

# 웹 개방성 확대를 위한 가이드라인

2013. 4.

한국인터넷전문가협회  
대외협력본부

<b>I. 조사개요</b>	<b>2</b>
1. 조사목적	2
2. 조사개요	2
3. 조사방법	2
<b>II. 세부내용</b>	<b>3</b>
1. 로봇배제표준(robots.txt) 내용 확인	3
1) 로봇배제표준이란?	3
2) 로봇배제표준 사용방법	3
3) 로봇배제표준 확인 방법	4
4) 확인 결과 예시	4
5) 가이드라인	5
2. 메타태그 속성(noindex, nofollow) 확인	5
1) noindex : <meta name="robots" content="noindex">	5
2) nofollow : <meta name="robots" content="nofollow">	5
3) noindex, nofollow 상세설명	6
4) 확인결과 예시	6
5) 가이드라인	7
3. 검색엔진 비 친화적 사이트 확인	8
1) ActiveX 사용	8
2) Image / Flash 위주의 웹 사이트	9
3) 검색엔진 비친화적 웹 사이트 확인방법	10
4) 가이드라인	13
4. User-agent Switcher를 이용한 검색 차단 확인	14
1) 검색로봇으로 가장해 해당 사이트 접근	14
2) User-agent Switcher 설정방법(크롬 기준)	14
3) 차단 사례 확인	17
4) 가이드라인	18
5. URL공개 또는 비공개 여부 확인	18
1) URL 변경여부를 직접 확인하는 방법	18
2) form 태그의 method 속성 중 get과 post 확인	20
3) 가이드라인	21
<b>III. 조사결과</b>	<b>23</b>
1. 조사현황	23
2. 결과분석	23
3. 조사결과 종합	23
4. 조사의 한계	24
5. 조사의 제언	24

# I 조사개요

## 1. 조사목적

- 교육 및 공공기관의 웹사이트 개방성 현황을 평가하여 공공정보 개방에 대한 인식제고 및 실질적 개방성 향상을 목적으로 함
- 대국민 정보 사이트의 개방성지수가 일정 수준을 유지하도록 권고하기 위한 기초 자료로 활용

## 2. 조사개요

조사대상	국내 국·공립 및 사립대학교 100개 정부산하 연구기관 등 공공 정보사이트 100개
조사내용	㉠ robots.txt 파일을 통한 차단 확인 ㉡ noindex / nofollow 태그 확인 ㉢ ActiveX / Image / Flash 사용여부 확인 ㉣ User-agent 기반으로 접근 차단 확인 ㉤ URL 공개 또는 비공개 여부 확인
조사기간	2013년 1월 21일 ~ 2013년 2월 20일(총 1개월)
검수방법	조사를 담당한 2개의 팀이 각자 조사한 결과를 교환해 검수진행

## 3. 조사방법

- ㉠ robots.txt 파일을 통한 차단 확인
  - 각각의 웹 사이트의 홈페이지에서 robots.txt 파일 조사
- ㉡ noindex / nofollow 태그 확인
  - 조사대상 웹 페이지에 접속 후 소스보기를 확인해 아래와 같이 메타 태그에 **content="noindex"**를 설정했는지 확인

```
<meta name="robots" content="noindex">
```

- ㉢ ActiveX / Image / Flash 사용여부 확인
  - 각각의 웹 사이트 메인 및 서브페이지에 접속해 ActiveX / Image / Flash 사용 여부 조사(1개 사이트에서 메인 포함 5개 페이지 조사)
- ㉣ User-agent 기반으로 접근 차단 확인
  - 브라우저의 User-agent Switcher 기능을 이용해 해당 웹 서버가 검색로봇을 차단하는지 여부 확인
- ㉤ URL 공개 또는 비공개 여부 확인
  - 게시판의 게시물 목록과 게시물 보기 페이지를 접속할 때 해당 URL이 변경되는지 확인하고 URL만으로 목록과 특정 게시물을 접근할 수 있는지 확인

## 1. 로봇배제표준<sup>1)</sup>(robots.txt) 내용 확인

### 1) 로봇배제표준이란?

- ① 로봇배제표준은 웹 사이트에 로봇이 접근하는 것을 방지하기 위한 규약으로, 일반적으로 접근 제한에 대한 내용을 robots.txt에 기술한다.
- ② 검색엔진의 정보 수집용 로봇이 웹 서버에 접근하면 수집용 로봇은 서버의 최상위 디렉터리에서 robots.txt 파일을 불러내 문서에 대한 수집 허용 혹은 차단 여부를 확인한다.
- ③ 정보 수집용 로봇은 robots.txt에 기술된 내용을 바탕으로 웹사이트의 페이지를 수집한다. 중요 관리자 폴더 혹은 계정 정보가 들어있는 디렉터리에 대하여 정보 수집용 로봇의 접근을 각각 따로 설정하여 보안을 강화할 수 있다.
- ④ 최상위 디렉터리에 robots.txt 파일이 없다면, 모든 문서에 대하여 검색 로봇의 접근이 허용된다.

### 2) 로봇배제표준 사용방법

- ① 문서 접근 완전 허용 : 검색 로봇이 모든 문서에 접근하도록 허용

User-agent: *	#User-agent에서 *은 모든 로봇을 지칭한다.
Allow: /	#모든 디렉터리에서 로봇의 접근을 허용한다.

- ② 문서 접근 완전 차단 : 검색 로봇이 모든 문서에 접근할 수 없도록 차단

User-agent: *	
Disallow: /	#모든 디렉터리에서 로봇의 접근을 차단한다.

- ③ 문서 접근 부분 차단 : 구글 검색 로봇에 대하여 /admin/, /tmp/ 이하의 모든 페이지에 검색 로봇의 접근 차단

1) 이 규약은 1994년 6월에 처음 만들어졌고, 아직 이 규약에 대한 RFC는 없다. 이 규약은 권고안이며, 로봇이 robots.txt 파일을 읽고 접근을 중지하는 것을 목적으로 한다. 따라서, 접근 방지 설정을 하였다고 해도, 다른 사람들이 그 파일에 접근할 수 있다.(출처 : 위키백과, <http://bit.ly/YB031d>)

```
User-agent: Googlebot
Disallow: /admin/
Disallow: /tmp/
```

### 3) 로봇배제표준 확인 방법

- ① 조사 대상 웹 사이트에 접속합니다. ex)<http://www.sutra.re.kr/>
- ② robots.txt는 웹사이트 서버의 최상위 디렉터리에 존재함으로 브라우저 주소창의 해당 웹사이트의 주소 뒤에 "/robots.txt"를 추가입력하고 엔터키를 누릅니다.
- ③ 페이지에 나타나는 결과에 따라 로봇을 배제하는지 여부를 확인할 수 있습니다.

