INTRODUCTION TO LIS AND TEXT RETRIEVAL

연세대 문헌정보학 석사과정 홍수린

INTRODUCTION TO LIS

- 문헌정보학(Library and Information Science)
- 문헌정보학의 세부 갈래
- iSchools

문헌정보학Library and Information Science

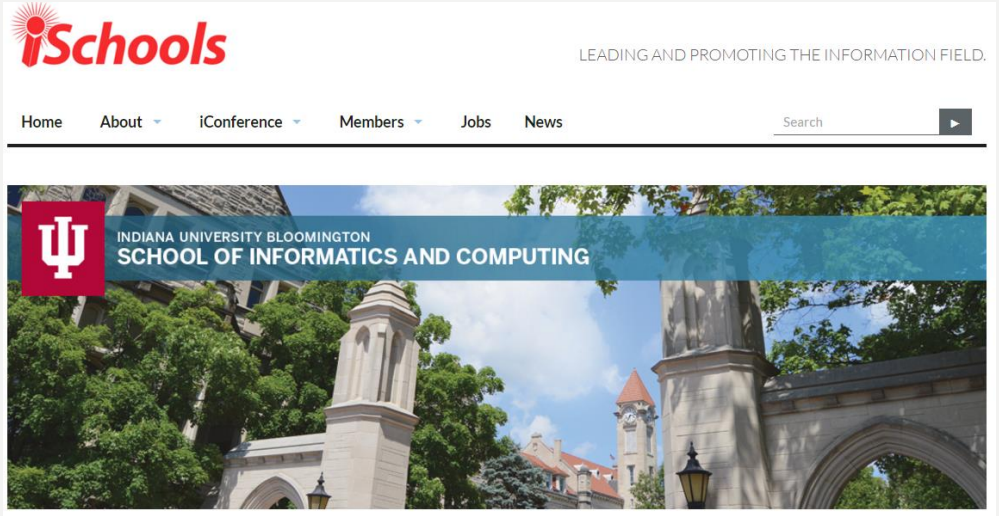
- 정보의 발생부터 수집·정리·분석·보존·축적·이용까지 정보에 관련된 이론과 원리, 방법과 기술을 과학적으로 연구하는 학문
- 대한민국의 경우 전통적인 도서관학과 1960년대 이후 발달한 정보학, 도서의 고증과 해석을 중심으로 하는 서지학, 기록물의 관리와 보존에 대한 기록관리학이 합친 학문

문헌정보학의 세부 갈래

- 정보 시스템
  - 정보검색(Information retrieval)
  - 데이터베이스
  - 멀티미디어 인터페이스
- 기록관리 및 시스템 경영
  - 기록관리학(archives management)
  - 도서관경영론(library management)
  - 디지털도서관
  - 서지학(Bibliography)
  - 자료조직론(Information organizatoin: Cataloging, Classification)
  - 장서관리·개발론 (Collection management · development)
- 정보제공
  - 정보 이용자 연구
  - 정보행동(Information behavior)
  - 참고봉사(Reference service)
- 정보과학
  - 텍스트 마이닝(Textmining)
  - 온톨로지(Ontology)
  - 사회정보학(Social informatics)
  - 학술정보커뮤니케이션(Scholarly communication)

iSchools

- 정보학 분야를 선도하는 인재 양성을 위한 교육기관(Information school) 연합체
- 정보기술(information technology), 도서관 과학(library science), 정보학(informatics), 정보과학(information science) 등 다양한 트랙(track)
- 컴퓨터과학, 정보학, 사회학 등 다양한 배경을 가진 학생들이 입학
- 세계적으로 65개 교(school, college, department) 참여
- 한국에서는 서울대학교, 성균관대학교, 연세대학교 3개 교 참여
- http://ischools.org/



INTRODUCTION TO TEXT RETRIEVAL

- 정보검색(TREC, 정보검색모형)
- 자동색인
- 검색 성능 평가(재현율, 정확률)
- 정보검색모델: VSM, 확률 모델
- 랭킹알고리즘: 페이지랭크

