Handling irrelevant data using weighted entropy and harmonic function

Yi ZHU
CSE@CUHK

Outline

- Motivation
- Semi-supervised Learning
- Minimize the Weighted Entropy
- Application
- Experiment Result

Motivation

- Classification with irrelevant data or noise;
- Unbalanced situation
- Personalized recommendation

Semi-supervised Learning

- Labeled data set: $L = (x_1, y_1), \ldots\ldots, (x_l, y_l)$
- Unlabeled data set: $U = x_{l+1}, \ldots\ldots, x_n, n = l + u$
- Binary label: $y_L \in \{0,1\}$

Semi-supervised Learning

- Graph $G = (V, E)$
- V: n instances;
- E: two instances are connected if they are similar with each other;
- Weight: represent the similarity between two instances.

Semi-supervised Learning

- Radial Basis Function (RBF)

$$w_{ij} = \exp(-\frac{1}{\sigma}\sum_{d=1}^{m}(x_{id} - x_{jd})^2), x \in R^m$$

Semi-supervised Learning

- Harmonic function
  - W: weighted matrix
  - D: $D_{ii} = \sum_{j=1}^{n} w_{ij}$
  - Laplacian matrix of a Graph:

$$L = D - W$$

$$L = \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix}$$

Semi-supervised Learning

$$f = \begin{bmatrix} f_l \\ f_u \end{bmatrix}$$ is the label of all the instances,

the solution would be:

$$f_l = y_L$$

$$f_u = -L_{uu}^{-1} L_{ul} y_L$$

Minimize the Weighted Entropy

- Our goal:
  - Query less irrelevant instances
  - Unbalance situation
  - Be more helpful for SSL

Minimize the Weighted Entropy

- Entropy

$$H^*(p) = \sum_{i=1}^{n} \sum_{y_i=0,1,2} p^*(y_i|L) H(\frac{[sgn(f_i) = y_i]}{n})$$

$$p^*(y_i|L)$$

$$sgn(f_i)$$

$$H(p) = -p\log(p)$$

Minimize the Weighted Entropy

$$p^*(y_i = j|L) \approx (f_i)_j, j = 0, 1, 2,$$

$$(f_i)_0 + (f_i)_1 + (f_i)_2 = 1, i = 1, \ldots, n$$

Minimize the Weighted Entropy

$$p_i^*(y_k = j | \{L, x_k\}) \approx f_i^{+\{x_k, y_k\}}, j = 0, 1, 2$$

$$f_u^{+\{x_k, y_k\}} = f_u + (y_k - f_k) \frac{(L_{uu}^{-1})_k}{(L_{uu}^{-1})_{kk}}$$

Minimize the Weighted Entropy

$$\hat{H}^{+\{k\}}(f) = \sum_{i=1}^{n} \sum_{j=0,1,2} (f_i)_j H_{i,j}^{+\{k\}}$$

$$H_{i,j}^{+\{k\}} = -f_{ij}^{+\{x_k\}} \log(f_{ij}^{+\{x_k\}})$$

Minimize the Weighted Entropy

$$\hat{H}^{+\{k\}}(f) = \sum_{i=1}^{n} \sum_{j=0,1,2} \lambda_j (f_i)_j H_{i,j}^{+\{k\}}$$

$$\lambda_0 = \lambda_1 = 1 - \lambda_2$$

Minimize the Weighted Entropy

Denote $(f_i)_j H_{i,j}^{+\{k\}}$ as $\mathbb{H}_{i,j}^{\{+k\}}$

$\hat{H}^{+\{k\}}(f)$ as $\mathbb{H}^k(f)$

$$\mathbb{H}^k(f) = \sum_{i=1}^{n} ((1-\lambda)(\mathbb{H}_{i,0}^{+\{k\}} + \mathbb{H}_{i,1}^{+\{k\}}) + \lambda \mathbb{H}_{i,2}^{+\{k\}})$$