※ 이 모든 과정을 한 번에 진행하려면 해당 웹 사이트 주소 뒤에 /robots.txt를 복사한 후 엔터키를 누릅니다. ex)<http://www.sutra.re.kr/robots.txt>

### 4) 확인 결과 예시

- ① <http://www.copyright.or.kr/robots.txt> 한국저작권위원회



→ robots.txt가 존재 하지 않음으로 모든 로봇에 대한 모든 문서 접근 허용

- ② <http://www.inje.ac.kr/robots.txt> 인제대학교



→ 모든 로봇에 대하여 /PDG/, /clife/ 이하의 페이지만 접근을 차단하고 나머지는 허용

- ③ <http://suwon.ac.kr/robots.txt> 수원대학교



→ 모든 로봇에 대한 문서 접근 완전 차단

## 5) 가이드라인

- ① 검색 로봇을 차단하는 이유에 대해 회사 내부의 정책이 있는 경우 정책 내용에 따라 적절히 활용
- ② 검색 로봇 차단과 관련해 특별한 정책을 수립하지 않은 경우 보안과 트래픽을 고려해 부분 차단을 적절히 활용하는 방안 검토 필요
- ③ 트래픽과 보안 관련 정책을 우선 수립하고 이에 맞는 적절한 로봇배제 표준을 적용할 경우 콘텐츠에 대한 접근권을 향상시켜 사용자 노출 빈도를 높일 수 있다는 점을 고려해 전체 차단보다는 부분 차단 권고

## 2. 메타태그 속성(noindex, nofollow) 확인

### 1) noindex : <meta name="robots" content="noindex">

- ① 검색엔진이 문서정보를 알 수 있도록 안내하는 역할을 하는 meta tag의 content 속성 중 noindex는 해당 페이지에 대한 색인을 제한하는 역할을 합니다.
- ② 색인은 인덱스 또는 찾아보기라고 정의하기도 하는데 로봇이 색인을 한다는 것은 해당 페이지에 대한 간략한 정보를 요약 저장하는 것을 의미합니다.
- ③ 로봇이 색인을 못하게 되면 해당 페이지에 대한 정보가 저장되지 않기 때문에 검색에서 제외됩니다.

### 2) nofollow : <meta name="robots" content="nofollow">

- ① meta tag의 content 속성에 nofollow를 설정하게 되면 검색로봇이 해당 페이지에 수록된 링크를 따라 갈수 없도록 제한합니다.
- ② nofollow 속성은 페이지 수준의 메타 태그에서 사용되며 페이지의 외부 링크에 대해 추적, 즉 크롤링 하지 않도록 검색엔진에 지시하는 역할을 합니다.
- ③ nofollow 속성을 사용할 경우 검색로봇은 타겟 링크를 삭제하기 때문에 효율적인 검색을 방해할 수 있습니다.

### 3) noindex, nofollow 상세설명

① <meta name="googlebot" content="noindex">

→ Google 로봇을 제외한 다른 모든 로봇이 사이트의 페이지에 대해 색인을 생성하도록 허용합니다.

② <meta name="robots" content="noindex, follow">

→ 해당 페이지의 색인은 차단하고, 해당 페이지에 수록된 링크들을 따라갈 수 있도록 한다.

③ <meta name="robots" content="index, nofollow">

→ 해당 페이지의 색인은 허용하나, 해당 페이지에 수록된 링크들은 따라갈 수 없다.

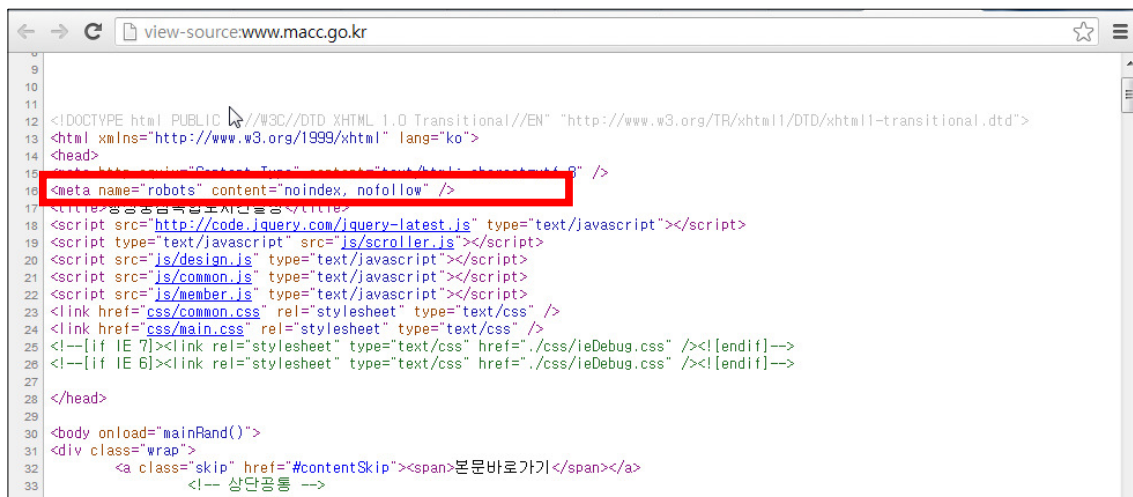
④ <meta name="robots" content="noindex, nofollow">

→ 해당 페이지의 색인을 차단하고, 해당 페이지에 수록된 링크들 또한 따라갈 수 없다.

⑤ noindex, nofollow 태그가 없을 경우 검색로봇은 색인이 가능하고 링크 또한 따라 갈 수 있다.

### 4) 확인결과 예시

① <http://www.macc.go.kr/> 행정중심복합도시건설청



→ 해당 페이지의 색인을 차단하고, 해당 페이지에 수록된 링크를 따라갈 수 없다.

② <http://snu.ac.kr/> 서울대학교



→ 해당 페이지의 색인을 허용하고, 링크들 또한 따라 갈 수 있다.

## 5) 가이드라인

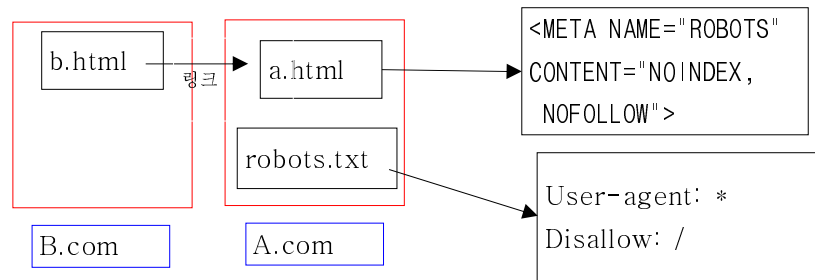
- ① 웹사이트의 소스코드 메타태그의 기능을 정확히 이해하고 적절히 사용해 불필요한 검색제한 방지
- ② 무조건적인 색인 차단과 보안 강화보다는 합리적이고 효율적인 보안정책과 콘텐츠 노출 범위에 대한 기준 마련을 통해 세밀하게 적용할 필요 있음
- ③ 필요한 경우 noindex로 색인은 허용하고 nofollow로 링크만 차단하는 방법으로 콘텐츠의 존재 여부는 공개하고 실제 정보에는 접근하지 못하도록 조절할 수 있음

※ robots.txt와 noindex를 사용했음에도 검색결과에 노출되는 경우

- robots.txt로 검색 수집용 로봇을 차단하고 각 페이지마다 noindex 메타태그를 추가했음에도 불구하고 해당페이지가 검색결과에 노출되는 경우가 있습니다.

- ① 예를 들어 A.com/a.html에 noindex 메타태그가 있고 A.com/robots.txt에서 로봇을 차단합니다.
- ② 검색수집이 완전 허용된 B.com/b.html에 A.com/a.html의 링크가 수록되어 있다면 검색결과에 A.com/a.html이 노출이 됩니다.
- ③ A.com/robots.txt에서 로봇검색을 차단했기 때문에 검색로봇은 A.com/a.html에 들어있는 noindex 메타태그를 확인 할 수 없고 결과적으로는 A.com/a.html에서 색인을 차단하고 있다는 사실을 인지하지 못합니다.
- ④ 하지만 b.html을 검색해 a.html을 색인에 포함시키게 됩니다.





