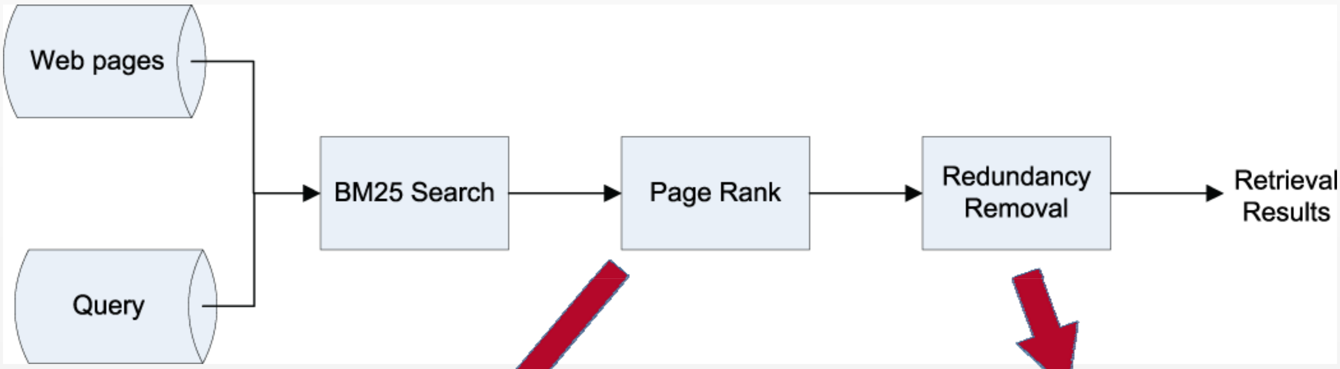
Redundancy Removal to Selectively Diversify Information Retrieval Results

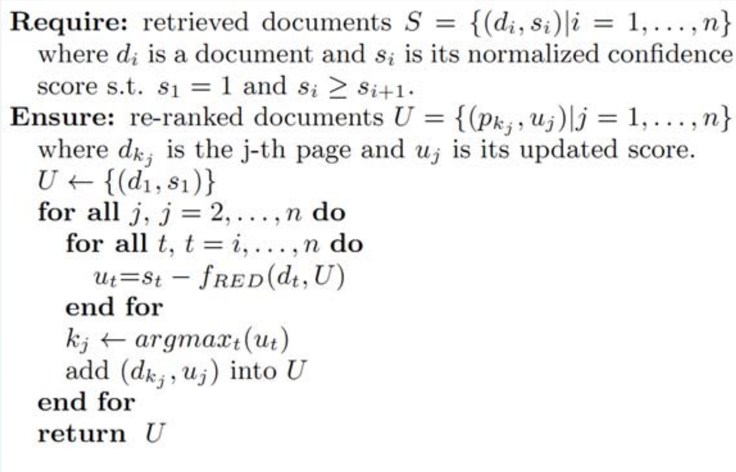
Xiaolin Wang, Hai Zhao, Baoliang Lu
BCMI Lab, Shanghai Jiao Tong University
arthur.xl.wang@gmail.com {zhaohai; blu}@cs.sjtu.edu.cn}

The Brain-like Computing and Machine Intelligence Lab (BCMI) of Shanghai Jiao Tong University takes part in the NTCIR-9 Intent Chinese subtask. A redundancy removal (RedRem) algorithm is proposed to diversify the top-N retrieval results.



$$\mathrm{score}(d,q) = \frac{\mathrm{BM25F}(d,q)}{\max_{d' \in D} \mathrm{BM25F}(d',q)} + \lambda \frac{\mathrm{PR}(d)}{\max_{d' \in D} \mathrm{PR}(d')}$$

where
d is a document
q is a query
BM25F is BM25F similarity score
PR is SogouT-Rank score



```
Require: retrieved documents $S = \{(d_i, s_i) | i = 1, \ldots, n\}$
  where $d_i$ is a document and $s_i$ is its normalized confidence
  score s.t. $s_1 = 1$ and $s_i \geq s_{i+1}$.
Ensure: re-ranked documents $U = \{(p_{k_j}, u_j) | j = 1, \ldots, n\}$
  where $d_{k_j}$ is the j-th page and $u_j$ is its updated score.
  $U \leftarrow \{(d_1, s_1)\}$
  for all $j$, $j = 2, \ldots, n$ do
    for all $t$, $t = i, \ldots, n$ do
      $u_t = s_t - f_{RED}(d_t, U)$
    end for
    $k_j \leftarrow argmax_t(u_t)$
    add $(d_{k_j}, u_j)$ into $U$
  end for
  return $U$
```

where $$f_{\mathrm{RED}}(d,U) = \alpha \frac{|d \cap U|}{|d|} + \beta \frac{\{w \mid w \in d, w \notin U\}}{|d|}$$

Runs	Page retrieval	Page rank?(λ)	RedRem? (α,β)	I-rec@10	D-nDCG@10	D#-nDCG@10
SJTUBCMI-D-C-1	BM25F	Yes(0.4)	Yes(0.1,-0.9)	0.6038	0.2654	0.4346
SJTUBCMI-D-C-2	BM25F	Yes(0.4)	No	0.6008	**0.3317**	**0.4663**
SJTUBCMI-D-C-3	BM25F	No	No	0.5856	0.3288	0.4572
SJTUBCMI-D-C-4	BM25	Yes(0.4)	Yes(0.1, -0.9)	0.6108	0.2756	0.4432
SJTUBCMI-D-C-5	BM25	No	Yes(0.0,-0.9)	**0.6228**	0.2816	0.4522