**Spider Trap**

**Spider trap (or crawler trap)** là một tập hợp các trang web có thể cố ý hoặc vô ý được sử dụng để khiến trình thu thập thông tin web hoặc bot tìm kiếm thực hiện vô số yêu cầu hoặc khiến trình thu thập thông tin có cấu trúc kém gặp sự cố. Trình thu thập dữ liệu web còn được gọi là trình thu thập dữ liệu web, từ đó tên này được bắt nguồn. Bẫy nhện có thể được tạo ra để "bắt" spam bots hoặc các trình thu thập thông tin khác làm lãng phí băng thông của trang web. Chúng cũng có thể được tạo ra một cách vô ý bởi các lịch sử dụng các trang động với các liên kết liên tục trỏ đến ngày hoặc năm tiếp theo.

**Các kỹ thuật phổ biến được sử dụng :**  
Tạo cấu trúc thư mục sâu vô hạn định như: http: //example.com/bar/foo/bar/foo/bar/foo/bar / ...   
Các trang động tạo ra số lượng tài liệu không giới hạn để trình thu thập thông tin web theo dõi. Ví dụ bao gồm lịch và thơ ngôn ngữ được tạo theo thuật toán. Tài liệu chứa nhiều ký tự, làm hỏng trình phân tích từ vựng phân tích cú pháp tài liệu. tài liệu có id phiên dựa trên cookie được yêu cầu. Không có thuật toán để phát hiện tất cả các bẫy nhện. Một số loại bẫy có thể được phát hiện tự động, nhưng các bẫy mới, chưa được phát hiện sẽ nhanh chóng phát sinh.

Bẫy nhện khiến trình thu thập thông tin web đi vào một cái gì đó giống như một vòng lặp vô hạn, làm lãng phí tài nguyên của trình thu thập dữ liệu, làm giảm năng suất của nó và, trong trường hợp trình thu thập thông tin viết kém, có thể làm hỏng chương trình. Trình thu thập thông tin lịch sự luân phiên yêu cầu giữa các máy chủ khác nhau và không yêu cầu tài liệu từ cùng một máy chủ nhiều hơn một lần trong vài giây, có nghĩa là trình thu thập thông tin web "lịch sự" bị ảnh hưởng ở mức độ thấp hơn nhiều so với trình thu thập thông tin "bất lịch sự".

Ngoài ra, các trang web có bẫy nhện thường có tệp robots.txt thông báo cho các bot không được vào bẫy, do đó, một bot "lịch sự" hợp pháp sẽ không rơi vào bẫy, trong khi một bot "bất lịch sự" không quan tâm đến cài đặt robots.txt sẽ bị ảnh hưởng bởi cái bẫy.

**Làm thế nào để bạn giải quyết một cái bẫy nhện?**

Giải quyết một cái bẫy nhện có thể là một điều khó khăn. Trong trường hợp của chúng ta, chúng ta hoàn toàn không muốn / đề xuất / được lập chỉ mục, vì vậy chúng ta đã chặn hoàn toàn bằng robots.txt. Trong những trường hợp khác, bạn không thể làm điều đó một cách dễ dàng. Đối với điều hướng theo từng khía cạnh, bạn phải suy nghĩ rất lâu và khó xem bạn muốn Google thu thập thông tin và lập chỉ mục các khía cạnh nào.

Nói chung, có ba loại giải pháp:   
1. Chặn (một phần của) các URL bằng robots.txt.   
2. Thêm rel = nofollow và noindex, theo dõi trên các tập hợp con cụ thể của các liên kết và trang và sử dụng rel = canonical một cách khôn ngoan.   
3. Khắc phục cái bẫy bằng cách không tạo ra vô số URL nữa.