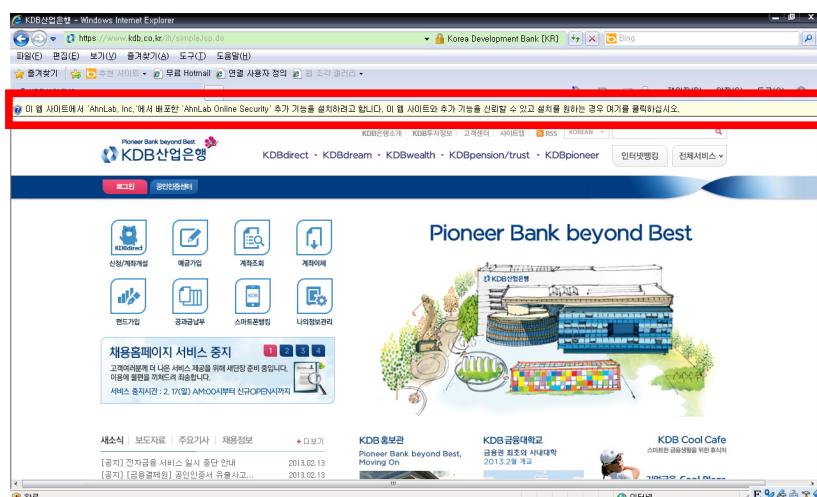
- 위의 사례에서 a.html의 색인을 막으려면 robots.txt를 수정해 로봇이 검색을 할 수 있도록 허용한 후 중요 페이지에 대해서는 noindex, nofollow 메타태그를 추가하여 색인할 수 없다는 것을 알려주어야 합니다.

### 3. 검색엔진 비 친화적 사이트 확인

#### 1) ActiveX 사용

- ① 비표준 기술인 ActiveX 과다 사용은 웹 호환성 및 보안문제를 발생시킴
- ② 국내 웹 서비스 환경이 IE에 최적화돼 공인인증 등 IE에 적합한 기술로 개발 및 보급되는 쏠림현상 발생
- ③ ActiveX 위주의 서비스 환경은 급변하는 모바일 환경에서 호환성이 떨어지는 불편 초래
- ④ 사용 예시

- <http://www.kdb.co.kr/> KDB산업은행



→ 메인페이지 접속 시 ActiveX를 설치하라는 메시지가 나타남

## 2) Image / Flash<sup>2)</sup> 위주의 웹 사이트

- ① Image나 Flash를 활용해 본문 텍스트 등을 표현할 경우 검색용 로봇이 이미지 안의 글자를 인식할 수 없기 때문에 검색이 불가능함
- ② 콘텐츠를 Image나 Flash로 표현하면 해당 내용이 색인되지 않는 것은 물론, 시각 장애인의 웹 접근성을 떨어뜨릴 수 있으므로 대체 텍스트 등 부가적인 장치를 추가해야 하는 번거로움 발생
- ③ 사용 예시

- <http://www.ksoi.org/> 한국사회여론연구소



→ 텍스트처럼 보이지만 드래그 혹은 마우스 오른쪽 버튼을 통해 보면 이미지파일임을 알 수 있음

- <http://www.smu.ac.kr> 상명대학교



→ 플래시 플레이어 설치되지 않은 브라우저에서는 콘텐츠가 제대로 표시되지 않음

- 2) 어도비사에서 만든 플랫폼으로 다양한 확장성과 화려한 효과 등으로 대부분의 PC에서 사용중인 확장 프로그램이다. 한 때 거의 모든 웹사이트에서 사용했을 정도로 인기를 끌었으나 HTML의 최신규격인 HTML5의 등장과 웹표준 지향적인 분위기에서 점차 사라져가고 있는 추세이다.

### 3) 검색엔진 비친화적 웹 사이트 확인방법

#### ① ActiveX<sup>3)</sup> 확인 방법

- ActiveX를 설치하지 않은 인터넷 익스플로러나 크롬 등의 웹브라우저로 해당 사이트를 방문했을 때 정상적인 웹페이지가 아닌 ActiveX 설치 페이지가 나타나면 ActiveX에 의한 검색엔진 비친화적 사이트로 분류
- 메인 화면을 포함해 로그인, 결제 등 ActiveX 사용 빈도가 높은 메뉴를 직접 확인해 ActiveX존재 여부 확인

#### ② Image 위주의 웹 사이트 확인방법


- Chrome 브라우저에서 Web Developer Extension을 설치한 후 웹 사이트를 방문하면 이미지 위주의 웹 사이트는 내용이 보이지 않으며, 대체 텍스트 항목을 체크한 경우 대체 텍스트를 사용하고 있는지도 확인할 수 있음

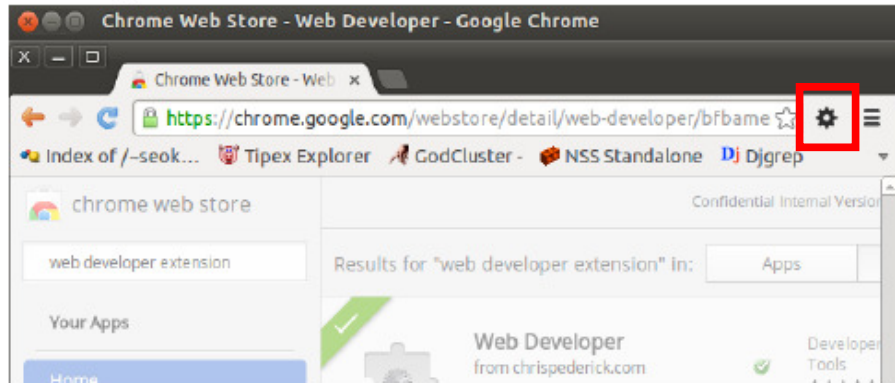
※ 설치방법 : 구글 등의 검색사이트에서 "Chrome Web Developer Extension"으로 검색하여 다운로드


---

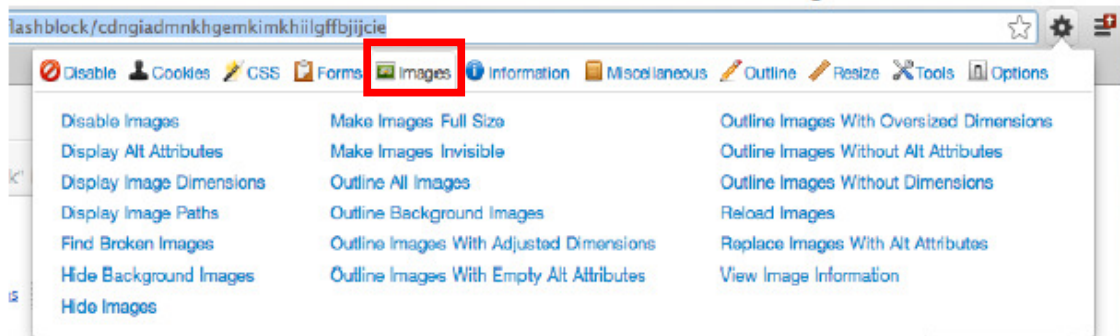
3) 마이크로소프트사에서 개발한 기술로 기존 응용프로그램에서 작성된 문서 등을 웹과 연결시켜 그대로 사용할 수 있도록 도와주는 기능을 한다. 초창기에는 인기가 많았지만 무분별한 응용프로그램의 자동설치와 스파이웨어 등 악성코드의 위험과 인터넷 익스플로러의 의존도 감소로 점차 사라져가고 있는 추세이다. 게임 및 은행 사이트 등에서 자동으로 설치되는 키보드 보안 프로그램 혹은 결제 프로그램 등이 ActiveX의 대표적인 예이다.

