hci+d lab.

Week 07 • 소셜네트워크 데이터마이닝과 분석

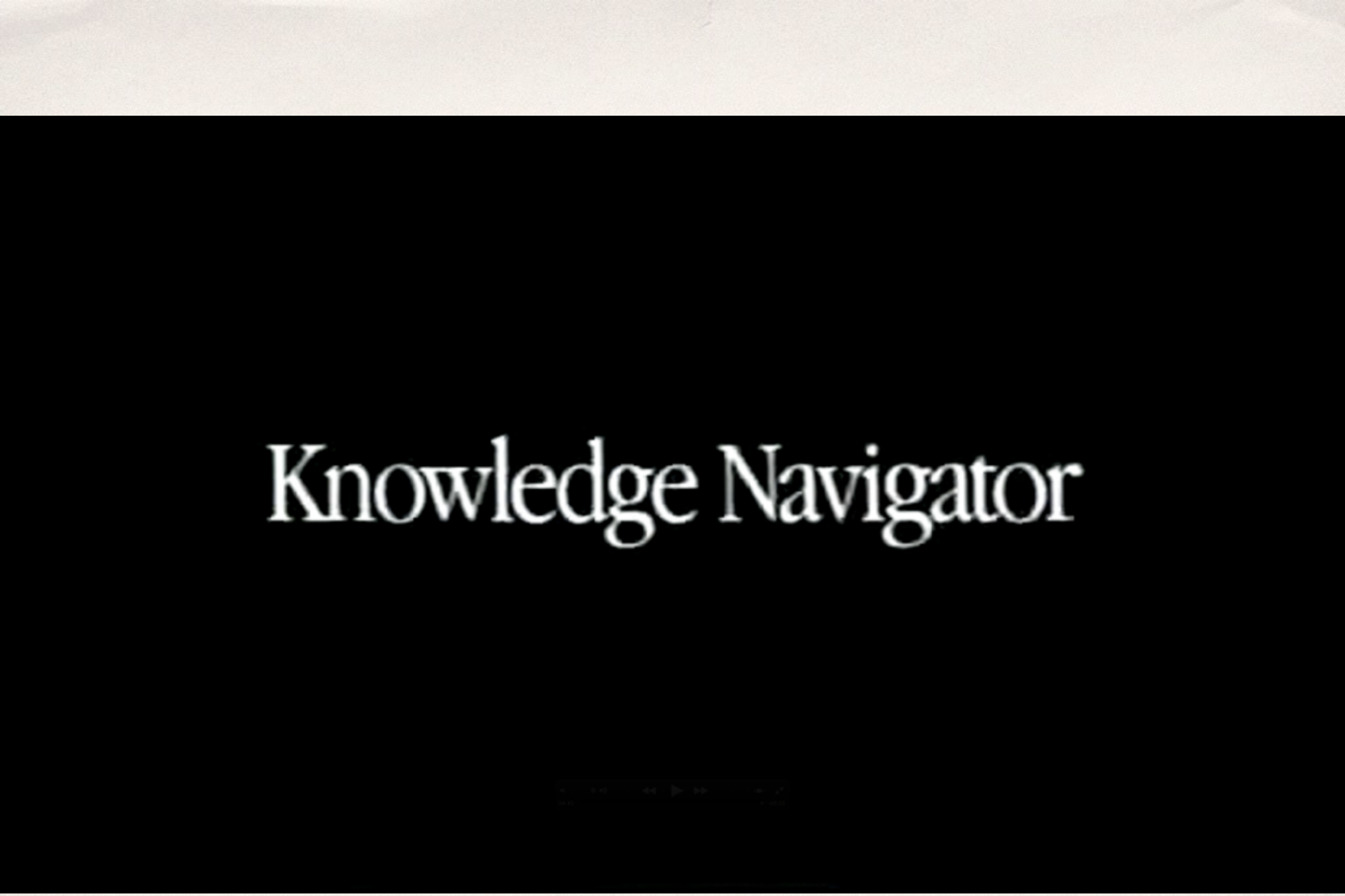
Information Retrieval

Joonhwan Lee
human-computer interaction + design lab.

