Chapter 18
Review of Link Structure Based Ranking Algorithms and Hanging Pages

Ravi P. Kumar  
Curtin University, Australia

Ashutosh K. Singh  
National Institute of Technology, India

Anand Mohan  
National Institute of Technology, India

ABSTRACT

In this era of Web computing, Cyber Security is very important as more and more data is moving into the Web. Some data are confidential and important. There are many threats for the data in the Web. Some of the basic threats can be addressed by designing the Web sites properly using Search Engine Optimization techniques. One such threat is the hanging page which gives room for link spamming. This chapter addresses the issues caused by hanging pages in Web computing. This Chapter has four important objectives. They are 1) Compare and review the different types of link structure based ranking algorithms in ranking Web pages. PageRank is used as the base algorithm throughout this Chapter. 2) Study on hanging pages, explore the effects of hanging pages in Web security and compare the existing methods to handle hanging pages. 3) Study on Link spam and explore the effect of hanging pages in link spam contribution and 4) Study on Search Engine Optimization (SEO) / Web Site Optimization (WSO) and explore the effect of hanging pages in Search Engine Optimization (SEO).

INTRODUCTION

Hanging pages can be compromised by spammers to induce link spam in the hyper link structure of Web and can be a threat to Web site security. By identifying hanging pages and handling hanging pages reduces the threat in Cyber Security. With the rapid growth of WWW and the users’ demand for knowl-

DOI: 10.4018/978-1-5225-0105-3.ch018
edge, it has become more difficult to manage information on the WWW and satisfy user needs. Users are looking for better IR techniques and tools to locate, filter and extract the necessary information. Most of them use IR tools like search engines to find information from the WWW. Generally, many Web users do not see beyond the top few pages of the search results (Broder, 2002; Jansen, Spink, Bateman & Saracevic, 1998; Silverstein, Marais, Henzinger & Moricz, 1999). Therefore, search engines need to produce the relevant results within the top few pages, or they will decline in popularity. Web users are not only looking for relevant information also but also for authoritative sources, i.e. trusted sources of correct and authentic information, like getting the information direct from the home page of a company (Borodin, Roberts, Rosenthal & Tsaparas, 2005). Hence, in current Web searches, there is a shift from relevance to authoritativeness, and the main task of the search engine ranking algorithms have also shifted to finding and ranking the more authoritative Web documents.

With the above shift from relevancy to authoritativeness, the link structure of the Web plays a very important role in sourcing for authoritative documents. Through the hyperlink structure, the Web offers a rich context of information. Here, a link from page a to b denotes an endorsement for the quality of page b. Therefore, the Web can be imagined as a network of recommendations which contains information about the authoritativeness of the pages. Based on this concept, Kleinberg (1999a) and Brin and Page (1998) introduced the HITS and PageRank link analysis algorithms, where hyperlink structures are used to rank Web pages.

Search engines and their ranking algorithms play a very important role in extracting relevant information from the World Wide Web (WWW); however, searching for relevant information in this huge Web is a challenging task due to the non-standard structure of the Web, complex styles of different Web data, the exponential growth, dynamic nature of the Web and the unfair treatment of relevant hanging pages by the search engines.

This chapter focuses on one of the problems in extracting relevant information from WWW, i.e. problems with hanging pages and retrieving the relevant hanging pages from WWW (Kumar, 2014). This chapter starts with comparing different types of link structure based ranking algorithms followed by introducing the hanging pages, explore the effects of hanging pages in ranking Web pages and compare the existing methods to handle hanging pages. Next link spam is introduced and the effect of hanging pages in link spam contribution is described in detail. Finally Search Engine Optimization is introduced, challenges of SEO, SEO optimization techniques and the effect of hanging pages in SEO are described in detail with examples and experiments.

**BACKGROUND**

Hanging page is a page that does not have any forward or outgoing links in the Web. Hanging page can be also called dangling page, zero-out-link page, dead end page, sink page etc. For uniformity and consistency purpose, the term ‘hanging page’ has been used throughout this chapter. Most of the ranking algorithms used by search engines just ignore the hanging pages. However, hanging pages cannot be just ignored because they may have relevant and useful information like pdf, ppt, video and other attachment files. This chapter exposes the various problems associated with hanging pages in link structure based ranking algorithms. This chapter also discusses some former methods for dealing with hanging pages. Hanging pages are one of the hidden problems in link structure based ranking algorithms because: