3 HITS and Related Algorithms

Hyperlink Induced Topic Search (HITS) is a representative of algorithms that reveals Web page relationships conveyed by hyperlinks. It aroused investigation on constructing Web page community from hyperlink information. As the beginning of Web community discussion, this chapter discusses this hyperlink-based algorithm, its improvements, variations, and related issues. Some in-depth analyses of HITS are presented as well. Sect. 3.1 gives the original algorithm of HITS. The stability issues of HITS are discussed in Sect. 3.2, which is the basis of further discussions in this chapter. Randomized, subspace and weighted HITS are discussed respectively in Sect. 3.3, 3.4 and 3.5. In Sect. 3.6 and 3.7, two algorithms are discussed, which incorporate page content information to improve HITS. Before discussing other HITS related algorithms, Sect. 3.8 gives an in-depth analysis of HITS from matrix analysis point of view to reveal some features of HITS. After that, Sect. 3.9 discusses a special case of "nullification" in HITS, and gives another approach to avoid this abnormality and improve HITS accordingly. Sect. 3.10 gives another way to improve HITS by eliminating noise pages from the page base set of HITS, rather than directly tuning the HITS. In the last section of this chapter Sect. 3.11, a stochastic approach SALSA is discussed to improve HITS.

3.1 Original HITS

The Hyperlink Induced Topic Search (HITS) algorithm is essentially a link-based approach that intends to find authority and hub pages from a link induced Web graph. Authorities are those pages that provide the best source of information on a given topic, while hubs are those pages that provide collections of links to authorities (Kleinberg 1998). Hubs and authorities exhibit a mutually reinforcing relationship: A good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs. The algorithm was implemented by IBM Almaden Research Centre in Clever, a prototype search engine, and was named the "Clever Algorithm" in an article of *Scientific American*. This algorithm name is also used in some research papers.