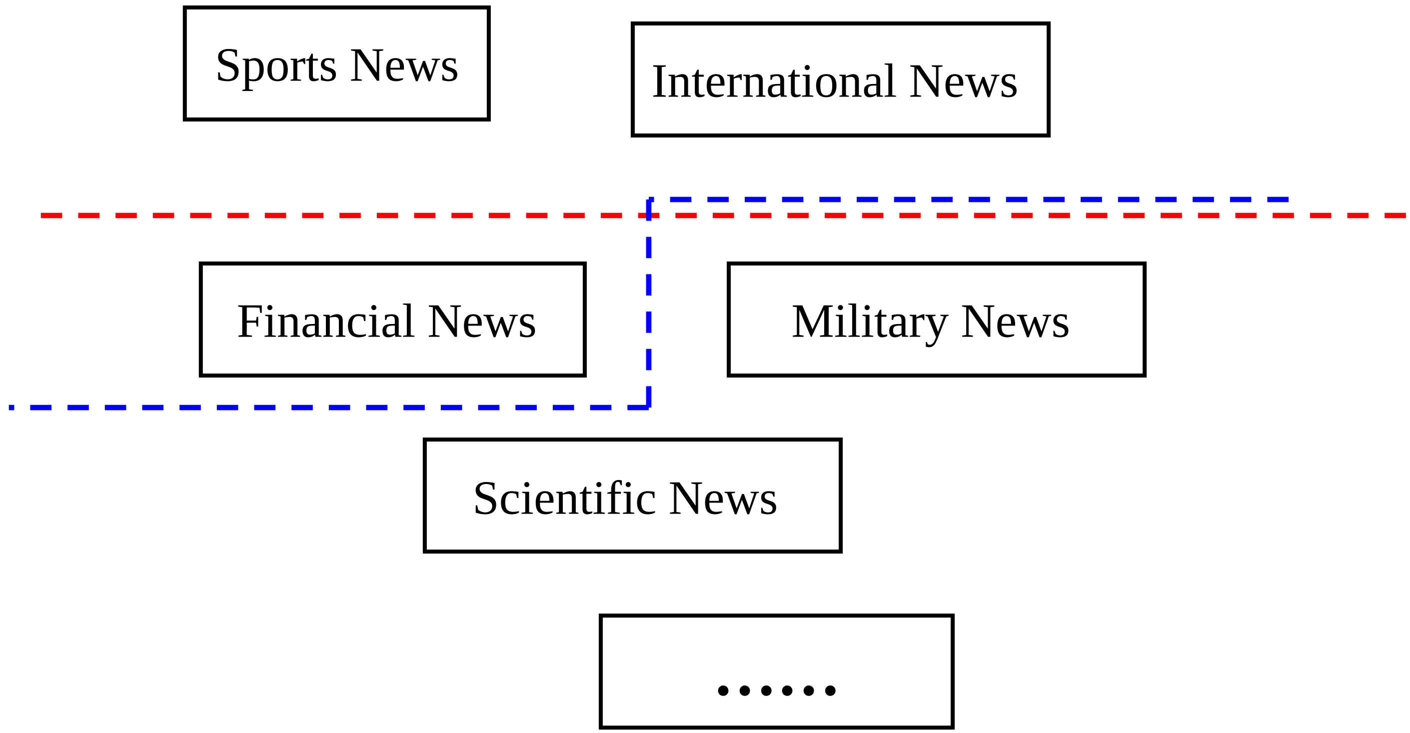
Minimize the Weighted Entropy

$$k = \arg\min_{k'} \mathbb{H}^{k'}(f)$$

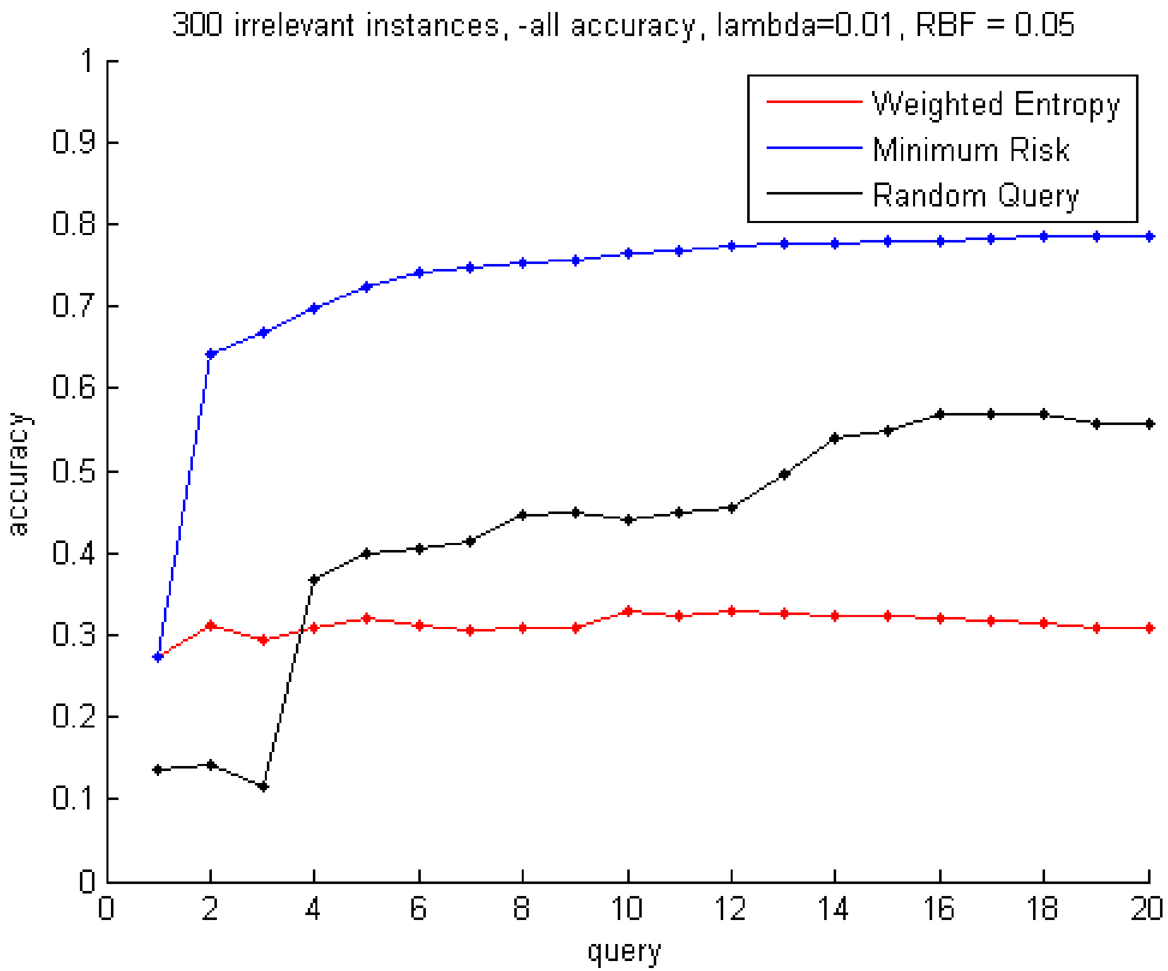
Application

- Personalized new recommendation
  - Initialize as user’s preference
  - Query relevant news
  - More precise classification

