Daum 웹크롤링 뉴스검색 기술 가이드

2009

본 문서는 웹뉴스 크롤링 제휴를 하고 있는 매체사들에게 기술적인 지원을 위해 만들어졌습니다.
따라서, 제휴매체사를 제외한 다른 개인 또는 단체에게 배포되는 것을 금합니다.

1. 웹뉴스 크롤링 체크리스트

웹뉴스 크롤링 등록전 아래 질문에 **'예'** 항목이 나온다면 제작가이드를 참고하셔서 **반드시** 관련사항을 수정하여 주시기 바랍니다. 등록과정에서 같은 사항이 발견되면 등록처리가 매우 지연되거나 미노출 될 수 있습니다.

1. 강력한 방화벽이 작동하고 있다.	예	아니오
→ 방화벽이 크롤러 접근을 막는다면 서비스가 불가능합니다.		
2. robots.txt로 크롤링을 막고 있다	예	아니오
→ robots.txt로 크롤링을 막는 사이트에 대해서는 크롤링을 진행하지 않습니다.		
3. 일일 허용 트래픽이 너무 적다.	예	아니오
→ 이로 인해 검색노출시 서버가 다운된다면 크롤링을 정지할 수 있습니다.		
4. 뉴스 리스트 페이지가 없다.	예	아니오
→ 목록페이지가 없으면 크롤링이 불가능합니다.		
5. 각 뉴스의 URL이 숨겨져 있다.	예	아니오
→ 각 뉴스의 링크주소는 모두 중복되지 않는 고유의 값을 가지고 있어야 합니다. (자바스크립트도 수집 가능함)		
6. 뉴스 제목과 등록시간, 본문 중 없는 요소가 있다.	예	아니오
→ 3가지 필수요소중 1가지라도 없다면 수집되지 않습니다.		
7. 본문내 사진 이미지 경로가 상대경로로 되어있다	예	아니오
→ 이미지가 상대경로일 경우 썸네일이 표시되지 않습니다.		
8. 등록시간 형식이 일정치 않다. (예를 들어 '년/월/일'과 '년/일/월/시/분/초' 혼재)	예	아니오
→ 등록시간이 일정치 않을 경우 뉴스가 수집되지 않습니다		
9. 카테고리에 속하지 않는 기사가 있다	예	아니오
→ 카테고리별로 수집이 이루어지기 때문에 카테고리에 포함되지 않는 뉴스는 수집되지 않습니다.		

2. 제작가이드

웹뉴스 크롤링 서비스의 보다 안정적인 서비스를 위해 다음 작업을 점검하고 있습니다.
아래 사안들이 입점 결정이후에 웹페이지에 지속적으로 반영된다면 보다 안정적인 뉴스수집이 가능합니다.

크롤링 필수요소 : ① 뉴스 리스트 페이지, 뉴스 뷰페이지(② 제목, ③ 시간, ④ 본문)

1 뉴스 리스트 페이지는 카테고리별로 URL을 적어주셔야합니다

사회 : www.news.co.kr/section=1
정치 : www.news.co.kr/section=2
경제 : www.news.co.kr/section=3

2 다음과 같은 주석을 뉴스 뷰페이지 HTML소스안에 추가해주셔야 합니다.

- 주석은 해당 데이터 앞뒤에 위치하여 합니다. (다음페이지 참조)
- 아래의 주석은 사이트 개편시 변경된 구조에 맞게 계속 적용하셔야 합니다.
- 아래 태그가 올바로 적용되고 있는 경우에는 뉴스 누락이 줄어들 수 있습니다.

