비밀에 집중하라, 그리고 공유하라

씨랭크 비밀문서

2016년 네이버 컴퍼런스

프로그래머 문호영 각색

누구나 알고
또 누구나
알지못하는
씨랭크

이 내용을
돈으로
값을
매기지말라

네이버
알고리즘의
비밀

I. 콘퍼런스 개요

행사명: 2016년 네이버 컴퍼런스
주제: 검색서비스가 사용자에게 계속 사랑을 받으려면
일시: 2016년 10월
발표 세션:

- 1교시: 웹수집 · 분석 (김기동)
- 2교시: 웹스팸과의 전쟁
- 웹검색과 랭킹 (김상범)
- 네이버 웹검색 서비스 (김종범)
- 네이버가 알려주는 웹검색 공략 (김종범)
- 내 사이트와 연관된 채널 (이다해)
- 웹마스터도구 활용법 (홍상현)

중요 메시지

- “빠짐없이, 빠르고, 정확하고, 공정하게” 검색결과를 제공해야 한다.
- 웹공간은 방대하고, 품질관리(스팸 대응 포함)가 핵심 과제이다.
- 검색엔진 랭킹은 다양한 시그널과 기계학습을 통해 결정되며, 공개가 쉽지 않다.
- 사이트 운영자는 웹표준 준수, 컨텐츠 품질관리, 구조화 데이터 활용 등을 통해 노출 기회를 높일 수 있다.

목표

- 검색서비스 엔지니어와 운영자, SEO 전문가를 대상으로 네이버 검색 시스템의 전반을 공유
- 웹수집, 웹스팸 대응, 랭킹 알고리즘, 웹마스터도구 활용법 등 심층 노하우 제공
- 검색 서비스 품질 제고 및 웹생태계 발전

II. 1교시: 웹수집 · 분석 (김기동)

1. 웹수집 시스템 개요

- 웹 공간은 상상을 초월할 정도로 방대하며, 품질이 낮은 정보(스팸/저품질)도 많음.
- 네이버 관측 기준 약 1조 개 규모(전체 웹의 1/100 수준)
- 크롤러(Crawler) “Sunny”를 통해 페이지를 수집 및 분석

2. 규모와 품질의 문제

- 규모: 전 세계적으로 3억 개 이상의 도메인, 웹사이트 10억 개 이상 추정
- 품질: 모아 놓으면 절반 이상이 불필요한(스팸, 저품질) 정보

3. 우선순위 결정 알고리즘: IDEAL Crawler

- CrawlPriority(c, c', m) $= w \times NKCE(c,m) + (1-w) \times NKACE(c',m)$
  - c: 대상 URL의 클래스
  - c’: 소스(링크 출처) URL의 클래스
  - m: 모니터링 시간

- NKCE(c, m) 주요 지표

  - 얼마나 자주: RFC(Recently Fetched URL Cache)
  - 얼마나 많이: Crawl Ticket Count
  - 어떻게 확장: Link-Extension Policy

II. 1교시: 웹수집 · 분석 (김기동)

4. 웹공간 & 웹페이지 분석

- White Network 기반 분석
- 문서 외부의 링크, 공유(레퍼) 등 트러스트 링크 개념
- 스팸/저품질 사이트를 걸러내기 위한 분류 기준
- 해외/학술정보 등도 적극 수집·분석을 시도

5. 연구 성과

- 2017년 4월 19일부터 약 5.5배 확장(White Site 확대)
- 사용자 체감 개선: 의미 있는 결과 다수 확보
- 앞으로도 “씨랭크(C-Rank)” 등 지표개발을 통한 도전적 연구 지속

III. 2교시: 웹스팸과의 전쟁

1. 나쁜 놈들 전성시대: 스팸의 위협

- 스팸은 돈이 되는 곳에 몰려들며, 빠르게 진화
- 검색결과에 등장하면 사용자가 서비스 품질에 불만을 가지게 됨

2. 스팸대응 실패사례

- Tumblr 스팸: 무시하다가 뒤늦게 대응, 서비스 품질 크게 하락
- '070 스팸전화': 인터넷전화 전체 이미지 훼손, 가입자 급감
- 한메일 스팸: 스팸메일로 인해 주요 사용자층 이탈, 네이버 메일로 이전