## 1단계

Web Developer Extension을 설치 후 우측 상단의  모양 아이콘을 클릭하면 여러 가지 메뉴가 나오는데 그 중 "Images" 탭 선택

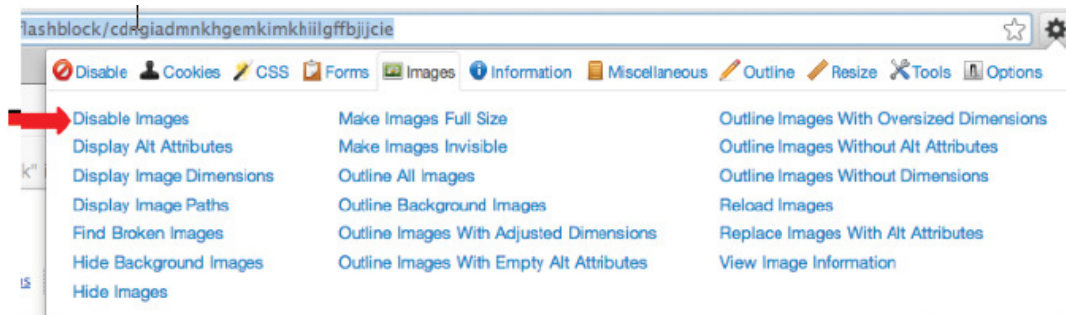


 아이콘을 클릭하면 여러가지 메뉴가 나오게 되는 데 그중 "images" tab 을 선택 합니다.



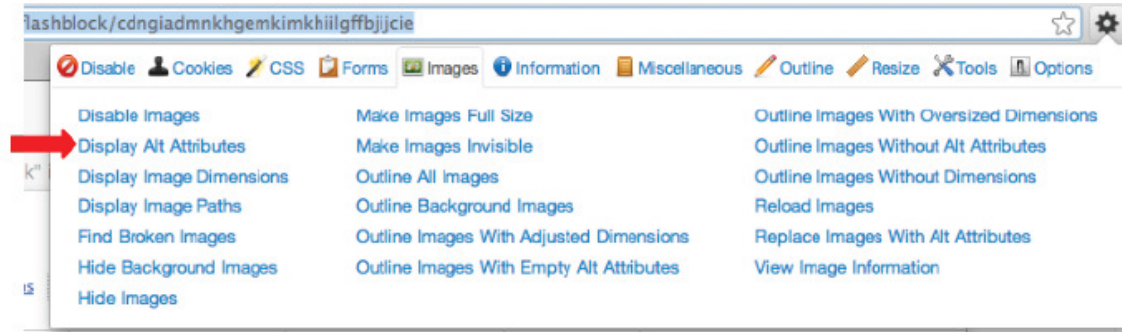
## 2단계

Images 항목 중 첫 번째 "Disable Images"를 선택합니다.  
(“Disable Images”를 설정하면 해제하기 전까지는 모든 사이트의 이미지가 보이지 않음)



### 3단계

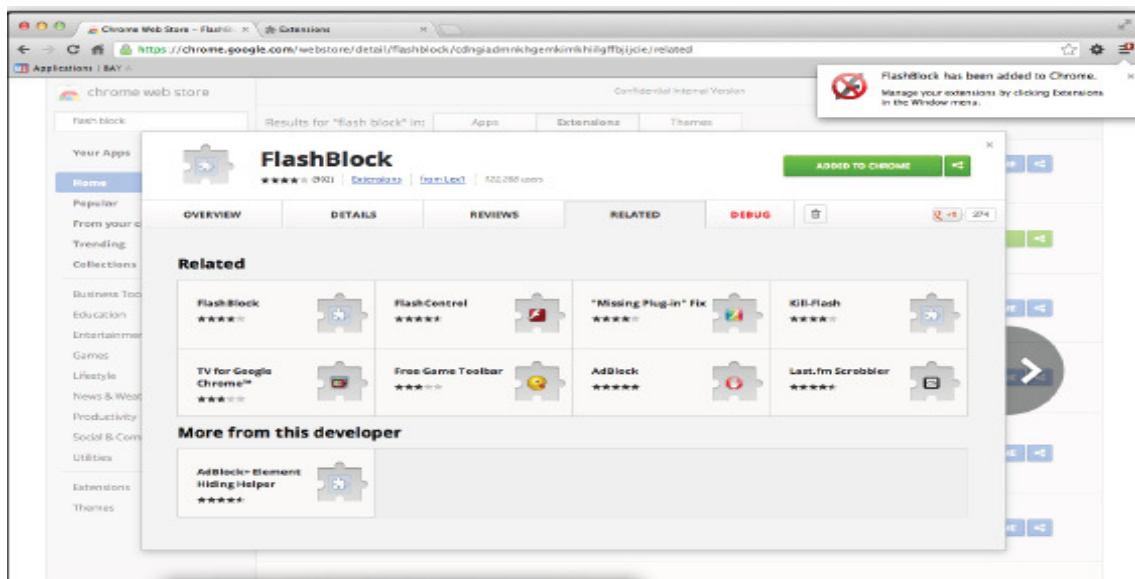
Images 항목 중 두 번째 "Display Alt Attributes"를 선택할 경우 이미지 대체 텍스트가 나타나 해당 이미지에 대한 대체 텍스트를 제공하는지 확인할 수 있습니다.



### ③ Flash 위주의 웹 사이트 확인방법

- Chrome 브라우저에서 Flash Block Extension을 설치한 후 웹 사이트를 방문하면 플래시로 제작된 서비스가 실행되지 않음

※ 설치방법 : 구글 등의 검색사이트에서 "Chrome Flash Block Extension"으로 검색하여 다운로드



#### 4) 가이드라인

- ① 웹사이트 제작 시 가급적이면 ActiveX를 사용하지 않고, 웹 표준을 준수하는 방식으로 제작하는 것이 바람직 함
- ② ActiveX를 반드시 사용해야하는 경우에도 공개된 페이지에 대해서는 ActiveX의 사용 없이도 정상적인 정보가 표시될 수 있는 방향으로 제작
- ③ 이미지 위주의 웹사이트가 필요한 경우, 사진에 대한 설명을 Alt태그로 나타내고 본문의 내용 전체를 Alt태그에 입력하여 최소한의 검색이 가능하도록 조치
- ④ 플래시의 경우에도 반드시 사용해야한다면 중요 콘텐츠를 직접 표현하는 방식을 피하고 네비게이션, 시각자료 등 웹 수집이 되지 않아도 관계 없는 내용을 중심으로 표현  
(플래시의 경우도 Alt태그를 이용해 대체문구를 만들어 검색이 가능하도록 조치)