제목 : <!--DAUM_TITLE--><!--/DAUM_TITLE-->
시간 : <!--DAUM_TIME--><!--/DAUM_TIME-->
본문 : <!--DAUM_CONTENTS--><!--/DAUM_CONTENTS-->

※ 주의사항 ※
주석을 삽입하실 경우 text 와의 사이에 다른 코드가 존재해서는 안됩니다.
다음 페이지 샘플 코드 참조

2. 제작가이드

< 잘못 적용된 사례> 붉은색은 오류원인임

<!--DAUM_TITLE→<!--제목--> 소녀시대 Gee 돌풍 <!--/DAUM_TITLE-->

<!--DAUM_REGDATE-->등록시간 : 2009년 1월 1일 KST<!--/DAUM_REGDATE-->

<img src="...">><!--DAUM_CONTENTS-->소녀시대는 신곡 Gee를 발표했다..... 끝<!--/DAUM_CONTENTS-->

< 정상 적용된 사례>

① 제목 태그의 시작, 마감사이에는 다른 태그가 들어올 수 없습니다.

<!--제목--><**!--DAUM_TITLE-->**소녀시대 Gee 돌풍<**!--/DAUM_TITLE-->**

② 등록시간은 년/월/일/시/분/초 단위만 가능하며, '등록시간, update' 등의 TEXT는 인식이 불가합니다.

단, 시간 표시 형식은 제한없음 2009/1/1, 2009-1-1, 2009년 1월 1일... 모든 형식 가능

등록시간<**!--DAUM_REGDATE-->** 2009년 1월 1일<**!--/DAUM_REGDATE-->**KST

③ 이미지, 기자명은 본문태그안에 포함되어야 합니다.

<**!--DAUM_CONTENTS-->**<img src="...">소녀시대는 신곡 Gee를 발표했다..... 이다음 기자<**!--/DAUM_CONTENTS-->**

※ 이미 입점된 매체들의 페이지 소스를 참고로 하시기 바랍니다.

<sample> http://www.eyeng.com/news/?m=1&category=0101&kind=&mode=view&no=1635

3 개별적인 크롤링 제외 설정 (제휴기사는 반드시 적용)

• 기사별로 크롤링을 제외시키기 위해서는 다음과 같은 조치가 필요합니다.
기사본문 안에 띄어쓰기하지 않고 white font로 **(뉴스검색제공제외)**라고 입력 (실제 기사에서는 안보임)
본문 내 위치는 관계없으며, 본문 태그 안에만 위치해 있으면 됩니다.
<!--DAUM_COMTENTS--> (뉴스검색제공제외) <!--/DAUM_CONTENTS-->
<!-- DAUM_COMTENTS--> (뉴스검색제공제외) <!--/DAUM_CONTENTS-->

4 이미지는 절대경로로 표시

• 크롤링된 본문 내의 HTML소스중에서 <img src="">형태의 소스는 반드시 절대경로(http://~)로 되어 있어야 함.

5 등록시간은 년월일 시분으로 표시

• 등록시간은 뉴스에 표기되어있는 시간을 기준으로 합니다. (전송된 시간 기준이 아님)
• 등록시간은 기본적으로 년월일 시분 까지 표시됩니다.
년월일만 표시하는 경우 시분을 제외하고 년월일만 표시됩니다.

3. 검색 도움말

- 다음 뉴스검색에서는 보다 정확한 검색결과를 위해 다음과 같은 검색패턴을 제공해드리고 있습니다.
- **언론사명, 기자명으로 바로 검색하시는 경우 정확한 결과를 얻기 어렵습니다**. (이는 대부분의 검색사이트에서 마찬가지입니다.)
- 등록되어있는 매체명+아래 패턴 외 특정 키워드에 대해 기사 노출은 불가능합니다.
- 다음 뉴스는 600여개의 매체가 입점되어있고 시시각각 뉴스가 추가되기 때문에 간단한 키워드로는 원하는 매체를 찾기 어렵습니다. 점검을 하시는 경우에는 한단어보다는 뉴스의 제목 전체를 입력하셔서 찾아보시길 권해드립니다.

