The Ever-changing Labyrinth: A Large-scale Analysis of Wildcard DNS Powered Blackhat SEO

The authors in this paper investigated a new type of Blackhat Engine Optimisation (SEO), namely *spider pool*. The *spider pool* utilizes cheap domains with low RP to construct fully connected link-networks and poison long-tail keywords. Compared with the other SEO infrastructures, spider pools is much more flexible as well as cheaper. In order to build a link network that trap a search-engine crawler, the owners of spider pools abuse wildcard DNS to create virtually infinite sites and construct complicated loop structure to force the search-engine crawler to visit them relentlessly.

The authors firstly infiltrated one popular spider pool called *super spider pool* (in short for SSP) to better understand the bussiness model, features and operational details of a spider pool. They found four features of SSP, namely (1) wildcard DNS usage, (2) content generation, (3) link structure and (4) site free-ride. Then, they developed a classifiler to differentiate the types of domains in the sitemap automatically. After analysis of SSP, the authors implemented a scanner, aiming to discover and measure SEO domains of spider pools from an internet-wide view. To identify a SEO domain, the authors used a crawler to visit the sitemap of the root folder twice and the domain showing different set of hyperlinks is considered the SEO domain.

I am quite confused with the details of modeling and building the classifer. Especially, the author did not tell what kind of classifier they used and how they built the classifier.