3. 스팸피해와 대응 중요성

- 스팸을 방치하면 서비스 전반의 신뢰도 하락
- 고비용 · 고도화된 기술과 인력이 필요

4. 스팸의 유형

- Web Contents Spam: 자동생성 콘텐츠, 워드 샐러드, 히든 텍스트 등
- 리다이렉트: URL을 중간에 바꿔치기
- 클로킹: 검색봇에는 정상페이지, 사용자에게는 스팸 페이지
- 해킹: 정상사이트를 해킹하여 스팸 페이지 삽입

III. 2교시: 웹스팸과의 전쟁

5. 스팸을 잡기 어려운 이유

- 스패머는 웹환경과 검색엔진을 매우 잘 이해하고, 끊임없이 로직을 우회
- 검색엔진은 기술적으로 '명품 vs. 고퀄리티 이미테이션' 식 구별이 필요

6. 스팸 분류 및 기술

- TextCNN: 성인, 도박, 보험 등 키워드 기반 필터링
- Grammar: 문법에 맞지 않는 문서 자동생성 콘텐츠 제거
- Information: 정보량(가격, 호텔이름, 연도 등) 부족 시 스팸 판단
- GRAPH: 사이트간 통계로 스팸 호스트/광고 링크 추출
  (성균관대 황지영 교수팀과 산학)
- 클로킹 스팸: JS 분석, NLU + GRAPH로 식별
- 해킹 스팸: 정상 사이트 내 스팸영역 기생 → 문서 Layout/클러스터링 분석

7. 결론

- 인기 있고 잘되는 서비스에는 반드시 스팸이 몰려든다.
- 지속적 기술개발 및 모니터링이 필요.

Ⅳ. 웹검색과 랭킹 (김상범)

1. 글로벌 검색 시장 현황

- Google: 대표적 선두주자
- Bing: 구글을 많이 모방
- Yahoo: Bing + 구글 일부 혼합
- Baidu: 중국 정부가 구글 철수 후 현지화
- Yandex: 러시아 대표, 구글이 추격
- Seznam: 체코 로컬 검색
- 네이버: 한국에서 지역서비스 · 쇼핑 강점, 웹검색 격차 고민

2. 웹검색이 어려운 이유

- 방대한 웹페이지(규모)
- 관리 불가능(품질 보장 어려움)
- 메타데이터나 웹표준 미준수 사이트 다수

3. 네이버가 지향하는 좋은 검색결과

- 검색평가 가이드라인: 5점(완벽)~1점(무관)
- 데이터셋 생성: 사람 손으로 정확한 답안을 제공(학습용) → 기계학습

Ⅳ. 웹검색과 랭킹 (김상범)

4. 랭킹시그널과 'Learning to Rank'

- 시그널(피처)이 곧 검색 품질.
- 공개 순간 무력화 위험 → 비공개 유지.
- 엔지니어는 매일 시그널 발굴, A/B 테스트 진행

5. 랭킹성능 평가

- DCG(Discounted Cumulative Gain) 등으로 정량 평가
- “랭킹 이상현상 = 기술적 한계” (조작만이 아님)

6. 결론

- 만족도 높은 검색엔진은 전 세계적으로 몇 개 안 남음
- 랭킹은 전형적 기계학습 문제
- 시그널 비공개 이유: 무력화 우려
- 아직 기술 한계 존재

V. 네이버 웹검색 서비스 (김종범)

1. 웹검색의 변화와 방향

- 90’s: 사이트 검색, 검색등록(신청 & 에디터 DB)
- 스팸 및 변질에 취약, 변경사항 수동대응 한계

2. 웹환경 변화와 네이버의 대응

- 웹표준 준수 유도, 자동수집 및 콘텐츠 이해(출처 분석) 기술
- 사이트 영역 개선 & 통합 진행

3. 웹문서·사이트 탭 통합

- 2017.12: “웹문서 + 사이트” → “웹사이트” 탭으로 통합
- 2018년 1분기: 통합검색에서도 동일 UI 적용

4. 다변화된 UI와 노출 방식

- 리뷰(별점), 지역(지도 핀), 방송(재생정보) 등 리치한 정보 제공
- 서브링크 확대: 사이트맵/메뉴 자동 식별 → 브랜드 강화

5. 사용자 가치

- 생산자: 노출 기회 확대(브랜드 · 연관 질의 등)
- 소비자: 양질의 정보로 빠르게 이동

VI. (2부) 네이버가 알려주는 웹검색 공략 (김종범)

1. 만족도 결정 요소: 관련성 · 신뢰성 · 접근성

- 관련성: 사용자가 입력한 키워드와 문서 내용의 일치도
- 신뢰성: 사이트 운영기간, 신규콘텐츠 갱신, 외부 링크 평판 등
- 접근성: 사이트 속도, UI/UX 편의, 광고 과다 여부 등