1 매체명으로 검색(특정 매체의 뉴스 전체가 검색됨)

검색어 – <매체명> + 뉴스 예제) **ㅇㅇㅇ신문 뉴스, 다음인터넷뉴스 뉴스**

2 기자명으로 검색(특정 기자의 뉴스 전체가 검색됨)

검색어 – 기자 : <기자명> 예제) **기자 : 홍길동**

3 한 매체에서 특정 키워드로 검색(해당 매체의 기사에서만 키워드로 검색)

검색어 – <매체명> : <키워드> 예제) **MBC : 이효리 , KBS : 김연아**

4 한 주제에서 특정 키워드로 검색(해당 주제의 기사에서만 키워드로 검색)

검색어 – <주제명> : <키워드> 예제) **연예 : 이효리 , 스포츠 : 김연아**

4. 웹뉴스 크롤링의 처리 단계와 미노출 사례

1단계 : 메타정보 입력

입점등록이 된 매체들을 대상으로 수집 대상 카테고리의 URL주소와 뉴스본문페이지의 3요소(제목,본문,등록시간)을 구분할 수 있는 HTML정보를 입력하고 1일 이상 테스트하게 됩니다.

2단계 : 뉴스 카테고리에 접근

크롤러 서버는 카테고리별 뉴스 목록 페이지에 접근합니다.
입점이후 결정된 뉴스카테고리 경로로 크롤러 서버가 접근하여 리스팅 되어 있는 각 뉴스의 URL을 인식합니다.

3단계 : 뉴스 페이지 접근

위에서 수집한 뉴스페이지 주소에 접근하여 미리 입력된 메타정보대로 HTML소스를 분석하여 제목,본문,등록시간을 수집하게 됩니다.

4단계 : 수집된 정보를 필터링

임시조치 대상이나 규정에 맞지 않는 뉴스를 검색에 노출되지 않도록 제외합니다.

5단계 : 검색결과에 반영

필터링된 뉴스를 검색결과에 반영합니다.

수정작업

미노출 발생 1

URL 변경,
접근불가(방화벽)
짧은 시간내에 동시에
여러기사를 등록하는 경우

미노출 발생 2

3요소의 메타정보 불일치

미노출 발생 3

중복기사,제휴기사, 유료,
불법(저작권침해), 성인

썸네일 추출불가

이미지 상대경로인 경우
본문에 포함안되는 경우

4. 웹뉴스 크롤링의 처리 단계와 미노출 사례

1 뉴스 페이지(카테고리/본문)의 메타정보 변경

- 뉴스 카테고리와 본문 페이지의 URL주소가 변경된 경우.
- 뉴스 본문페이지의 HTML 정보가 변경된 경우 (제작가이드에서 권장하는 <--DAUM_xxxx--> 적용시 문제발생확률이 적어짐)
- 이상의 원인이 전체의 80%이상. (URL변경과 HTML변경시 웹뉴스검색 담당자에게 반드시 사전연락필요)

2 중복필터로 처리되는 경우

- 기사 본문이 다른 기사와 거의 비슷한 경우 최초 등록된 기사 외에 추가되는 기사는 중복으로 노출되지 않음

3 실제 기사등록시간과 표기된 시간이 차이가 날 경우

- 기사등록시간을 고의적으로 늦춰 올리는 경우 (최신기사로 앞에 노출되고자 미래의 시간으로 등록)

소녀시대 vs 원더걸스 2009.1.1 10:00 ← 실제 등록시간은 9:00
소녀시대가 5주간 정상을 차지하며 원더걸스의....

4 임시조치(블라인드 처리)에 해당

- 서비스 가이드 패널티가 부여되는 경우 참조

5 기사의 원문 링크가 잘못된 경우

- 시스템변경이나 URL변경등으로 기사의 원문링크가 잘못된 페이지로 연결되는 경우 삭제될 수 있습니다.

6 로그인이 필요한 기사일 경우

- 뉴스검색 컨텐츠로 부적합하거나 컨텐츠가 없는 카테고리, 로그인이나 유료결제를 요구하는 카테고리의 경우 수집하지 않습니다.

5. 뉴스 크롤링 FAQ

Q 뉴스가 일정시점 이후로 노출되지 않습니다.

A. 4-뉴스가 노출되지 않는 경우를 참조하시고, 뉴스페이지의 메타정보가 변경된 경우 Daum뉴스검색 담당자에게 변경요청 메일을 주십시오.

Q 뉴스가 일부 누락됩니다.