정보검색Information retrieval

- 정보검색: 다양한 정보원으로부터 이용자의 정보요구에 적합한 정보/지식을 찾아내기까지의 모든 과정*
- 1990년대 이후 웹이 정보검색의 대중화를 가져왔으며 이에 따라 정보검색의 내용도 크게 변화
- 검색(retrieval), 마이닝(mining), 필터링(filtering), TDT(Topic Detection and Tracking), …
- 자연언어 텍스트 처리를 기반으로 하는 여러 텍스트 분석 task 중 하나
- 검색결과의 정확률과 재현율을 향상시키는 전통적 task도 여전히 연구되고 있지만, 1990년대 중반 이후부터 이용자의 정보 요구를 만족시켜줄 수 있는 다양한 검색 관련 서비스들이 많이 연구되는 추세

* 정영미. (2012). 정보검색연구 증보판. 서울: 연세대학교 출판문화원.

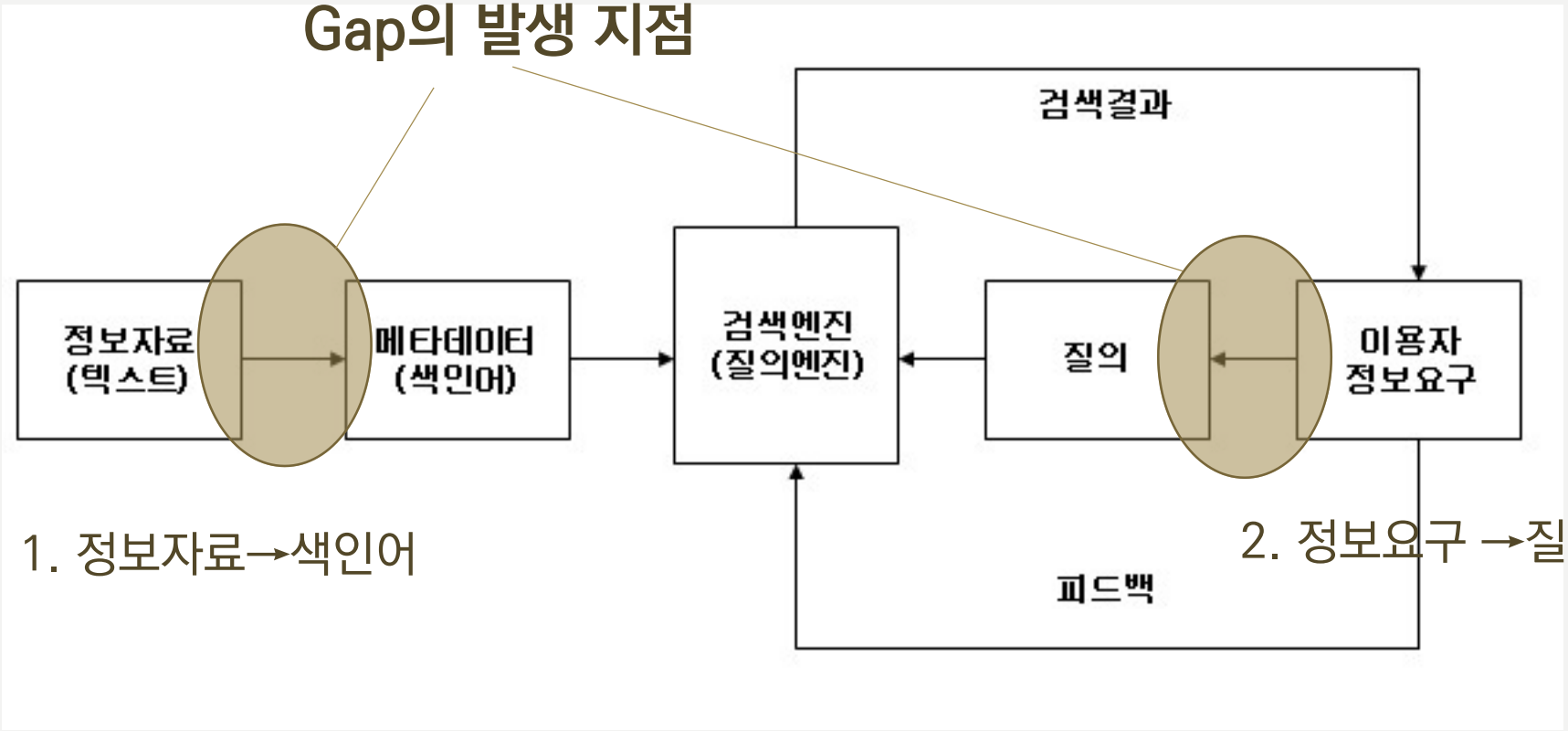
TREC(Text REtrieval Conference)

