Http 완벽가이드 #9,10(~p294)
웹로봇, Http 2.0

2023.10.14

1. 다음 설명 중 틀린 것은?

1. 좋은 루트 집합은 크고 인기있는 웹사이트, 새로 생성된 페이지목록, 자주 링크되는 잘 알려진 페이지 목록으로 구성되어 있다.
2. 복잡한 로봇은 방문 URL을 추적하기 위해 검색트리나 해시 테이블을 사용했을수 있다.
3. URL개수는 잠재적으로 무한하고, 존재비트배열은 유한개수의 비트만 존재해, 존재비트에 두 URL 매핑으로 충돌될 가능성이 존재한다.
4. 동적 콘텐츠로 인해 발생할 수 있는 무한 루프 문제 때문에, 많은 로봇이 URL 어딘가에 'cgi'라는 문자열을 포함한 사이트 크롤링을 거부한다.
5. 웹사이트들 전체에 걸쳐 너비 우선으로 스케줄링 하면, 순환의 영향을 최소화 할 수 있다.

정답 : 1
p249 자주 링크되지 않는 잘 알려져 있지 않은 페이지 목록으로 구성되어 있다.

2. 다음 설명 중 틀린 것은?

1. 웹 로봇이 루프와 중복을 피하기 위해 URL정규화, 너비우선 크롤링, 스로틀링, URL 크기제한을 할 수 있다.
2. 로봇개발자들에게 구현이 권장되는 식별 헤더에는 User-agent, From, Accept, Referer이 있다.
3. 가상 호스팅 문제로 Http/1.1은 Host 헤더를 사용할 것을 요구한다.
4. 모든 로봇은 200 OK 나 502 Bad Gateway 와 같은 HTTP 상태코드를 이해해야 한다.
5. <meta http-equiv="Refresh" content="1; URL=index.html"> 태그는 수신자가 문서를 마치 HTTP응 답 값이 "1; URL=index.html"인 Refresh HTTP헤더를 포함하고 있는 것처럼 다루게 한다.

정답 : 4
4. 404Not Found (p 262)
5. 무슨 말인가요? ㅠ

3. 다음은 무엇에 대한 설명인가?

이 방법은 웹 로봇이 중복을 감지하는 방법보다 직접적이다. 이것을 사용하는 로봇은 페이지의 콘텐츠에서 몇 바이트를 얻어내 체크섬(checksum)계산한다. 이것 생성용으로 MD5와 같은 메세지 요약함수가 있다.

정답 : 콘텐츠 지문
p258

4. 다음 설명 중 맞는 것은?

1. 로봇의 접근을 제어하는 정보를 저장하는 파일 이름은 robot.txt이다.
2. 로봇 차단표준은 RSI 위원회의 정책으로 마련되었다.
3. 사이트 전체에 대한 robot파일은 단 하나만 존재하며, 가상호스팅 되면 docroot에 각각 파일이 있다.
4. 서버가 404 Not Found HTTP코드로 응답하면 로봇의 접근을 제한한다는 뜻이다.
5. HTTP 503코드를 본 로봇은 리소스를 발견할때까지 리다이렉트를 따라간다.

정답 : 3
1. robots.txt
2. 임시방편으로 마련된 표준이다. (p266)
4. 로봇의 접근을 제한하지 않는 것으로 간주하고 어떤 파일이든 요청하게 된다.(p267)
5. 리소스 검색을 뒤로 미룬다. (p268)

5. 다음 설명 중 맞는 것은?

1. robots.txt 파일의 각줄은 포맷줄, 주석줄, 규칙줄 세 종류가 있다.
2. 특정 디렉토리만 크롤링 하지 못하게 하는 경우 robots.txt는 이를 표현할 수단을 제공하지 않는다.
3. robots.txt 파일의 장점 중 하나는 파일을 콘텐츠 작성자 개개인이 아니라 웹사이트 관리자가 소유하는 것이다.
4. 가장 널리 쓰이는 로봇 메타 지시자 두가지는 All, None이다.
5. NoARCHIVE는 로봇에게 캐시를 위한 원격 사본을 만들어서는 안된다고 말해준다.

정답 :
1. 포맷줄 - > 빈줄 (p268)
3. 단점
4. NoINDEX, NoFOLLOW
5. 원격사본이 아니라 로컬사본

6. 다음 설명 중 틀린 것은?

1. 2012년 SPDY를 기반으로 HTTP/2.0 프로토콜 설계가 결정되었다.
2. HTTP/2.0은 서버의 요청을 받지 않아도 클라이언트가 필요하다고 생각되는 리소스를 능동적으로 보내줄수 있다.
3. :content-length, :status 는 변경된 HTTP/2.0의 문법이다.
4. HTTP/2.0에서 모든 메세지는 프레임에 담겨 전송된다.
5. 멀티플레싱은 HTTP/2.0 커넥션을 통해 클라이언트와 서버 사이에 교환되는 프레임들의 독립된 양방향 시퀀스이다.

정답 : 5
스트림에 대한 설명이다.

7. 다음 설명 중 맞는 것은?

1. HTTP/2.0 에서는 하나의 커넥션에 여러 스트림이 동시에 열릴수 없다.
2. 스트림의 우선순위는 의무사항이다.
3. 스트림이 클라이언트에 의해 초기화 되면, 그 식별자는 반드시 짝수, 서버는 홀수여야 한다.
4. HTTP/2.0의 스트림 생성에는 TCP 패킷 생성이 우선한다.
5. 서버의 PUSH_PROMISE 프레임에 대해 클라이언트는 RST_STREAM 으로 거절할 수 있다.

정답 : 5

1. 있다.
2. 의무사항이 아니다.(p291)
3. 짝수 <> 홀수
4. TCP패킷 주고 받을 필요 없이 만들어진다.

생각해 볼것(p251)

웹로봇의 대규모 URL을 위한 적합한 자료구조는? 이유는?