오늘 다룰 내용

- Information Retrieval

1. Information Retrieval



Apple, 1987

Apple, 2011


Demo

Siri

- Personal Assistant application for iOS
  - voice recognition using **Natural Language Processing**
  - **agent**: answer questions, make recommendations, delegate requests to an expanding set of web services
  - **mashup**: integration with web services using their APIs
    - Yelp, Yahoo Local, CitySearch, OpenTable, Eventful, Movie Tickets, RottenTomatoes, New York Times, Bing, Google etc.

Information Retrieval 의 역사

- 현재 대부분의 정보 사이트는 검색을 기반으로 하고 있음
  → 검색엔진
- 검색엔진의 개념은 컴퓨터가 등장하기 이전 시대 부터 존재, 기본적인 개념들이 정리 됨 (문헌정보학의 영역)
- 컴퓨터의 등장 → 정보 검색 역사 시작
- 1945년 Vannervar Bush 의 논문에서 정보검색 (Information Retrieval)이라는 용어 제시, 이후 1세대 컴퓨터가 등장하던 시기에 미국에서 정보검색의 역사가 시작 됨.

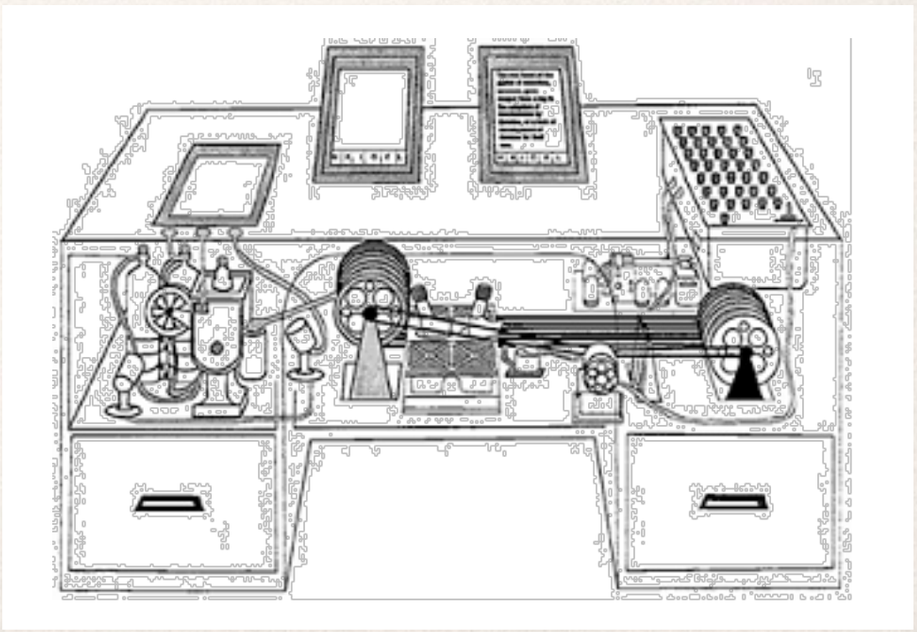
Information Retrieval 의 역사

- 1945 - 1955 (검색엔진의 태동기)
  - 정보검색에 대한 아이디어와 이론적 배경들이 제시 됨
  - 정보검색이라는 용어가 처음 사용 (Vannervar Bush)
    - “As We May Think (The Atlantic Monthly)”: 인간이 이루어 놓은 지식을 한데 보관하는 연구의 필요성 언급
    - 여러 세대에 걸쳐서 축적한 방대한 양의 자료들을 쉽고 빠르게 검색하여 이용할 수 있다면 인간의 지성은 크게 증가
      할 것이라고 주장 → Memex 장치 고안

Vannervar Bush, Ph.D.