- 1992년에 처음 시작된 text retrieval 관련 컨퍼런스
- 대규모의 실험 집단을 대상으로 수행된 검색 실험 결과를 비교 평가하고 의견을 공유하는 자리
- 초기에는 정적인 데이터를 대상으로 새로운 질의를 처리하는 ad hoc task(소급 검색)과 새로운 데이터를 대상으로 똑같은 질의를 처리하는 routing task(최신정보 검색) 두 가지 트랙(track)에서 연구 수행
- 1995년부터 텍스트 검색의 하부 문제에 초점을 맞추는 여러 트랙들이 추가되기 시작하고 일부는 종료됨
- 2016년 현재 8개의 live track 존재
- http://trec.nist.gov/

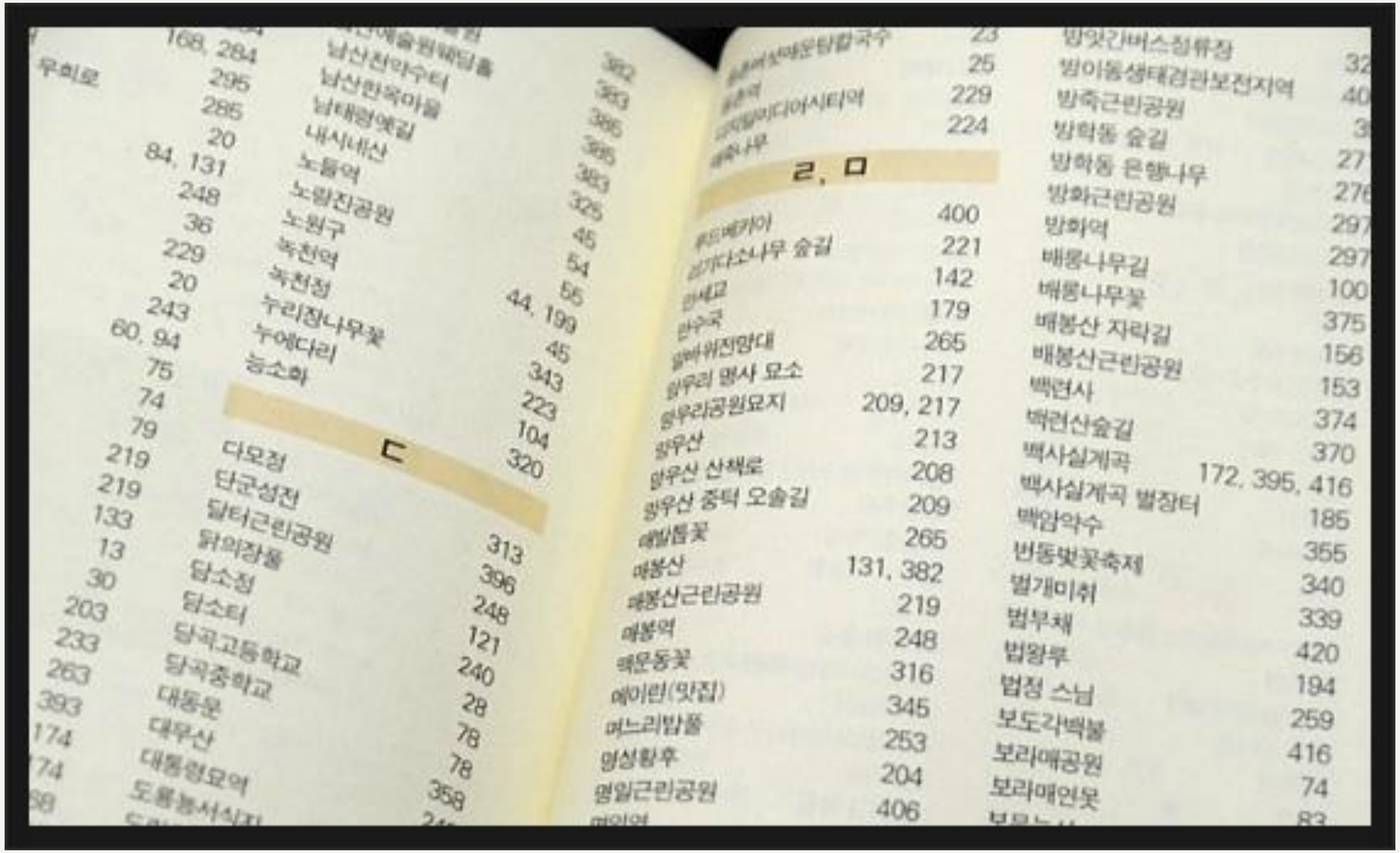
TREC's tracks(2016)

