(1) The incorporation of language-model-based framework resource descriptors allows us to calculate the similarity between collections or collections and queries based on more solid theoretical ground.

(2) topology reorganization protocol pulling semantically similar nodes together to form loose content clusters in distributed manner. We then analyze the impact of topological factors on IR efficiency.

(3) Two efficient query routing algorithms which take advantage of the language model and the underlying topology.

We evaluated these algorithms on TREC 100 and TREC VLC (Very Large Collection 1) respectively [7]. The large collections are split to hundreds of small sub-collections by

t this point, $n_i$ 's routing t ble is reorg nized ccording to the following rules:

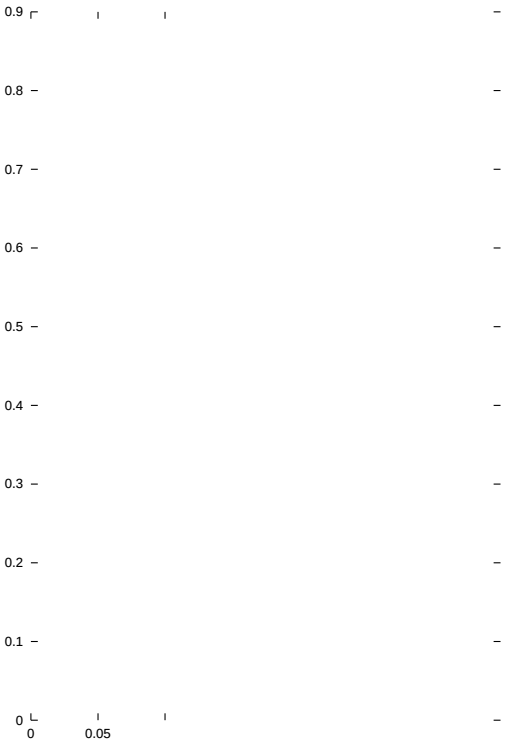
( ) W-% of its degree re design ted s its most simil r neighbors while the rest $(1-$W-$)$% neighbors re r ndomly pic ed from the set $\{n_{j,r_i}\} \bigcup_{n' \in n_{j,r_i}} \{n_{p,r_n}\}$. Choosing neighbors r ndomly is n effort to eep the gr ph well connected. s exp eriments show th t if ll the neighbors re chosen from the most simil r nodes, the result nt networ suffers from b d connectivity. Sp ecific lly, it cont ins m ny sep -

(KL) divergence to me sure the dist nce between collection models or collection models nd query models. The formul is:

$$D(p||q) = \sum_i p(\ ) \log \frac{p(\ )}{q(\ )}$$

Unfortun tely, this formul includes e ch possible word, so is very time-consuming. To speed the process, n pproxim te formul is used: [9]

$$D(p||q) = - \sum_{w \in \ \cap Q} p(\ \ |\theta_Q$$


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.05

0 –

–

re dy chieves l rge incre se in IR erform nce. r dient se rch scheme erforms best with low covered nodes level