Information Retrieval 의 역사

- 사진과 같은 정보들을 스캐닝하여 저장할 수 있는 스캐너 (왼쪽)
- 정보를 검색할 수 있는 타자기 (오른쪽)
- 검색된 결과를 출력해 주는 스크린 (가운데)
- 원통형의 장치 → 데이터를 저장하는 마이크로필름(micro film): 이 부분에 미세하게 저장된 데이터를 찾아 화면에 투사, 사용자가 볼 수 있게 한다.



Information Retrieval 의 역사

- 1960년대
  - 검색엔진에 대한 연구 개발이 이루어진 시대
  - 현재 거론되는 대부분의 검색모델들이 이 시대에 정립되었음
  - 대용량 정보 검색 시스템의 초기 모델들이 제시되고, 초기 정보 검색 시스템이 구현되기도 하였다.

Information Retrieval 의 역사

- 1970년대
  - 전자문서의 시대.
  - 개인용 컴퓨터의 등장과 워드프로세서의 개발로 처리해야 할 문서의 수와 양이 비약적으로 증대 되었다.
  - Paper-free office 환경에 대한 연구가 진행 → 전자문서가 종이문서를 대체하는 환경에 대한 연구 → 정보의 양이 증가
  - 디스크드라이브 발명 → 이 당시는 데이터 저장에 따른 비용이 비싼 편 (2000불/M) 이었지만, 대용량 검색시스템의 상용화를 이끌게 됨

Information Retrieval 의 역사

- 1970년대
  - Database Management System의 등장!
    - 단순한 계층모델의 DB에서 관계형 모델로 발전 (데이터의 연계)
    - Information 이 아닌 Data 중심
  - 초기의 정보검색은 인공지능의 한 분야로 인식되어 오다 70년대 들어 인공지능에서 분리, IR 의 영역을 독자적으로 구축

Information Retrieval 의 역사

- 1980년대
  - 본격적으로 전문 검색엔진 등장 시작
  - 컴퓨터 성능의 비약적인 향상과, CD-ROM 등 다양한 미디어의 등장으로 하드웨어적인 요건이 좋아짐 → 대용량 멀티미디어 데이터베이스 시스템 등장 → 정보의 양 증가
  - 검색에 대한 사용자의 요구 증대 → 도서관 위주의 검색 기술이 지속적으로 발전

Information Retrieval 의 역사

- 1990년대 이후
  - Vannervar Bush 박사의 검색에 대한 아이디어가 발표된 지 45년 만인 1990년에 **Archie 라는 최초의 검색엔진**이 등장
    - McGill 대학교의 Alan Emtage 라는 학생이 개발
    - FTP에 올려진 파일들을 검색할 수 있게 함: 사용자가 찾고자 하는 파일명을 입력하면 전세계 익명 FTP 서버를 통해 파일을 검색
    - Archie 는 전세계의 익명 FTP 서버를 주기적으로 접속, 모든 디렉토리와 파일들을 거대한 색인 형태로 Archie 서버에 저장 → 사용자가 Archie 에서 특정한 파일을 검색하면 서버의 색인 목록에서 이를 찾아 다시 원래 그 파일이 저장된 주소를 출력

Information Retrieval 의 역사

- 1990년대 이후
  - 1990년 봄, 미네소타 대학의 Mark McCahill → Gopher 라는 시스템 개발
    - 정보를 주제별로 메뉴를 구성하여 이용할 수 있도록 한 시스템
    - 인터넷에 익숙하지 않은 사람들도 쉽게 정보를 찾을 수 있게 함
- 1990년 겨울, 유럽 물리입자 연구소 (European Organization for Nuclear Research, CERN) 의 Tim Berners Lee 박사가 World Wide Web을 개발
  - Hyper Text Transfer Protocol (HTTP) 와 Hyper Text Markup Language (HTML) 을 개발 → 정보를 링크를 통해 서로 연결할 수 있는 방법을 개발