Past tracks (25~)	Live tracks (8)
ad hoc, routing task(1992)	Clinical Decision Support Track
Filtering Track(1995)	Contextual Suggestion Track
Cross-Language Track(1997)	Dynamic Domain Track
Question Answering Track(1999)	Live QA Track
Web Track(1999)	OpenSearch Track
Video Track(2001)	Real-Time Summarization Track
Novelty Track(2002)	Tasks Track
Genomics Track(2003)	Total Recall Track
Terabyte Track(2004)	
Microblog Track(2006)	
Legal Track(2006)	
Crowdsourcing Track(2011)	
…	

정보검색 모형


Gap의 발생 지점
검색결과
정보자료
(텍스트)
메타데이터
(색인어)
검색엔진
(질의엔진)
질의
이용자
정보요구
1. 정보자료→색인어
2. 정보요구 →질의어
피드백

자동색인



다모정	313
단군성전	396
달터근린공원	248
닭의장풀	121
담소정	240
담소터	28
당곡고등학교	78
당곡중학교	78
대동문	358

ㄹ, ㅁ

리기다소나무 숲길	221
만수국	179
말바위전망대	265
망우리 명사 묘소	217
망우리공원묘지	209, 217
망우산	213
망우산 산책로	208
망우산 중턱 오솔길	209
매발톱꽃	265
매봉산	131, 382
매봉산근린공원	219
매봉역	248
매이란(맛집)	345
며느리밥풀	253
명성황후	204
명일근린공원	406
방화역	297
배롱나무길	297
배롱나무꽃	100
배봉산 자락길	375
배봉산근린공원	156
백련사	153
백련산숲길	374
백사실계곡	370
백사실계곡 발장터	172, 395, 416
백암약수	185
번동벚꽃축제	355
벌개미취	340
범부채	339
법왕루	420
법정 스님	194
보도각백불	259
보라매공원	416
보라매연못	74

자동색인(cont.)

- 색인(indexing): 개개의 정보자료의 특성을 표현하는 데이터 요소를 추출하여 각 정보자료를 표현하는 작업. 색인 결과 추출된 데이터 요소를 색인어(index term) 또는 메타데이터(metadata)라고 함
- 웹 검색엔진이나 온라인 DB에서는 색인의 결과로 색인 DB가 생산됨
- 자동색인(automatic indexing): 컴퓨터에 입력된 문헌의 텍스트를 분석한 후 문헌의 내용을 대표할 수 있는 단어나 단어구를 일정한 기준에 의해 추출하여 색인어로 선정
  1. 단어들을 주제어와 비주제어로 구분
  2. 주제어를 모두 색인어로 선택하거나 이들 가운데 핵심 주제어만을 색인어로 선정
  3. 선정된 색인어에 가중치를 반영하지 않는 binary model과 이를 반영하는 weight model로 나뉨
- 가중치 표현의 예시: $w_{ij} = tf_{ij} * log(N/df_j)$
- 문헌 표현의 예시: $D_i = (t_1, w_{i1}; t_2, w_{i2}; \ldots; t_n, w_{in})$