2. 검색 = 수집 + 색인 + 랭킹

- 수집: 웹표준, 기계가 이해하기 쉬운 구조 → 필수 조건
- 랭킹: 사용자 만족도, 스팸 여부, 외부 평판 등 여러 시그널 복합

3. 웹마스터도구 활용

- 로봇 접근, 사이트맵 제공, 오류 상황(404, 3xx 등) 체크
- 품질지표 모니터링 및 개선

4. 랭킹과 가이드라인

- 키워드 반복보다 주제 충실도, 정보 정확성이 중요
- 외부 링크(신뢰 링크) 확보, 브랜드화된 페이지 구축

5. 소문 & 질문

- “대규모 사이트만 우대?” → 사실 아님
- “내부 서비스가 우선노출?” → 공개 시그널 부재, 기술부족 문제
- “사이트 모든 문서가 노출?” → 전부 보장 불가(수요와 효율)
- “웹사이트 최적화는 어렵다?” → 웹표준 지식 · 장기간 선호도 누적 필요

VII. 내 사이트와 연관된 채널, 네이버 검색에 알려주세요 (이다해)

1. 대표사이트 연관채널 노출

- 구조화 데이터(schema.org) 마크업을 통해 연결
- 페이스북, 인스타그램 등 SNS 계정 → 검증 후 굴비처럼 묶어 노출

2. 연관도 검증 및 변질 감지

- 도메인 소유 권한 확인, 해킹 · 판매로 인한 변질 파악

3. 서브링크(사이트 내부 메뉴) 구현

- 웹표준 준수: 자바스크립트 링크 최소화, HTML 구조화
- GNB · LNB 구성 & 사이트맵: 검색로봇이 쉽게 파악
- 사이트 활성화: PV · 외부평판 중요

VIII. 사이트 품질향상을 위한 웹마스터도구 활용법 (홍상현)

1. 검색시스템 이해: 수집 → 색인 → 노출

- 웹마스터도구: 사이트 소유자가 검색로봇 수집상태, 오류, 품질보고 등 확인 가능

2. 사이트 품질 진단 및 개선

- robots.txt / 메타태그: 검색허용 여부 설정
- sitemap.xml: 하위 페이지 구조와 변경빈도, 중요도 표기
- HTML 마크업: 웹표준 준수, 프레임 지양, 주요 텍스트는 이미지 대신 텍스트로
- 링크빌딩: 앵커텍스트를 정확히, 중복URL 관리(캐노니컬 등)

3. 구조화 데이터

- 리치 스니펫: 검색결과에서 풍부한 정보 제공
- schema.org / JSON-LD / 마이크로데이터 등 도입 검토
- 채널 연동(블로그, SNS 등) 시 마크업으로 사이트와 연결

4. 사이트 품질향상을 위한 기본 가이드

- 브랜드에 맞는 도메인
- 페이지 경량화 (속도 최적화)
- HTTP 상태코드 적절 사용(404, 3xx, 5xx 등)
- 모바일 · 멀티 디바이스 대응
- 낚시성 키워드 지양, 사용자 선호도 장기간 누적
- 전문 에이전시 협업 또는 SEO 전문가 컨설팅

이 글을 마무리하며..

검색 엔진은 방대한 웹정보를 신뢰도 높게 연결하는 핵심 인프라입니다.
네이버는 웹문서를 더 많이, 더 정확히, 더 깨끗하게 수집·분석하기 위해
크롤러 기술, 스팸대응, 랭킹 알고리즘, 웹마스터도구 등을 지속적으로
고도화하고 있습니다.

이 자료를 통해 검색 생태계에 대한 폭넓은 이해와 함께,
각 웹사이트가 검색결과에서 더 좋은 성과를 낼 수 있길 바랍니다.

감사합니다.

시크릿 키코드

(복사해서 이용하세요)

PTJBPZJBRUFBJLPZTPLLZRAUA