Information Retrieval 의 역사

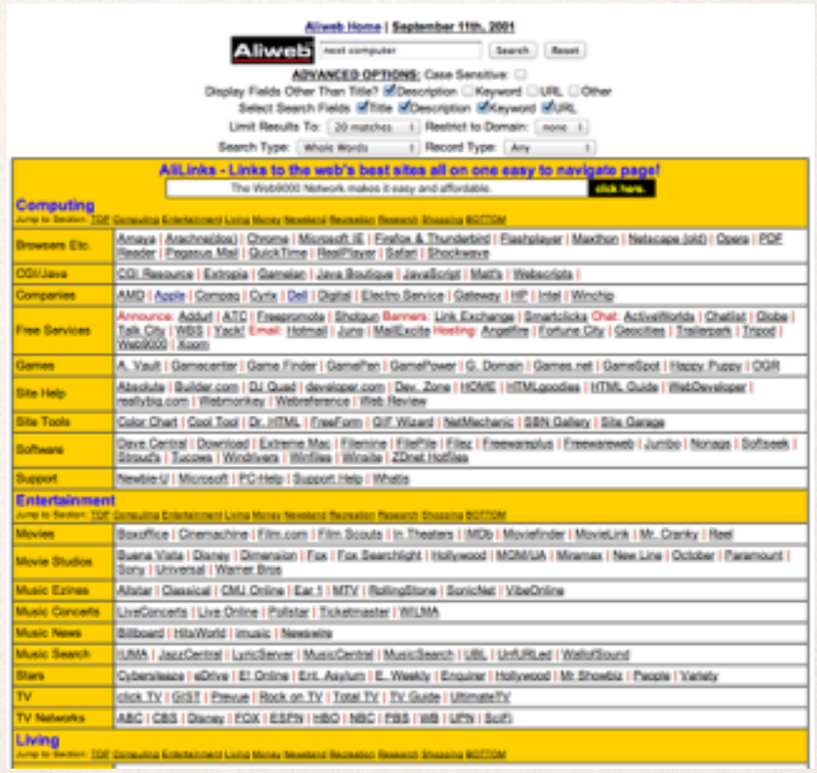
- World Wide Web
  - 팀 버너스 리는 전세계 물리학자들이 공통된 기기나 소프트웨어 없이 데이터를 공유할 수 있어야 한다고 생각 → A large hypertext database with typed links 라는 논문을 발표
  - 1990년, NeXT 워크스테이션에서 hypertext database 시스템을 개발 → WWW 라고 명명
  - 팀 버너스 리 → 최초의 웹 브라우저와 웹 서버 개발, 연구소 내의 전화번호부와 같은 정보를 웹으로 구축
  - 스탠포드에 의해 미국에 도입된 이후 급속도로 보급 (1991)

Information Retrieval 의 역사

- 1993년 6월, MIT의 Matthew Gray 가 최초의 웹 로봇을 개발
- World Wide Web Wanderer
  - WWW 크기를 측정하기 위해 사용
  - Perl 을 기반으로 작성되었고, “Wandex” 라는 인덱스를 생성
  - 1995년 후반까지 사용

Information Retrieval 의 역사

- 1993년 10월, Martijn Koster → ALIWEB 이라는 최초의 웹 검색엔진을 개발
- Archie 와 유사한 구조
- 웹 로봇을 사용하지 않고, 웹 사이트 관리자로 부터 각 사이트의 특정한 형식의 인덱스 파일을 받아 구축
- http://www.aliweb.com/
  (최근 검색어로는 검색 X)



Information Retrieval 의 역사

- 1993년 12월 JumpStation 개발
  - 최초로 웹 페이지를 찾기 위해 웹 로봇을 사용, 스스로 인덱스 작성
  - 사용자로부터 쿼리를 받기 위해 최초로 검색 창 (web form)을 사용
  - 웹 검색엔진의 필수 구성요소(crawling, indexing, searching)을 모두 갖춘 검색엔진
  - 서비스를 종료하기 전 까지 1500개 정도의 서버를 수집