자동색인: 통계적 특성에 기반한 모델

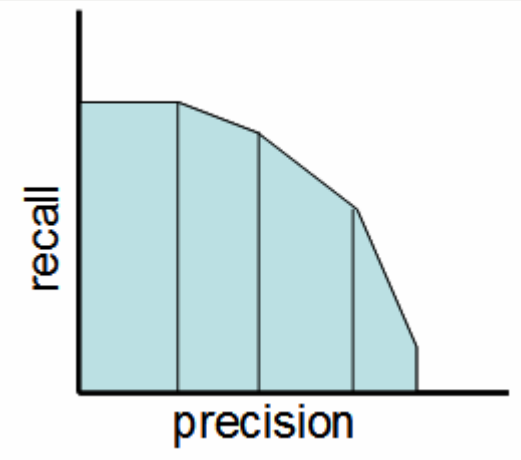
- 텍스트에 출현한 단어의 통계적 특성에 의해 추출하는 Luhn(1957)의 기법이 오늘날 상업적 시스템에서 가장 보편적으로 사용됨
- 너무 빈번히 나타나는 고빈도 단어는 일반적 단어이므로 주제어로서의 가치 없음
- 너무 빈도가 낮은 저빈도 단어 또한 무의미어로 분류
- 출현빈도/출현확률에 근거: 문헌분리값(term discrimination value), 신호량 가중치(signal weight), 적합성 가중치(relevance weight) 등
- 출현빈도에 따른 확률분포 이용: 포아송 분포 모형, 2-포아송 분포 모형, 점유분포(occupation distribution) 기반 단어 집중도 모형 등

검색 성능 평가: 재현율, 정확률

- 재현율(recall) = $\frac{\text{검색된 적합문헌 수}}{\text{총 적합문헌 수}}$
- 정확률(precision) = $\frac{\text{검색된 적합문헌 수}}{\text{검색된 문헌 수}}$

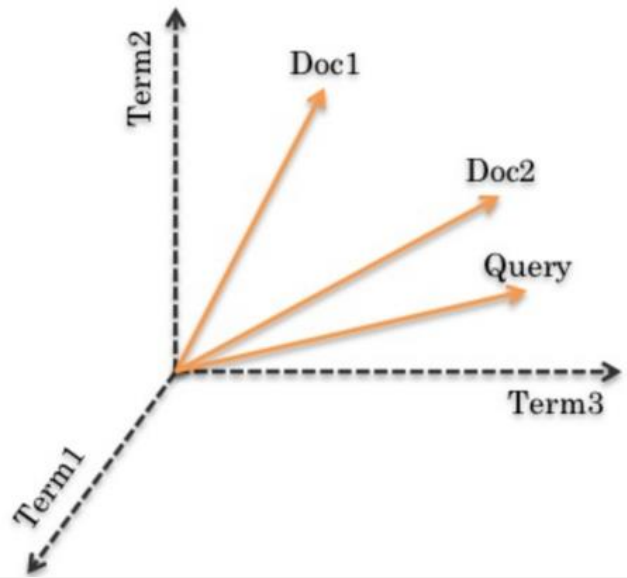
$$F_1 = \frac{2P\mathrm{R}}{P + \mathrm{R}}$$

- 재현율은 검색의 완전성을, 정확률은 정확성을 측정하는 척도, F-measur는 위의 F_1 주로 사용
- 실제 운영 중인 시스템에서 재현율을 측정하기는 매우 어려움: 가상의 적합문헌 집단을 구성하여 '상대적 재현율' 파악
- 재현율과 정확률은 일반적으로 반비례 관계.
  시스템의 목적에 따라 최적화를 시켜야 함



정보검색모형: 벡터공간모델(VSM)

