Weeks 12–13 Summary: Link Analysis, PageRank, and HITS

Z3003 Data Mining: Scaling to Large Data (Part 2 · Dr. Mahendran V)

1. PageRank Power Iteration, Dead Ends, and Spider Traps

For column-stochastic transition matrix M ($M_{vu} = 1/\deg_{\text{out}}(u)$), random surfing obeys $r^{(t+1)} = \beta M r^{(t)} + \frac{1-\beta}{N}\mathbf{1}$ with typical damping $\beta \approx 0.85$. Dead ends (zero out-degree nodes) are replaced by uniform columns $1/N$, and random teleportation prevents spider traps from absorbing all probability mass.

2. Topic-Sensitive PageRank and HITS

Restricting the teleportation vector to a topic seed set S produces personalized authority scores. Kleinberg's HITS algorithm iteratively updates authority weights $a = L^{\top}h$ and hub weights $h = La$ (normalized after each step) to rank mutually reinforcing hub-and-authority pages.