Information Retrieval 의 역사

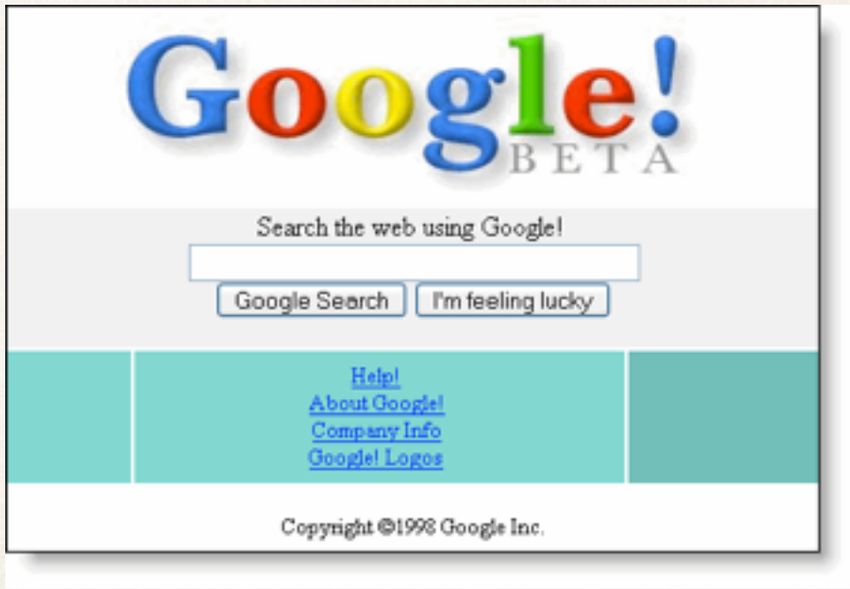
- 1994년, 스탠포드의 Jerry Yang 과 David Filo 가 Yahoo! 를 개발
  - 박사 학위 논문 작성 시 필요한 사이트들을 쉽게 찾아보기 위해 웹사이트를 분류한 후, 목록을 만듦 → Jerry and David’s Guide to the World Wide Web
  - 초기에는 대학교 내의 네트웍에 개설한 페이지 였으나, 트래픽이 마비될 정도로 인기를 끌자 따로 호스팅을 통해 본격적인 서비스 시작
  - 1995년에 야후 법인 설립
  - 다른 검색엔진과는 달리 키워드를 입력하면 야후 자체에서 제작한 웹 디렉토리를 통해 정보 검색

Information Retrieval 의 역사

- 1994년 4월, 워싱턴 대학교의 Brian Pinkerton 이 WebCrawler 를 발표
  - 최초로 전체 페이지를 인덱싱하는 crawler
  - 이를 기반으로 Lycos, Infoseek, OpenText 와 같은 여러 종류의 검색 엔진들이 개발
    → 검색엔진 역사의 시작!
- 1994년, Carnegie Mellon 에서 Lycos 개발
  - 검색어와의 관련성(relevance)를 순서대로 검색결과를 정렬 → 검색 품질 향상
  - 1996년 11월 6천만개가 넘는 문서를 인덱싱 → 당시 검색엔진 중에 최고의 성능

Information Retrieval 의 역사

- 1997년, 스탠포드의 Larry Page 와 Sergey Brin 이 구글이라는 검색엔진을 개발, 스탠포드의 서버를 통해 공개
  - PageRank 검색 알고리즘에 기반: 웹 페이지의 가치는 그 페이지를 링크한 백링크 수에 관련이 있다
- 기존의 검색엔진보다 월등히 앞선 검색 결과를 보여 줌
- 1998년 Google Inc. 라는 이름으로 창업
- 1998년 말, 6억 페이지의 인덱스 보유



