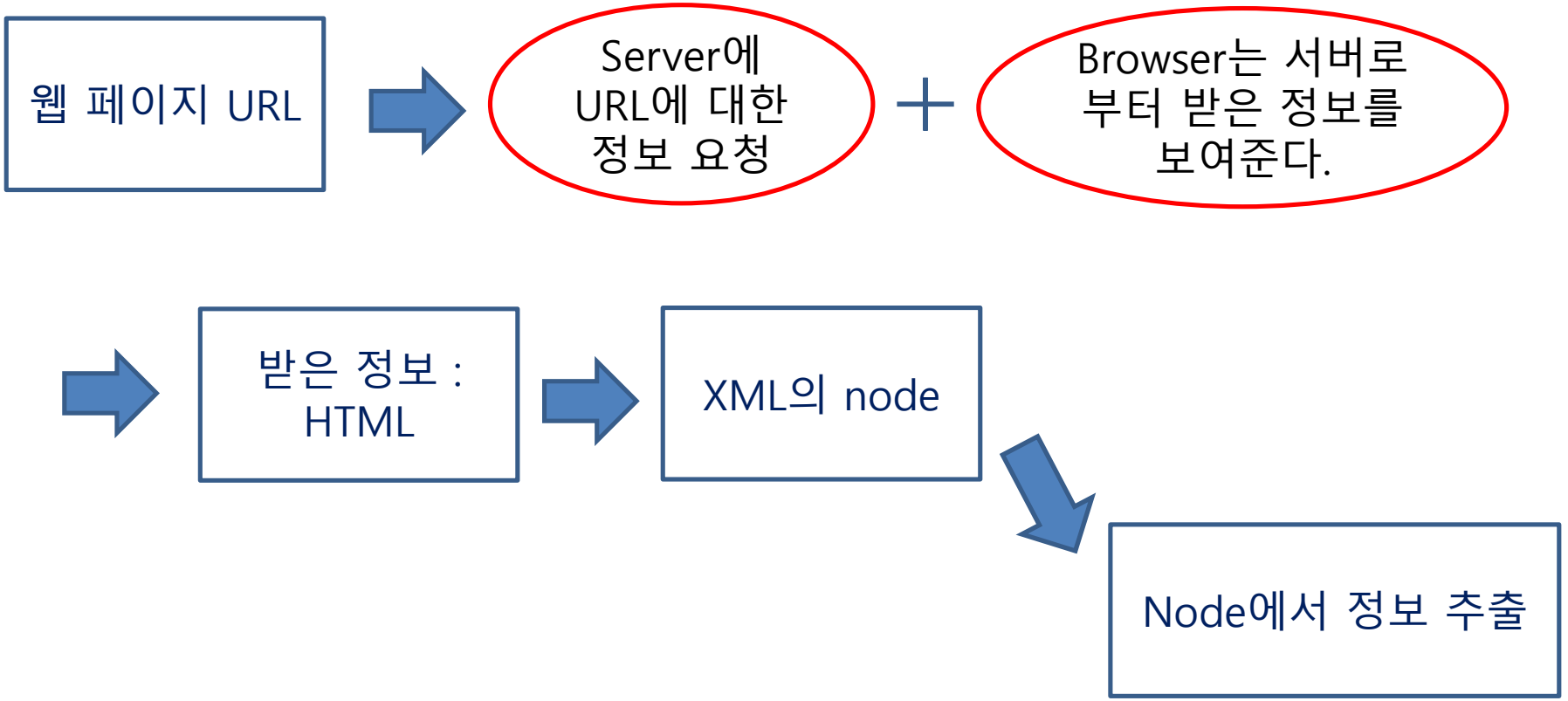
R을 활용한 크롤링

참고 : https://mrchypark.github.io/getWebR

크롤링이란?

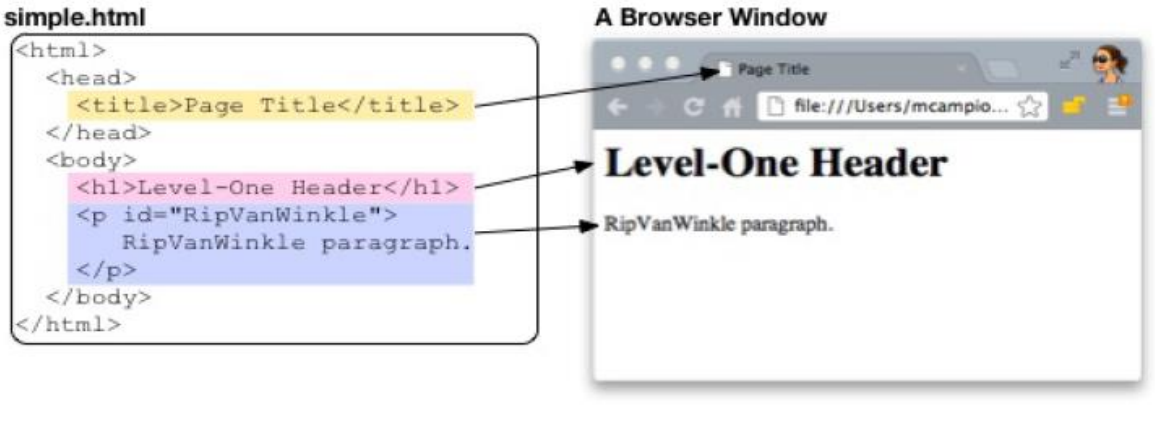
Crawling 혹은 scraping은 웹 페이지를 그대로 가져온 후 가져온 웹 페이지에서 데이터를 추출해내는 행위를 말한다.