#### ※ Alt태그

Alt태그는 이미지를 설명하는 목적으로 사용되는데, 일반 사용자에게는 Alt태그의 내용이 보이지 않지만 검색로봇은 Alt태그 속의 내용을 검색할 수 있음

#### ※ Alt태그 사용 예

- <http://www.snu.ac.kr> 서울대학교



```

364         <li class="tab01"><strong>서울대 60년사</strong></li>
365         <li class="tab02"><a href="/historyphoto/1940/1">사진으로 보는 역
366 사</a></li>
367     </ul>
368     <div class="ab0101_btn"><a href="/about/snu60/index.htm" title="새창
369 으로 열림" target="_blank"></a></div>
371
372     <dl id="intro1ist" class="write_box">
373         <dt class="historyfirst"><a href="#this" title="기원 起源 열고/
374 달기"> <strong><!--<em>0단계</em--></strong>
376 <span><em>달기</em></span> </a> </dt>
379         <dd class="historyfirst nodisplay">
380             <div class="textinterbox">
381                 <div class="imgright">
382                     
384                     <p><strong>법관양성소 1회 졸업생 이준 열사 (왼쪽)와<br/>지
385 석명 의학교 초대 교장</strong></p>

```

→ 위와 같이 이미지로 되어 있는 텍스트 정보는 Alt태그에 들어있는 본문 내용 덕분에 검색엔진 수집용 로봇이 이미지화 되어있는 텍스트까지도 수집을 가능하게 한다.



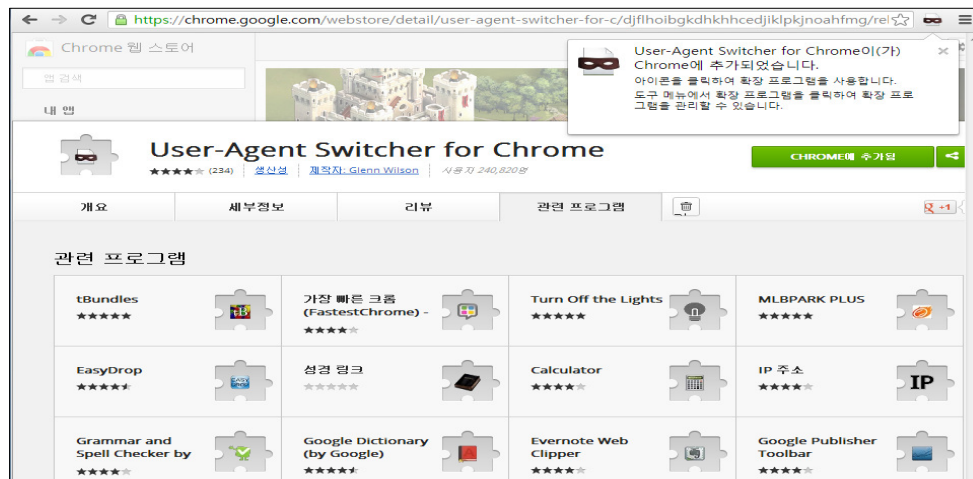
#### 4. User-agent Switcher를 이용한 검색 차단 확인


##### 1) 검색로봇으로 가장해 해당 사이트 접근

- ① 브라우저의 User-agent Switcher 기능을 이용해 설정을 마친 후 검색을 진행하면 해당 사이트에서는 마치 검색로봇이 접근한 것처럼 인식함
- ② 조사자의 PC가 웹 사이트에 접근할 때 해당 사이트가 로봇을 차단하고 있다면 차단 메시지나 나오므로 로봇검색 차단 여부를 쉽게 알 수 있음

##### 2) User-agent Switcher 설정방법(크롬 기준)

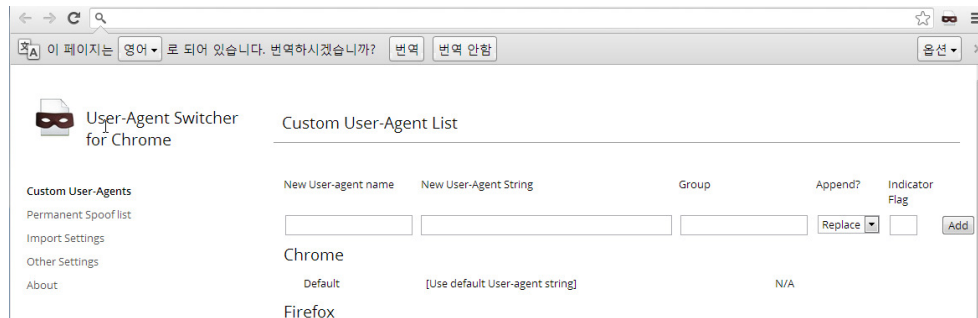
- ① 구글 등의 검색사이트에서 "User-agent Switcher for Chrome"으로 검색하여 다운로드하고 확장 프로그램 추가



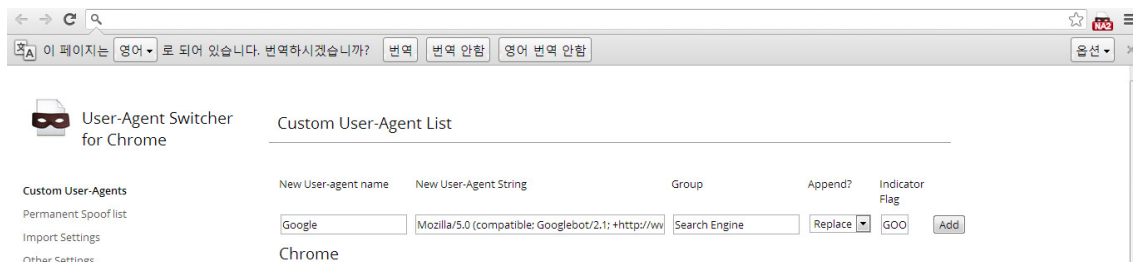
- ②  아이콘을 클릭하고 Settings를 누릅니다.



③ 아래와 같이 "Custom User-agent List"화면으로 이동하게 됩니다.



④ 여기에 다음과 같이 세 개의 User-agent를 추가합니다. 각각의 User-agent를 입력하고 Add버튼을 누르면 User-agent가 추가됩니다.(오 타 없이 입력해야 함)



- User-agent 입력가이드

[구글봇]↕

New Useragent name↕	Google↕
New Useragent String↕	Mozilla/5.0(compatible; Googlebot/2.1; +http://www.google.com/bot.html)↕
Group↕	Search Engine↕
Append↕	Replace↕
Indicator Flag↕	GOO↕

[네이버봇 1]

New User-agent name	Naver01
New User-agent String	Mozilla/4.0 (compatible; NaverBot/1.0; http://help.naver.com/customer_webtxt_02.jsp)
Group	Search Engine
Append	Replace
Indicator Flag	NAV01







[네이버봇 2]

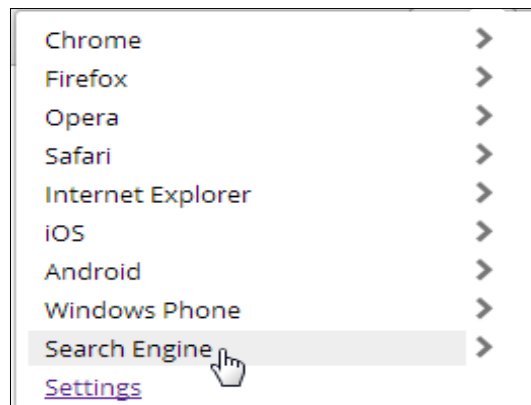
New User-agent name	Naver02
New User-agent String	Yeti/1.0 (NHN Corp.; http://help.naver.com/robots/)
Group	Search Engine
Append	Replace
Indicator Flag	NAV02

⑤ 화면의 가장 아래에서 다음과 같이 설정이 됐는지 확인합니다.