A. **4.** 뉴스가 노출되지 않는 경우를 참조하시고, **누락된 기사의 URL주소와 기사 제목**을 첨부하여 웹뉴스 담당자 메일로 보내주십시오

Q 썸네일 이미지가 나오지 않습니다.

A. 2-4 이미지 절대경로 표시를 참조하시고, 상대경로로 되어 있다면 절대경로로 변경하여 주십시오.

Q 뉴스가 너무 늦게 노출이 됩니다.

A. 웹크롤링 서비스는 기본적으로 1시간 내외로 노출되고 있습니다. 빠르게 노출되는 뉴스들은 크롤링 제휴방식이 아닌 매체들입니다.
대부분 특별한 수정사항이 없을시 발생했다면 시스템오류가 아닌 단순 딜레이 현상일 수 있습니다.
만약 최대 6시간이 넘거나 중간중간 누락되는 기사가 있을 경우 Daum 뉴스검색 담당자에게 검토요청메일을 주시기 바랍니다.

Q 매체명으로 검색시 다른 매체 뉴스가 나옵니다.

A. **3-1** 매체명으로 검색을 참조하시고, 매체명뒤에 "뉴스" 붙여 검색하시면 정확한 결과를 보실 수 있습니다.

Q 기자명이 나오지 않습니다.

A. **3-2** 기자명으로 검색을 참조하시고, 성함앞에 "**기자 :**" 를 붙여 검색하시면 정확한 결과를 보실 수 있습니다. 그래도 노출되지 않는 경우
5-2 제작가이드를 참조하시고, 기자명이 본문소스안에 있는지 확인해 주십시오.

5. 뉴스 크롤링 FAQ

Q 이미지 아래 표시된 이름으로 검색되지 않습니다.

A. 이미지에 포함된 캡션 텍스트의 경우 검색지원이 되지 않습니다. 본문안에 별도로 표기해주시기 부탁드립니다.

Q 칼럼 저자 이름으로 검색되지 않습니다.

A. 본문 아래쪽은 기자명으로 혼동되어 저자명이 제외될 수 있습니다. 이 경우 칼럼 저자명을 본문 위쪽에 표기해주시기 주시면 검색될 수 있습니다.

Q 동영상 뉴스가 나오지 않습니다.

A. 본문이 없는 동영상뉴스나, 팝업창, 뷰어 형식으로 보여지는 동영상 뉴스의 경우 수집에서 제외되게 됩니다.

Q 카테고리 분류가 다르게 나옵니다.

A. 웹크롤링은 URL기반으로 수집되기 때문에 섹션페이지에 접근하였을때 URL에 카테고리 구분표시가 없는 경우 카테고리가 다르게 분류될 수 있습니다. section=sports, cate=100 같은 형식으로 해당 섹션기사임을 표기하여 주시면 정상적으로 카테고리 분류가 가능합니다.

<sample> http://www.eyeng.com/news/?m=1&category=0101&kind=&mode=view&no=1635

6. 문의 메일

1 **웹뉴스에 대한 모든 문의메일은 24시간 뉴스센터로 연락주시기 바랍니다.**

24시간 뉴스센터 | 전화 : 080-677-3355 | 이메일 : http://media.daum.net/info/newscenter24.html (이메일문의하기 선택)

2 **빠른 검토와 답변을 위해 의도에 따라 문의 메일 형식을 반드시 지켜주시기 바랍니다.**

문의 형식이 지켜지지 않은 경우 처리가 늦어질 수 있습니다.

3 **문의 내용에 따른 처리시간은 다음과 같습니다.** 최대한 빠른 처리로 지원해드리겠습니다

도메인 변경, 제휴 종료 – 5일 이내 | 미노출처리, 섹션(카테고리) 추가or수정, 페이지 수정 – 3일 이내 | 일반문의 – 2일 이내

1) 미노출 문의 메일 형식

From: OOO
Sent: Thursday, August 28, 2008 1:48 PM
To: 웹뉴스 담당자
Subject: [문의] 기사 미노출

1. 매체명 : ＯＯ 뉴스
2. 미노출 범위 : 일부기사
   제목 http://xxx.xxxx.xxxx/aaaa/bbbb.html
   제목 http://xxx.xxxx.xxxx/aaaa/ccccc.html
3. 미노출 시기 : 3일전

+ 추가 내용

From: OOO
Sent: Thursday, August 28, 2008 1:48 PM
To: 웹뉴스 담당자
Subject: [문의] 기사 서비스 중지

1. 매체명 : ＯＯ 뉴스
2. 미노출 범위 : 전체
3. 미노출 시기 : 2일전

+ 추가 내용

서버 교체후 뉴스가 검색되지 않습니다.

