11741 Reading Summary, Ch 21

Xin Qian (xinq@cs.cmu.edu)

Web search places great emphasis on link analysis. Link analysis for web search is the intellectual decedent of the field of citation analysis. A simple measurement of a web page's quality by the number of in-links is not robust to link spam phenomenon since people might set up several web pages pointing to a target page to boost its in-link count. Link analysis is useful in crawl suggestion.

Treating the web as a graph underlies two intuitions: first, the anchor text is a good page descriptor; second, a hyperlink from A to B is an endorsement by author of page A to page B. Anchor text bridges the gap between the original content/terms in a web page and the external description of the page. Current web search engines assign considerable weights to anchor text terms. However, there's sometimes orchestrated anchor text as a form of spam. Extended anchor text is often useful as the original anchor text, depending on the window width.

PageRank give every node in the web graph a numerical score from 0 to 1. The PageRank score can be combined as a feature into web search. PageRank is modelled as a random walker who visits some nodes more often than other, indicating these are important pages. Teleport operation solves the problem when a node has no out-links. A Markov chain consists of N states, corresponding to each web page. N by N transition probability matrix has each entry as a transition probability that depends only on the current state. Ergodic Markov chain means for all pairs of state, starting from state i at time 0, after a certain time T_0, the probability of being in the state j is greater than 0. The random walk along with teleporting generates a unique distribution of steady-state probabilities over the states. PageRank is a static measure of web page quality. Topic-specific PageRank teleports to a random web page non-uniformly, only a subset of web pages over which the random walk has a steady-state distribution. Personalized PageRank assumes that individual interest can be modelled as a linear combination of topic page distributions. Personalized PageRank vector can be constructed linearly from its underlying topic-specific PageRanks.

Dividing web pages into hub pages and authority pages, we might want to use the hub pages to discover the authority pages. We can iteratively compute the hub score and the authority score for every web page in the subset of the web containing good hub and authority pages. This is called HITS, the Hyperlink-Induced Topic Search. Sometimes good authority pages might not contain a specific query term. We include the root set of pages and the base set of pages to compile an adequate subset of the Web. Cross-language retrieval phenomenon are observed. Examples on the query japan elementary schools shows this phenomenon.