Search Engine

Google	Mozilla/5.0 (compatible; GoogleBot/2.1; +http://www.google.com/bot.html)	Replace	GOO	
Naver01	Mozilla/4.0 (compatible; NaverBot/1.0; http://help.naver.com/customer_webtxt_02.jsp)	Replace	NA1	
Naver02	Yeti/1.0 (NHN Corp.; http://help.naver.com/robots/)	Replace	NA2	

⑥ 위와 같이 한 번 설정 한 후, 다시 브라우저 화면으로 돌아가  아이콘을 클릭하면 아래와 같이 Search Engine 이라는 항목이 추가된 것을 확인할 수 있습니다.



⑦ Search Engine 메뉴를 클릭하면 조금 전에 추가한 Google, Naver1, Naver2가 추가된 것을 볼 수 있습니다. 여기서 Google을 선택하면

 아이콘이 으로 바뀝니다.

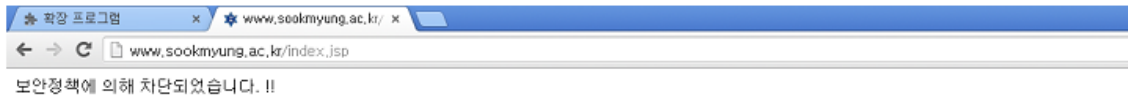
⑧ 이렇게 설정을 모두 마친 후 검색을 시작하면 마치 구글 검색로봇이 검색을 하는 것처럼 가장해 브라우징을 할 수 있게 됩니다. Naver1, 2의 경우에도 마찬가지로 네이버의 검색로봇의 입장에서 브라우징합니다.

User-agent 차단은 웹 서버에서 이루어지는 것이므로 검색 대상 사이트의 웹 페이지 중 하나만 확인하면 충분합니다.

다시 원래 상태로 돌아오려면 ⑥번 메뉴 Chrome메뉴에서 Default를 선택하면 됩니다.

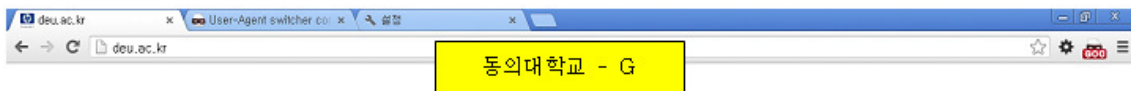
### 3) 차단 사례 확인

#### ① 숙명여자대학교 구글 차단화면



숙명여자대학교 - G

#### ② 동의대학교 구글 차단 화면



동의대학교 - G

비정상 접속으로 접속 차단 되었습니다.  
다시 접속하여 주십시오

동의대학교 전산정보원

#### ③ 상명대학교 네이버 차단 화면



상명대학교 - N2

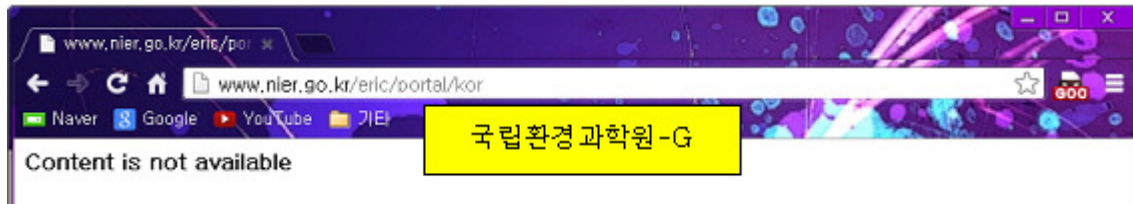
#### ④ 울산대학교 구글 차단 화면



불편을 드려 죄송합니다.  
기타 자세한 사항은 아래의 연락처로 연락부탁드립니다.  
웹방화벽 : Blocked Page 관련문의 : 052-259-1744,1733

울산대학교-G

#### ⑤ 국립환경과학원 차단 화면



### 4) 가이드라인

- ① User-agent의 접속을 차단은 방화벽이나 서버에서 이루어지고 소프트웨어마다 설정하는 방법이 다름
- ② 해당사이트의 보안담당자 혹은 웹마스터는 사이트의 환경 및 보안 정책에 알맞게 User-agent차단을 관리하고 있는지 확인할 필요 있음
- ③ 웹 사이트 관리자가 변경되거나 신규채용되는 경우 사내 보안 정책 등을 고려해 최소한 분기별 1회 User-agent정책을 조사해 불필요한 차단이 있는지 검토

### 5. URL공개 또는 비공개 여부 확인

#### 1) URL 변경여부를 직접 확인하는 방법

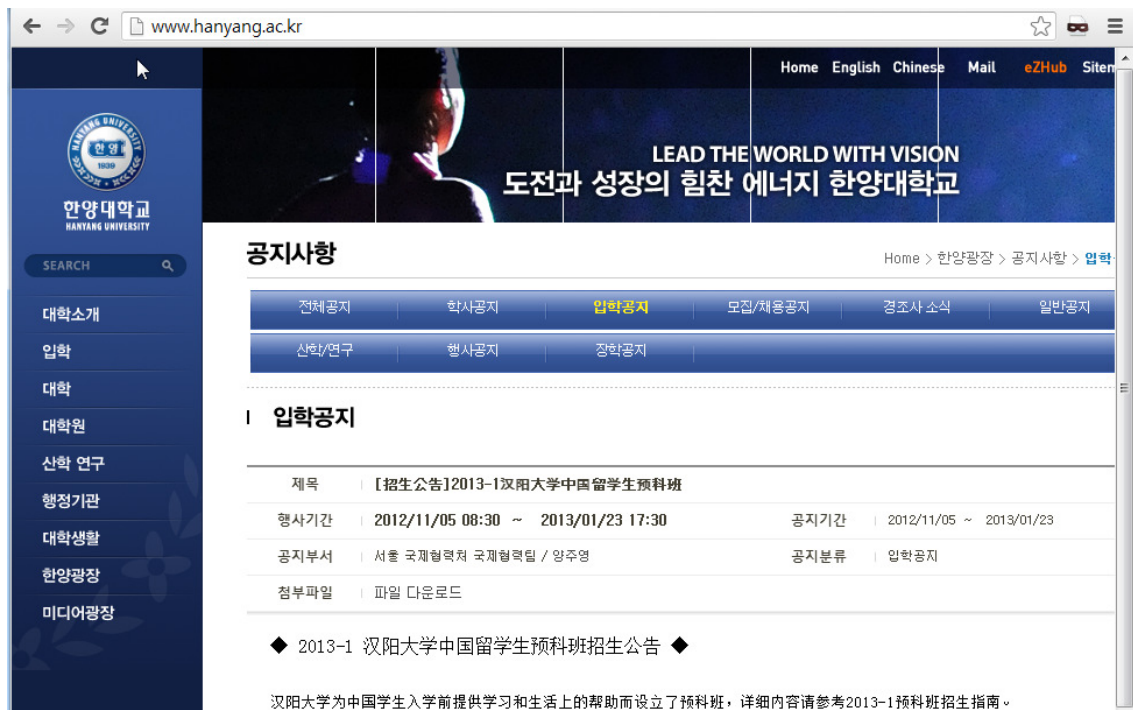
- ① 웹사이트와 그 안에 포함된 콘텐츠는 고유의 URL을 가지고 있기 때문에 URL만 알면 웹사이트의 특정 게시물에 접근할 수 있음
- ② 링크가 걸려있는 게시물이나 검색결과 페이지는 매번 URL이 바뀌는 것이 일반적이지만 필요한 경우 페이지가 변경됐음에도 불구하고 주소창의 URL이 변하지 않는 경우도 있음
- ③ 페이지가 변경됐지만 실제 주소가 바뀌지 않을 경우 검색로봇은 이를 인식하지 못해 검색이 제한될 수 있음
- ④ 게시판에서 게시물을 클릭했을 때 주소가 변경되는지 확인하는 등의 방법으로 해당 사이트가 URL을 공개하는지 확인 할 수 있음