7. 문의 메일

2) 섹션(카테고리) 추가,수정 메일 형식

From: OOO
Sent: Thursday, August 28, 2008 1:48 PM
To: 웹뉴스 담당자
Subject: [문의] 섹션 추가

1. 매체명 : O O 뉴스
2. 추가
   인터뷰 : http://xxx.xxxx.xxxx/aaaa/
   기획기사 : http://xxx.xxxx.xxxx/bbbb

+ 추가 내용

From: OOO
Sent: Thursday, August 28, 2008 1:48 PM
To: 웹뉴스 담당자
Subject: [문의] 섹션 url 변경

1. 매체명 : O O 뉴스
2. 변경
   인터뷰 : http://xxx.xxxx.xxxx/aaaa/
   기획기사 : http://xxx.xxxx.xxxx/bbbb

+ 추가 내용

3) 페이지 소스 수정 메일 형식

From: OOO
Sent: Thursday, August 28, 2008 1:48 PM
To: 웹뉴스 담당자
Subject: [문의] 페이지 수정

1. 매체명 : O O 뉴스
2. 수정안
   제목 :
   본문 :
   시간 :
3. 수정안 적용시기 : 2009년 12월 25일

+ 추가 내용

From: OOO
Sent: Thursday, August 28, 2008 1:48 PM
To: 웹뉴스 담당자
Subject: [문의] 페이지 수정

1. 매체명 : O O 뉴스
2. 수정안 : 다음 표준안
3. 수정안 적용시기 : 2009년 12월 25일

+ 추가 내용

8. 기사 삭제 와 수정

1 웹뉴스는 기사 수정이 자동반영되지 않습니다.

크롤링은 수집될 당시의 정보를 일정기간 계속 유지하기 때문에 수집된 이후의 기사의 변경내역에 대해서는
반응하지 못합니다. 따라서, 우선 2,3번 과 같이 삭제요청을 하신후에 후속조치를 취하셔야 합니다.
매체 사이트에서는 삭제되었지만 저희쪽에 삭제요청을 하지 않는 경우에는 기사노출에 따른 책임이 뒤따를 수 있습니다.

2 기사 문제시 뉴스센터로 꼭 삭제요청을 해주셔야 합니다. (삭제처리 20분내외)

뉴스센터는 이메일/전화로 24시간동안 연락이 가능하기 때문에 업무처리에도 많은 도움이 되어 드릴 것입니다.
그외에 뉴스와 관련된 문의도 무엇이든 가능합니다.

24시간 뉴스센터
전화 : 080-677-3355
이메일 : http://media.daum.net/info/newscenter24.html (이메일문의하기 선택)