검색엔진의 발전과 정보의 활용

- 1세대: Yahoo! 에 의한 디렉토리 방식
  - 야후는 제대로된 검색엔진은 현재도 가지고 있지 않음
  - 디렉토리에 기반한 정보의 분류를 통해 초기 정보의 바다에서 방황하는 사용자들을 위해 가이드를 제시
  - 정보 검색의 대상은 웹사이트의 주소 → Yellow Page
  - 초기에는 웹사이트 숫자가 적어 가능한 방식이었으나 웹사이트 숫자가 폭발적으로 늘어나게 되면서 웹사이트의 주소의 검색만으로는 정보를 제공할 수 없음
  - 다음, 네이버 등 국내 사이트 역시 초기에 비슷한 모델

검색엔진의 발전과 정보의 활용

- 2세대: 검색결과의 양적 비교 시대
  - 야후로는 원하는 정보를 찾기 어려울 정도로 정보의 양이 늘어나게 되자 각각의 페이지 내용을 찾아주는 검색엔진이 등장 → Lycos, AltaVista 등
  - 이 당시는 검색결과의 양적 비교가 검색의 품질을 대변하게 됨 → 좋은 검색엔진일 수록 찾아주는 정보가 많다! (Lycos 가 대표적인 선두주자)
  - 찾은 정보가 주어진 검색어와 어느 정도의 연관성이 있는지 여부를 평가 → Lycos

검색엔진의 발전과 정보의 활용

- 3세대: 검색결과의 질적 평가 시대
  - 웹문서의 양이 많아지자 검색된 결과에서도 원하는 정보를 정확히 찾아내는 것이 불가능해 짐 → 불필요한 문서를 구분하기 어려움
  - 검색엔진의 구조적 문제를 이용해 스팸 페이지 등이 검색결과로 많이 노출되는 문제점 발생 → 엉뚱한 문서에 검색에 자주 이용되는 키워드 삽입
  - 히트수, 방문수, 링크수, 추천수 등 다양한 방법을 통해 사람들이 많이 찾는 문서를 상위에 노출시키려는 알고리즘이 개발
  - PageRank 방식의 알고리즘을 사용한 Google 이 대표적. 국내의 검색엔진들은 대부분 2세대에 머물러 있음.

검색엔진의 발전과 정보의 활용

- 4세대: 개인화와 소셜검색?
  - 사람들이 많이 찾는 정보는 나에게도 훌륭한 정보?
    - e.g. apple price
  - 개인의 정보 검색 히스토리 등을 이용하여 검색된 정보 중에서 개인에 최적화된 정보를 선별해 줄 필요 (information filtering)
    - digital footprint 활용, 사용자의 검색 결과 평가
  - 소셜네트웍을 통해 결과물의 신뢰도 향상 (social filtering)
    - 음식점 리뷰, 영화평, 정치 컬럼 등

정보검색 시스템 구성요소

- 정보검색시스템
  - 정보 수요자가 필요하다고 예측되는 정보나 데이터를 미리 수집, 가공, 처리하여 찾기 쉬운 형태로 저장해 놓은 DB로 부터 정보를 신속하게 찾아내어 정보 수요자에게 제공하는 시스템

정보검색 시스템 구성요소

- 웹로봇
  - 웹에 존재하는 문서들을 가져오기 위해, 웹의 하이퍼텍스트 구조를 자동으로 추적하여 참조되어지는 모든 문서들을 검색하는 프로그램.
  - Crawler, spider, gatherer, worms, ants 등 다양한 이름으로 불린다.
  - 검색엔진의 가장 중요한 요소로, WWW 상의 모든 웹 문서의 인덱스 작업을 하는 것을 목적으로 한다
  - 웹로봇은 바이러스와 같은 역할을 할 수 있어서 사용에 주의해야 하며, robot exclusion standard 를 따라 로봇을 배제하고자 하는 경우 이를 지켜 주어야 한다.
    → naver 나 daum cafe 의 경우 (contents 및 privacy 보호)