⑤ URL확인 결과 예시

→ 한양대학교 홈페이지 주소 : <http://www.hanyang.ac.kr/>

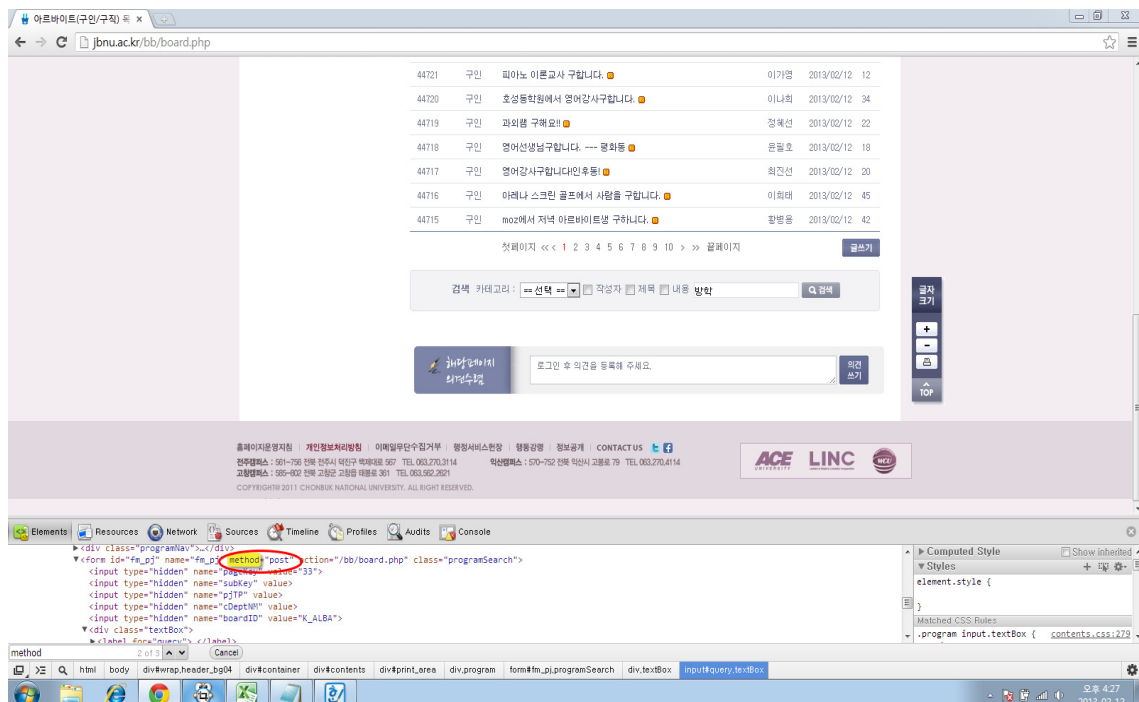


→ 한양대학교 홈페이지 내 공지사항 주소 : <http://www.hanyang.ac.kr/>



## 2) form 태그의 method 속성 중 get과 post 확인

- ① 변수를 사용하는 웹 페이지는 post 또는 get 방식을 사용해 서버에 변수를 전달함
- ② 웹 페이지가 form태그 method 속성 중 post방식을 사용하면 서버에 전달된 변수가 바뀌어도 웹 페이지의 주소는 바뀌지 않음
- ③ 내용이 다른 여러 개 웹 페이지의 주소가 바뀌지 않고 같은 주소를 사용할 경우 검색 엔진은 이 중 하나의 내용만 보여주게 됨
- ④ 따라서 post방식을 사용할 경우 검색로봇은 여러 개의 페이지 중 하나의 페이지만 검색할 수 있음
- ⑤ get방식을 사용하게 되면 변수가 웹 페이지의 주소에 표시되어 나타나기 때문에 페이지의 내용이 바뀔 때마다 웹 페이지의 고유 주소를 검색 결과로 보여주는 것이 가능함
- ⑥ 변수를 사용하는 웹 페이지를 검색엔진을 통해 보여주려면 get방식을 사용해야 하며 브라우저의 소스보기에서 어떤 방식인지 확인할 수 있음
- ⑦ 소스보기 확인 결과 예시(post방식)



### 3) 가이드라인

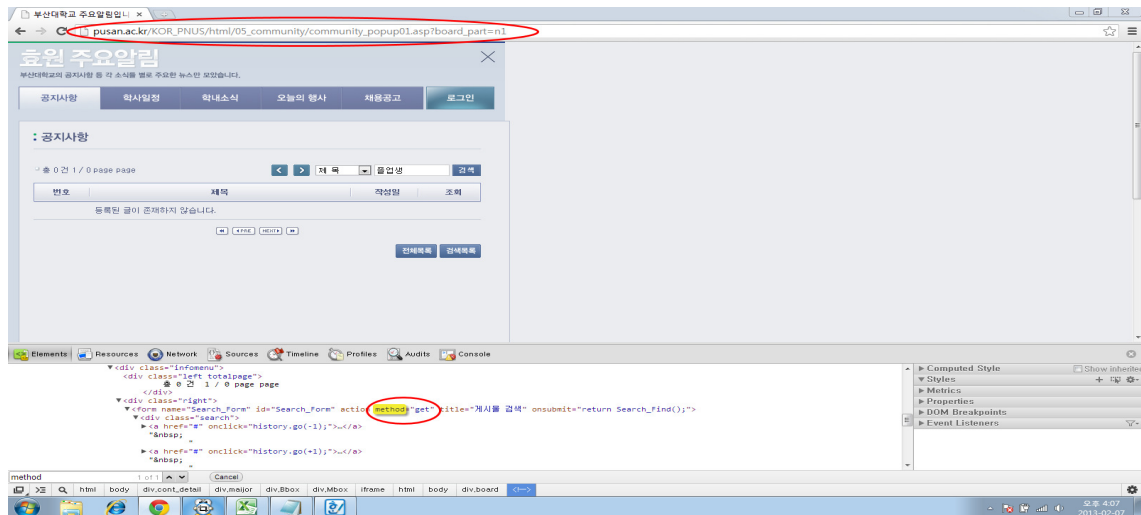
- ① POST방식이 파라미터를 URL이 아닌 HTTP 메시지의 body에 삽입하여 보내지만 이 메시지는 암호화되어 있지 않기 때문에 간단한 도구로 그 내용을 모두 볼 수 있다. POST를 쓴다고 반드시 보안에 좋은 것은 아니므로, 페이지가 변경됨에 따라 검색이 제한되는 상황을 최소화한다는 측면에서 가급적 get방식 이용
- ② 다만 get방식을 이용하면 클라이언트에서 서버로 보낸 자료가 URL에 모두 노출되기 때문에 비밀번호와 같은 개인정보를 get방식으로 보내는 것은 위험할 수 있으며 get방식으로 보낼 수 있는 자료의 양은 한계가 있기 때문에 내부 기준을 정해 적절히 사용

