(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) A topology reorganization protocol placing 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

At this point, n_i's routing table is reorganized according to the following rules:

( ) % of its degree are designated as its most similar neighbors while the rest $(1-$ $)$% neighbors are randomly picked from the set $\{n_{j,r_i}\} \bigcup_{n' \in n_{j,r_i}} \{n_{p,r_n}\}$. Choosing neighbors randomly is an effort to keep the graph well connected. Our experiments show that if all the neighbors are chosen from the most similar nodes, the resultant network suffers from bad connectivity. Specifically, it contains many deep -

(KL) divergence to measure the distance between collection models or collection models and query models. The formula is:

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

Unfortunately, this formula includes each possible word, so is very time-consuming. To speed the process, an approximate formula is used: [9]

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