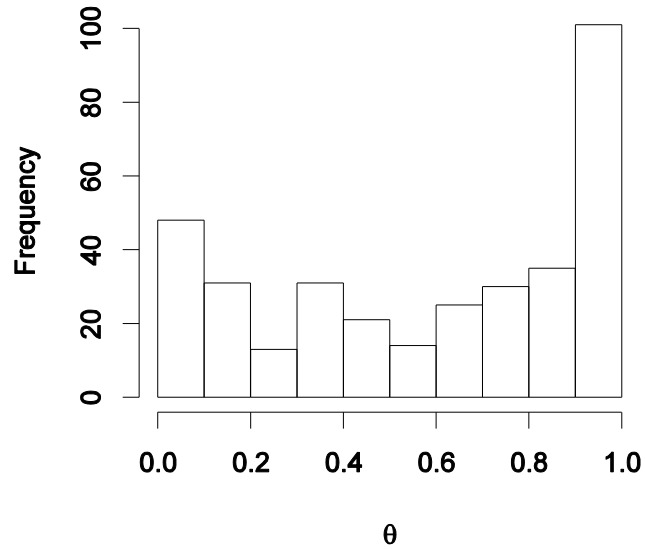


Hyperparameters

		Relevant				Non-Relevant			
		$\mathbb{E}[\theta r]$	$Var[\theta r]$	α	β	$\mathbb{E}[\theta \bar{r}]$	$Var[\theta \bar{r}]$	α	β
WSJ	Train	.6103	.1086	.7260	.4637	.0688	.0075	.5179	7.007
	Test	.5930	.1139	.6640	.4558	.0605	.0053	.5900	9.161
AP	Train	.5881	.1236	.5645	.3954	.0702	.0072	.5687	7.529
	Test	.5663	.1206	.5874	.4499	.0587	.0033	.9292	14.889
ROBUST	Train	.6108	.1001	.8397	.5352	.0366	.0036	.3262	8.575
	Test	.7230	.0952	.7974	.3056	.0370	.0022	.5647	14.69
WT10G	Train	.7434	.1035	.6270	.2165	.0429	.0062	.2429	5.422
	Test	.7380	.0986	.7089	.2517	.0476	.0043	.4589	9.182
ALL		.6321	.1124	.6756	.3932	.0539	.0052	.4743	8.327



Goodness of Beta Fit





IDF Experiments

- IDF-only ranking function

$$S(Q, D) = \sum_{i:d_i=1, q_i=1} (IDF_r + IDF_{\bar{r}})$$

- By eliminating TF and document length frequency we can directly quantify the impact of new IDF formulations



IDF Results

Assumption		AP		WSJ		ROBUST		WT10G	
IDF_r	$IDF_{\bar{r}}$	Train	Test	Train	Test	Train	Test	Train	Test
1	1	.0601	.0758	.1000	.1328	.1123	.0996	.0494	.0367
1	2	.0767	.0096	.1312	.1782	.1416	.1257	.0621	.0555 ^{$\alpha\beta$}
1	3	.0771	.0969 ^{α}	.1308	.1779 ^{α}	.1417	.1257	.0625	.0594 ^{β}
1	4	.0765	.0963	.1305	.1782	.1416	.1257	.0620	.0553
2	1	.0601	.0758	.1000	.1328	.1123	.0996	.0494	.0367
2	2	.0775	.0971	.1314	.1830 ^{$\alpha\beta$}	.1421	.1254	.0623	.0556
2	3	.0778	.0969 ^{α}	.1313	.1829 ^{$\alpha\beta$}	.1421	.1254	.0623	.0556 ^{$\alpha\beta$}
2	4	.0765	.0971	.1307	.1837 ^{$\alpha\beta$}	.1421	.1254	.0621	.0555
RSJ		.0766	.0967	.1301	.1767	.1405	.1255	.0610	.0553
RSJ Positive		.0766	.0967	.1296	.1783	.1405	.1255	.0610	.0553

Bold indicates the best formulation for each data set. The superscripts α and β indicate statistically significant improvements over RSJ and RSJ Positive, respectively, at the $p < 0.1$ level. Underlined superscripts are significant at the $p < 0.05$ level. Significance tests were only performed on the test sets.



TF.IDF Experiments

- Okapi TF

$$\hat{tf}(d_i, D) = \frac{(k_1 + 1) \cdot tf(d_i, D)}{k_1 \cdot (1 - b + b \cdot \frac{|D|}{|D|_{avg}}) + tf(d_i, D)}$$

- Ranking function

$$S(Q, D) = \sum_{i:d_i=1, q_i=1} \hat{tf}(d_i, D) \cdot (IDF_r + IDF_{\bar{r}})$$

- Does TF dampen IDF improvements?



TF.IDF Results

Assumption		AP		WSJ		ROBUST		WT10G	
IDF_r	$IDF_{\bar{r}}$	Train	Test	Train	Test	Train	Test	Train	Test
1	1	.1487	.1941	.2123	.2793	.1978	.2599	.1984	.1469
1	2	.1778	.2148	.2550	.3332	.2285	.2886	.2225	.1965 ^{β}
1	3	.1788	.2151	.2559	.3347 ^{$\alpha\beta$}	.2285	.2886	.2230	.1973 ^{β}
1	4	.1752	.2149	.2532	.3332	.2285	.2886	.2224	.1965 ^{β}
2	1	.1487	.1941	.2123	.2793	.1978	.2599	.1984	.1469
2	2	.1791	.2147	.2563	.3369 ^{$\alpha\beta$}	.2283	.2886	.2263	.2006 ^{$\alpha\beta$}
2	3	.1797	.2137	.2576	.3341	.2291	.2896	.2253	.1987 ^{β}
2	4	.1793	.2125	.2557	.3316	.2282	.2900	.2247	.2005 ^{$\alpha\beta$}
RSJ		.1778	.2149	.2550	.3332	.2279	.2892	.2225	.1965
RSJ Positive		.1781	.2147	.2147	.3332	.2283	.2888	.2217	.1947

Bold indicates the best formulation for each data set. The superscripts α and β indicate statistically significant improvements over RSJ and RSJ Positive, respectively, at the $p < 0.1$ level. Underlined superscripts are significant at the $p < 0.05$ level. Significance tests were only performed on the test sets.



Conclusions

- Derived a generalized IDF formulation
 - Can be used to derive new and improved IDF formulations
 - Provides better understanding of IDF
- Proposed several new IDF formulations that are more effective than current “best practice” IDFs
- Recommended IDF formulation:

$$IDF_{2,3}(d_i) = \underbrace{\log \frac{(1 - \lambda) \cdot \delta \cdot |C| + \lambda c(d_i, C)}{|C| - \lambda c(d_i, C) - (1 - \lambda) \cdot \delta \cdot |C|}}_{\text{modified } IDF_r \text{ assumption 2}} + \underbrace{\log \frac{|C| + \gamma}{c(d_i, C) + \gamma}}_{IDF_r \text{ assumption 3}}$$