## ※ iframe+get method

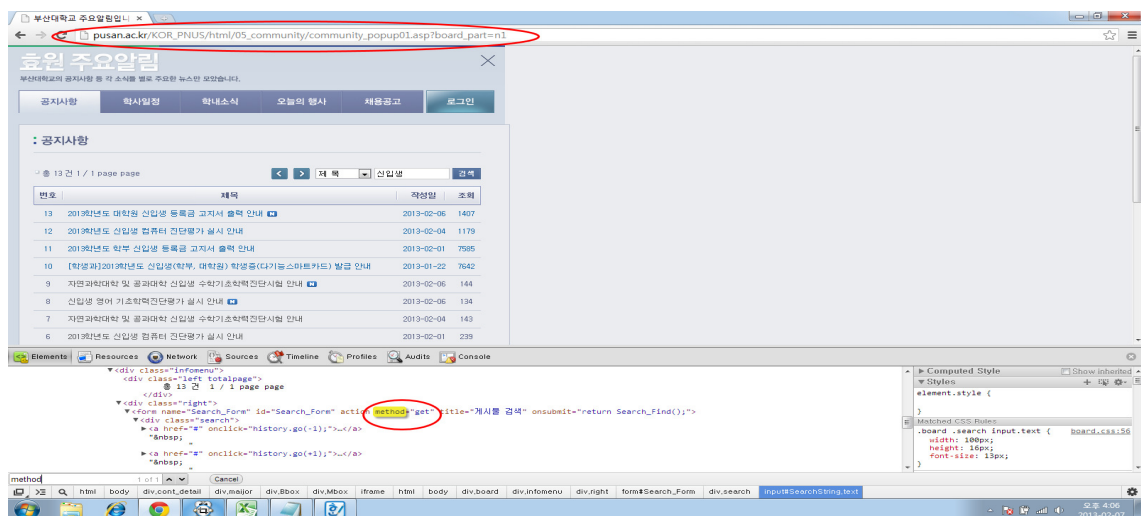
(:method는 get으로 되어 있으나 페이지 변환 시 URL이 바뀌지 않는 경우)  
예시)

사이트	주소 URL
부산대학교/ 공지사항	http://pusan.ac.kr/KOR_PNUS/html/05_community/community_popup01.asp?board_part=n1

→ 아래 예시 A, B처럼 get 방식으로 전송되는 웹페이지임에도 웹사이트 URL에 쿼리스트링(URL 주소 뒤에 입력 데이터를 함께 제공 하는 방법)이 붙지 않아 페이지 내용이 달라졌음에도 불구하고 URL은 변하지 않음  
예시 A)



예시 B)



### III

## 조사결과

### 1. 조사현황

조사항목	대학교	연구기관	합계	비율
㉠ robots.txt 차단	50개 (완전차단 32개)	35개 (완전차단 24개)	85개	42.5%
㉡ noindex / nofollow 태그 사용	4개	6개	10개	5%
㉢ ActiveX / Image / Flash 사용	58개	43개	101개	50.5%
㉣ User-agent Switcher 이용한 검색차단 확인	12개 (완전차단 2개)	4개 (완전차단 0개)	16개	8%
㉤ URL 비공개	36개	21개	57개	28.5%

\*조사기간 : 2013년 1월 21일 ~ 2013년 2월 20일

\*조사표본수 : 국내 국·공립 및 사립대학교 사이트 100개 / 정부산하 연구기관 등 공공 정보사이트 100개

### 2. 결과분석

- ㉠ 항목의 경우 페이지 전체를 차단하는 경우와 특정 페이지만 차단하는 경우로 나누어 조사를 실시했으며 대학의 경우 상대적으로 완전차단 비율이 높은 것으로 조사됨
- 5개 항목 모두 의미있는 차이를 보이지는 않았으나 연구기관이 대학에 비해 웹 개방성이 조금 높은 것으로 조사됨
- 조사대상 서비스의 42.5%가 로봇검색을 차단하고 있는 것으로 나타났으며 ActiveX / Image / Flash 등 검색 비친화적 요소를 포함하고 있는 사이트가 50.5%로 가장 높은 비율을 나타냄
- 인터넷 익스플로러 중심의 국내 웹 환경에 의한 특수성이 반영된 것으로 추정됨

### 3. 조사결과 종합

차단 개수	대학교	연구기관	합계	비율
5개	0개	0개	0개	0%
4개	2개	1개	3개	1.5%
3개	17개	7개	24개	12%
2개	31개	23개	54개	27%
1개	39개	37개	76개	38%
0개	11개	32개	43개	21.5%

- 조사대상 200개 사이트 중 약 78.5%는 어떤 형태로든 검색로봇의 접근을 제한하는 것으로 나타남
- 조사대상 항목을 모두 적용해 검색을 차단하는 경우는 없었으며 대학에서의 차단비율이 연구기관에 비해 상대적으로 높았음



#### 4. 조사의 한계

- 본 조사의 목적이 웹 개방성에 대한 실태를 파악함과 동시에 각각의 사이트가 처한 상황에 맞는 적절한 조치를 취할 수 있는 가이드라인을 제시하는 것임에도 불구하고 조사기관의 의도와 달리 웹 사이트 관리자가 적극적으로 개입하지 않은 경우 웹 개방성이 높은 것으로 나타날 가능성이 있음
- 본 가이드라인 조사항목 중 ㉔는 전수 조사가 아닌 해당 홈페이지를 포함한 5개 메뉴에 대한 샘플링 방식으로 진행되었으며 ActiveX와 같이 일정한 규칙 없이 필요에 따라 특정 웹 페이지에서 설치를 요구하는 경우를 모두 포함하지 못했기 때문에 전반적인 현황을 파악하는데 제한적으로 활용됨

#### 5. 조사의 제언

- 본 가이드라인 7페이지에 언급한 것과 같이 robots.txt와 noindex를 사용했음에도 불구하고 검색결과에 노출될 수 있기 때문에 로봇검색을 차단할 때 세심한 주의가 필요함
- 본 조사는 웹 개방성을 측정하기 위한 다양한 요소 중 일부를 이용하여 진행된 것이므로 검색엔진에 친화적인 웹 사이트 제작을 원하는 웹마스터는 본 가이드라인과 함께 구글에서 제공하는 검색 엔진 최적화(Search Engine Optimization) 가이드라인을 활용해 검색친화적인 웹 사이트를 구성하는 것이 바람직할 것으로 판단됨
- 본 조사에서 사용된 화면자료는 당시의 상황을 반영한 것이므로, 현재 시점에는 달라졌을 가능성이 있습니다.

※ 검색엔진 최적화 가이드라인

→ <http://support.google.com/webmasters/bin/answer.py?hl=ko&answer=35291>