RedirectHandling

Redirect handling in Nutch

This page is in construction but when completed will provide a comprehensive overview of redirect handling in Apache Nutch.

To begin with, we really want to define what HTTP URL redirects are, what types of problems they present for crawlers, and finally what Nutch does to address some of these problems. By the end of this tutorial, we should have addressed the complex and rather confusing area of redirects. For a whirlwind tour of this page please see the Table of Contents below.

Introduction

URL redirects as they are most commonly known (and hereby referred to in this document), in a high level sense play the role of temporarily or permanently redirecting an HTTP response recipient to a location other than the request URI. By doing this, it is possible to *easily* direct browsers, web crawlers, and subsequently users to your preferred domain (well this is true in theory anyway).

Some typical reasons for implementing URL redirects: (all courtesy of wikipedia)

- Similar domain names
- Moving a site to a new domain
- Logging outgoing links
- Short aliases for long URLs
- Meaningful, persistent aliases for long or changing URLs
- Manipulating search engines
- Satire and criticism
- Manipulating visitors
- Removing referer information

Types of URL Redirects

Problem Identification

Nutch & URL Redirects

Conclusion

URL Redirect FAQ's