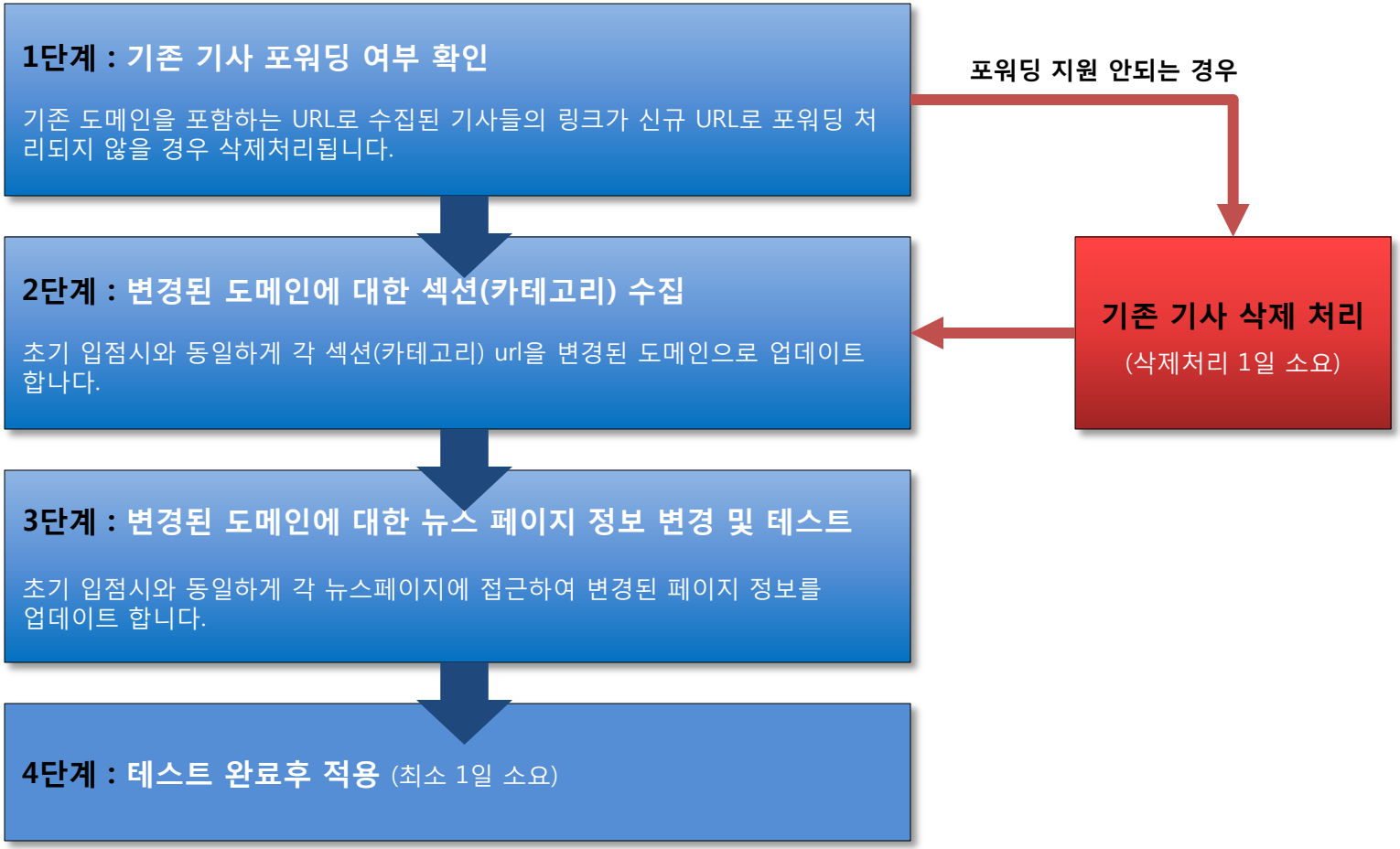
3 수정된 기사는 재등록이 되어야만 노출됩니다.

- **수정된 기사를 다시 등록하는 방법**
  기사삭제요청 -> **새로운 주소(url)로 수정된 기사 등록** -> 크롤링(자동) -> 웹뉴스 등록
  따라서 원문수정뿐만 아니라 수정된 뉴스를 새로운 주소로 올리셔야만 검색에도 노출되게 됩니다.

9. 도메인 변경 절차

- **기존 도메인 또한 유지하는 방안을 추천해 드립니다.**

  기존 도메인으로 링크가 유지되는 곳이 많기 때문에 잦은 변경이나 기존 도메인 삭제는 사이트에 좋지 않은 영향을 미치게 됩니다.
  기존 도메인이 삭제되는 경우 신규 도메인에서 수집처리 완료까지의 기간동안 크롤링은 중지됩니다.
  반면, 삭제하지 않는 경우 중지되는 시간을 최소화 할 수 있습니다.



감사합니다