정보검색 시스템 구성요소

- 스토리지
  - 검색엔진에서 사용할 색인용 데이타를 저장해 놓는 공간
  - 초기 검색엔진은 일반적인 DBMS 를 이용하여 데이터 저장 → 데이타가 대용량화 되고 알고리즘이 복잡해지면서 점차 파일시스템을 직접 제어하여 데이터를 저장하는 방식을 많이 사용 (전문 파일시스템 등장)
- 색인기
  - 웹로봇을 통해 수집된 문서들을 빠르고 정확하게 검색하기 위해 문서의 중요 키워드를 추출하고 키워드, 문서 간의 상관관계를 정의하여 스토리지에 저장하는 프로그램

정보검색 시스템 구성요소

- 형태소 분석기
  - 형태소 분석: 하나의 어절에서 의미를 갖는 최소 단위인 각 형태소를 분석해 내는 것.
  - 문서의 핵심 키워드를 추출하는 기본적인 시스템
    - 예: “영희가 소설책을 읽는다”
      - 자립형태소: 영희, 소설, 책
      - 의존형태소: 가, 을, 읽, 는, 다
      - 실질형태소: 영희, 소설, 책, 읽
      - 형식형태소: 가, 을, 는, 다
  - 초기 정보검색은 자립형태소 등을 주로 사용했으나, 최근에는 형태소의 구조적인 관계 및 의미관계까지 고려한 색인어를 추출
    → 자연어 검색으로 발전

정보검색 시스템 구성요소

- 스테머 (stemmer)
  - 어근 추출. 영어권에서 많이 사용.
  - 한국어 같은 경우, 어미변화가 심해 스테밍 알고리즘으로 처리 곤란 → 형태소 분석 주로 사용
- 검색기
  - 사용자가 입력한 질의를 색인기가 생성하여 둔 스토리지에서 정의된 검색모델에 의해 가장 유사한 문서를 추출하여 검색하여 주는 기능

정보검색 시스템 구성요소

- 랭커(ranker)
  - 검색 결과에 대한 순위를 매겨주는 기능.
  - 검색 모델 및 랭킹 알고리즘에 의해 검색된 결과물이 주어진 질의에 얼마나 부합하는지를 수치화 하는 것
- 질의 분석기
  - 사용자가 입력한 비 정규적인 질의를 파싱하여 검색에 적합하도록 정규화된 질의 문법으로 변환하는 기능
  - 형태소 분석기나 기타 분리기 등을 통해 질의에 꼭 필요한 키워드만 추출할 수 있도록 한다. (and/or 등의 논리 연산자 등으로 표현하기도 함)

검색기법

- Belkin, Nicholas J. and Croft, W. Bruce,"Retrival Techniques," ARIST, 22(1987), PP. 109-145
- 검색기법의 분류
  - Exact match
  - Partial Match
    - Individual, Feature-based
    - Individual, Structure-based
    - Network

검색기법

- Exact match techniques (완전일치 기법)
  - 초창기 대규모 검색시스템에서 사용하던 방식으로, 검색된 문서의 표현이 질문의 표현과 완전하게 일치된 것만 검색하는 방식
  - 공사계약서 → ‘건축하도급 공사계약서’, ‘인테리어공사 계약서 작성 시 주의사항’
  - 문제점
    - 이용자의 질의에 부분적으로 일치되는 경우의 텍스트들은 그 유용성에도 불구하고 검색에서 누락되는 경우가 발생
    - 검색된 텍스트들에 대한 적절성의 순위를 정할 수 없다

검색기법

- Partial match techniques (부분일치 기법)
  - 검색된 문서들의 표현이 질문의 표현과 부분적으로 일치된 경우의 것들을 검색하는 방식
    - 공사계약서 → ‘공사현장 감독계약서’ ‘건축공사 시방서’ ‘인테리어 공사를 하려는데 계약서는 어떻게 작성해야 하나요?’
    - individual: 질의를 개별 문서들과 비교하는 방식
    - network: 질의를 개별 문서들과 관련된 다른 문서들 간의 관계성에 중점을 두어 비교하는 방식

Questions?