앞으로 배울 R 크롤링의 순서



그럼 html 이란

xml양식으로 작성된 웹 브라우저들이 이해하는 표준 문서



HTML

```
<html>
<head>
  <title>제목입니다</title>
<body>
  <h1>H1 텍스트입니다.</h1>
  <p>P 텍스트입니다.</p>
</body>
```

head 또는 body 태그 각각의 의미가 정해져 있고 의미에 맞게 사용하여야 한다.

웹 페이지 표현에 치중

XML

```
<title>제목입니다</title>
<content>내용입니다</content>
<sender>발송자</sender>
```

XML에는 사전 정의 태그가 존재하지 않는다.

사용자가 임의로 태그 이름을 정의할 수 있다.

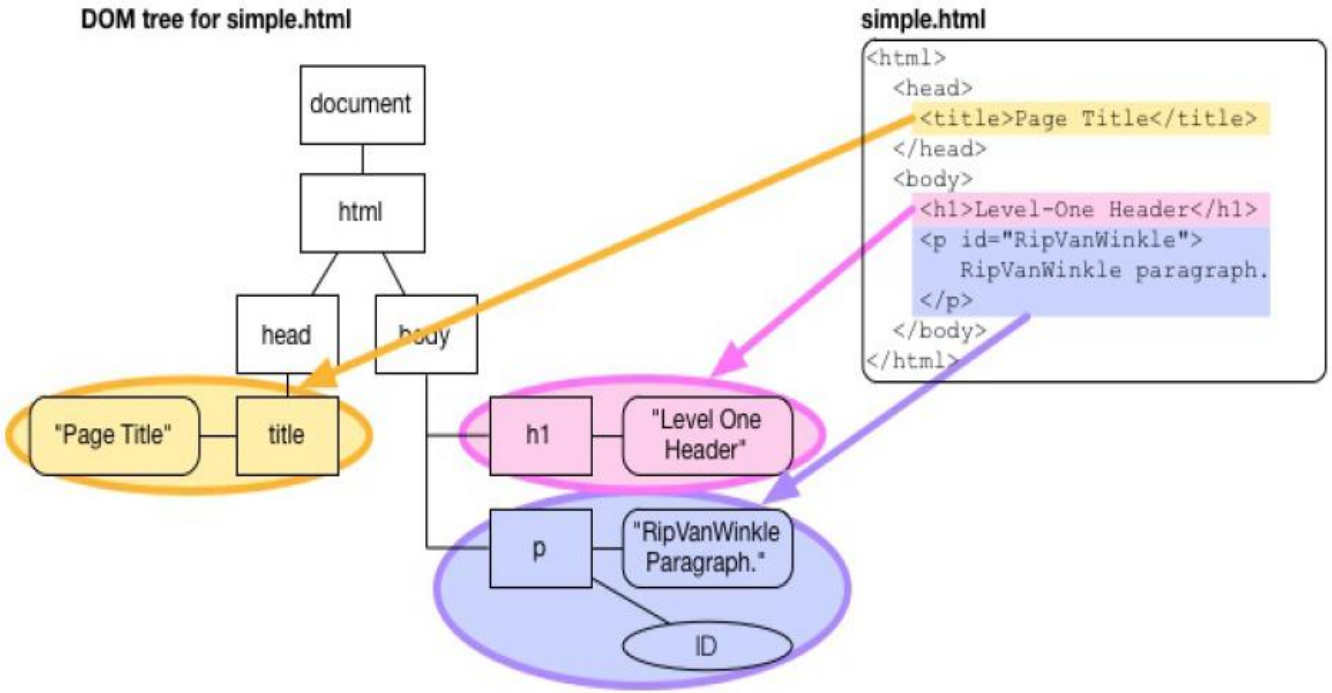
XML의 node를 다루는 패키지 : rvest

rvest

node, attr, text 만 기억하자

node란

html에서 tag라고 불리는 것.



attr 이란

attr은 attribute의 줄임으로 아래 예시로 tag의 attr1은 example1 임

```
<tag attr1="example1" attr2="example2"> 안녕하세요 </tag>
```

text 이란

text은 시작 태그와 종료 태그 사이에 있는 글자로, 아래 예시 기준 "안녕하세요" 를 뜻함

```
<tag attr1="example1" attr2="example2"> 안녕하세요 </tag>
```