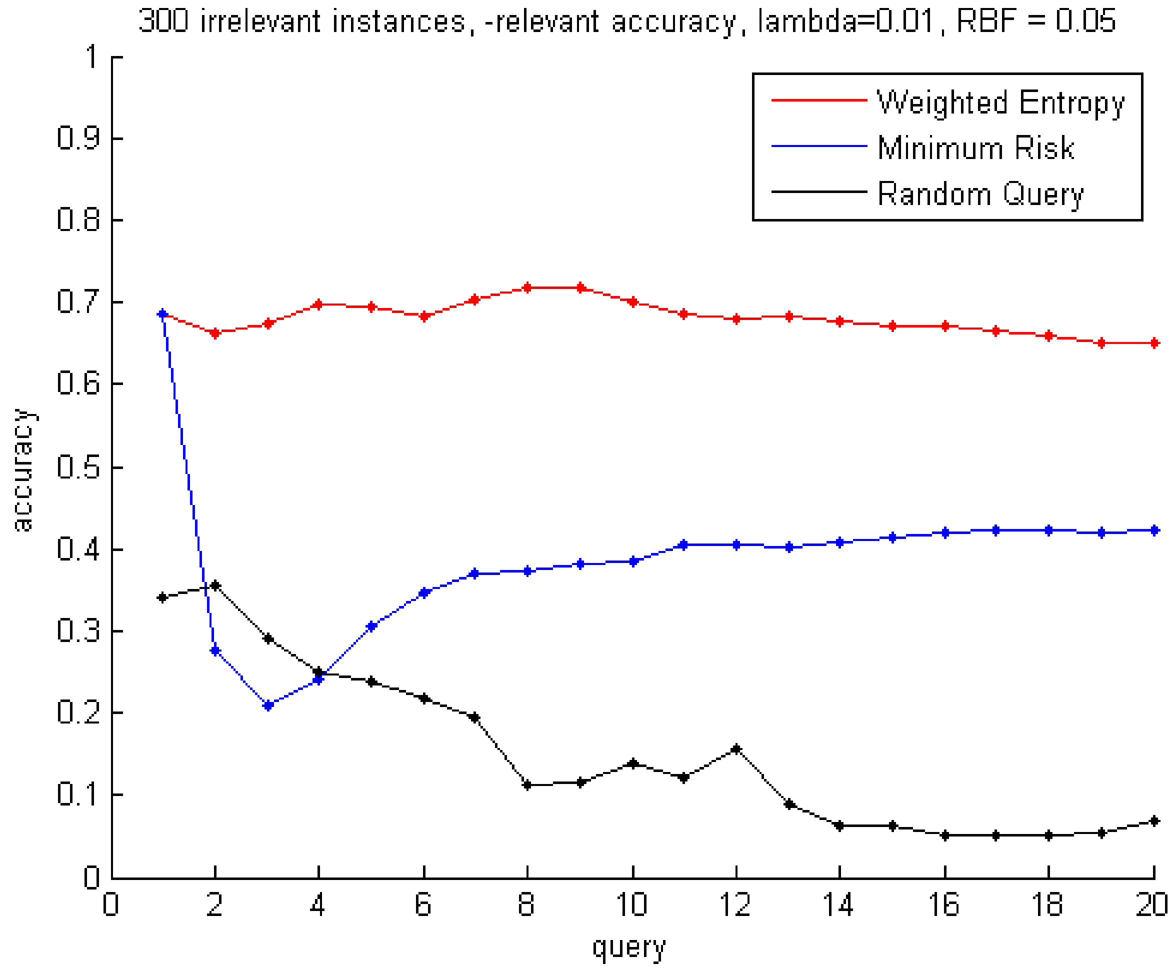
Example


Sports News
International News
Financial News
Military News
Scientific News
……

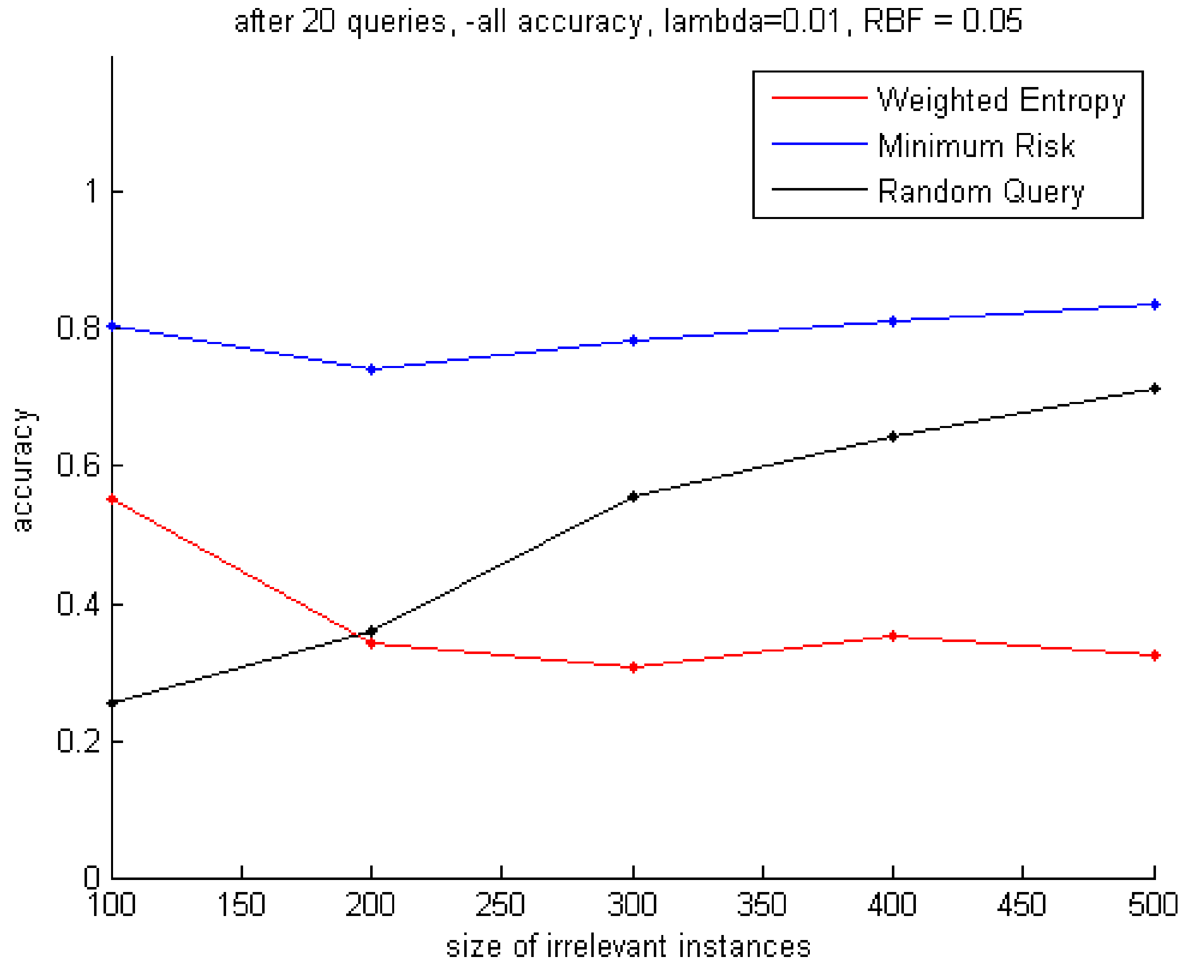
Experiment Result



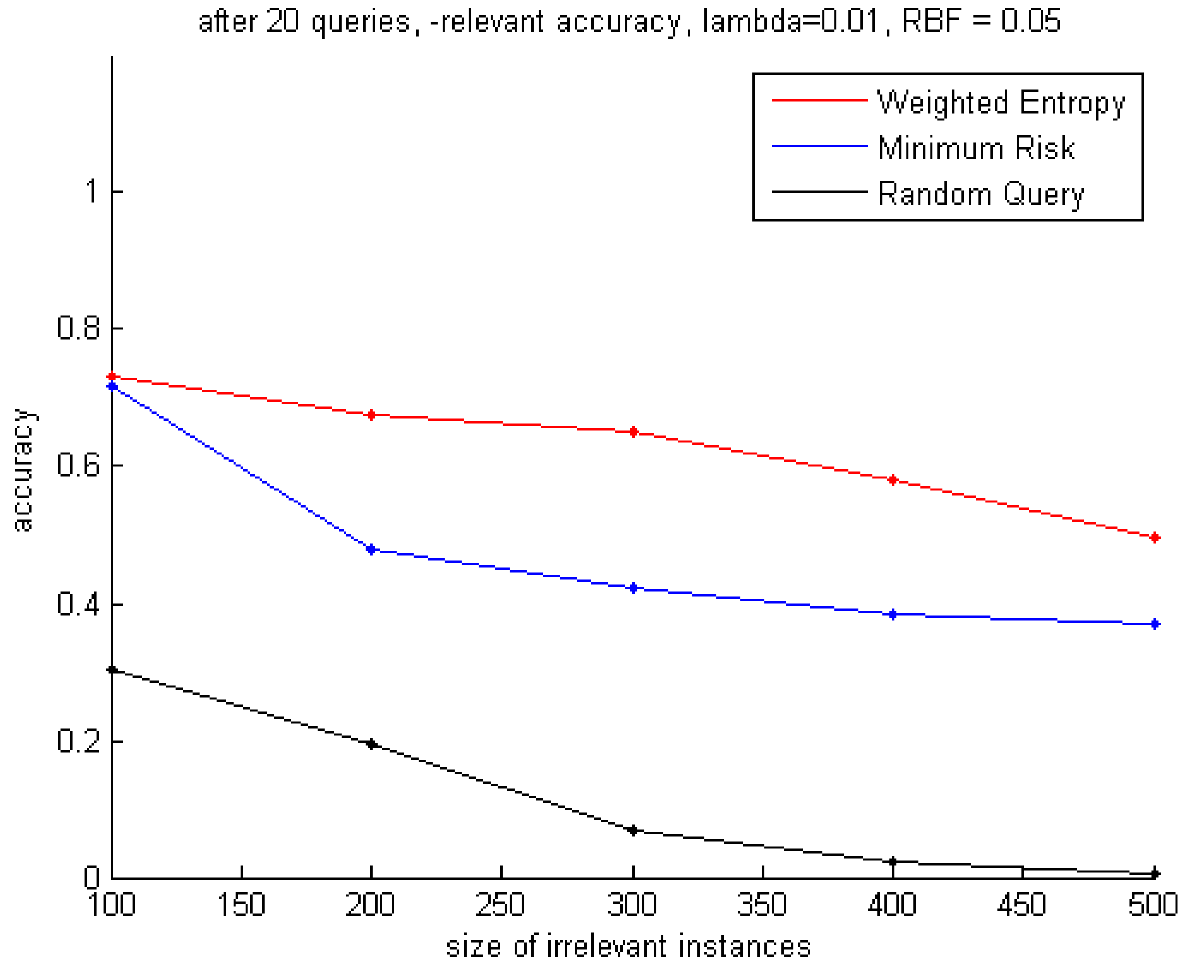
Experiment Result



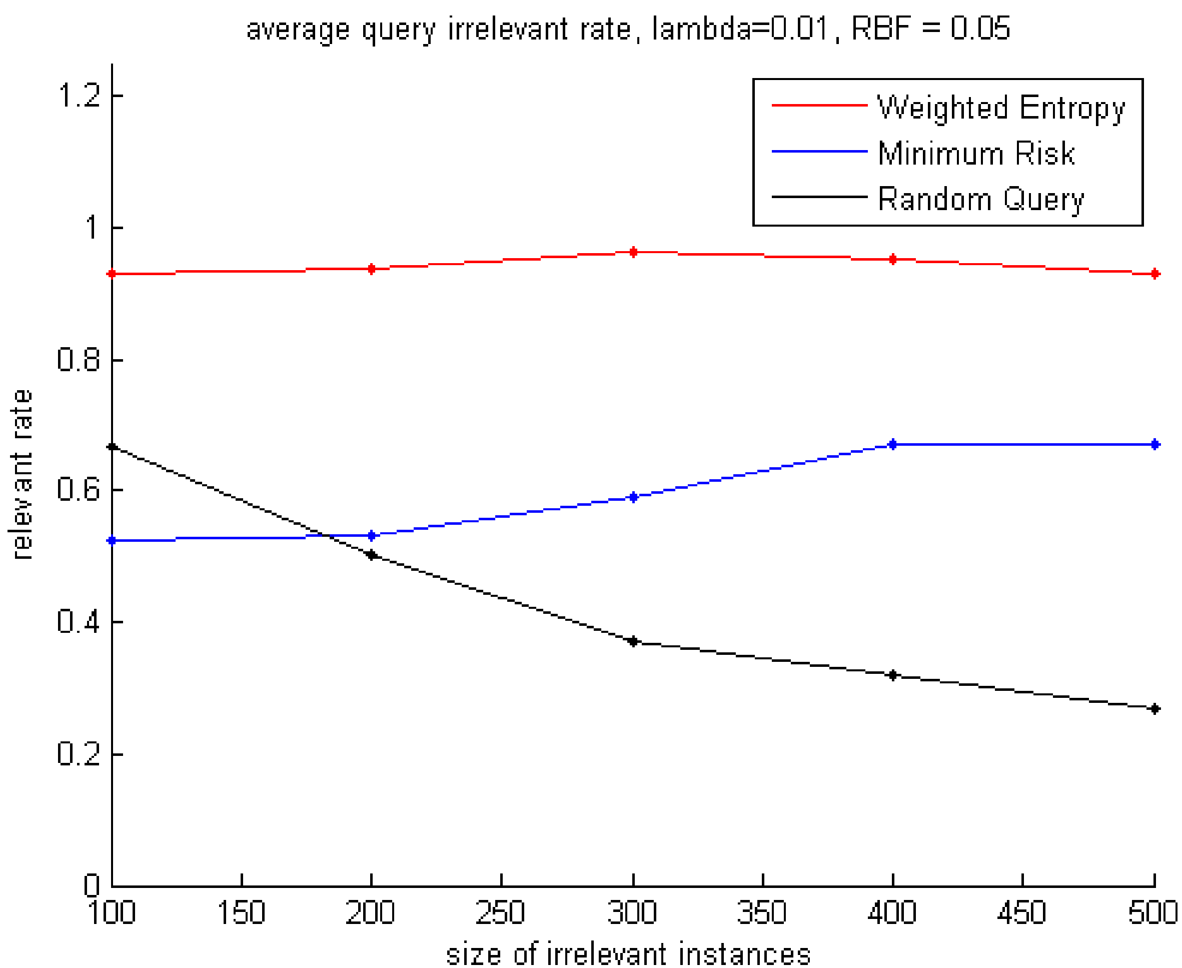
Experiment Result



Experiment Result



Experiment Result



Experiment Result

- To be continue…
  - adding a threshold to filter all irrelevant data
  - more data set

Thanks