- 문헌과 질의를 각각 용어 벡터 형태로 표현한 다음 두 벡터 간의 유사도를 산출하여 검색문헌을 순위화
- 문헌 DB 내 색인어의 수 N과 같은 N-차원 벡터로 문헌과 질의를 표현
- 색인어와 질의가 각각 가중치를 가지며, 질의와 부분적으로 일치하는 문헌들 검색 가능
- 유사도는 보통 코사인 유사계수 사용, 좀더 단순한 내적계수를 사용하기도 함



$$Cos(D_i, Q) = \frac{\sum_{j=1}^{t} d_{ij} * q_j}{\sqrt{\sum_{j=1}^{t} d_{ij}^2 * \sum_{j=1}^{t} q_j^2}}$$

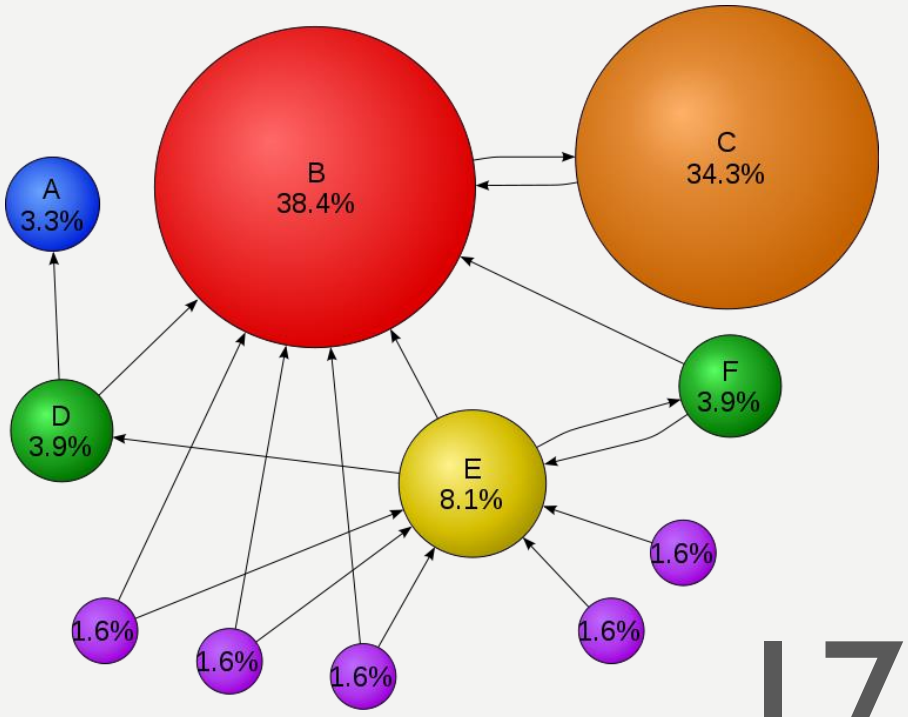
정보검색모형: 확률(probabilistic) 검색

- 문헌과 질의 간 유사도를 문헌이 질의에 적합할 확률로 검색
- 특정한 질의에 대해 각 문헌이 적합할 확률과 부적합할 확률을 산출하여 적합할 확률이 부적합 확률보다 큰 문헌을 검색함
- W_1(적합한 경우)와 W_2(부적합한 경우)로 나타낼 경우 문헌 X가 적합할 확률은 $P(W_1|X)$, 부적합할 확률은 $P(W_2|X)$가 됨
- 그런데 이 확률은 직접 산출하기 어려우므로 베이즈 정리를 이용함(확률검색의 기반)

$$P(X|w_1)P(w_1) > P(X|w_2)p(w_2) \quad \rightarrow \quad \frac{P(X|w_1)}{P(X|w_2)} > \frac{p(w_2)}{P(w_1)}$$

랭킹알고리즘: 페이지랭크(pagerank)

- 오늘날의 구글 검색엔진의 핵심이 되는 알고리즘(1998*)
- 인터넷이 발달함에 따라 웹 페이지 사이의 연결 관계가 검색 순위화에 유용하게 사용될 수 있음에 주목하고 기존의 학술지 인용 관계(citation relationship)를 하이퍼링크 관계에 적용
- 어떤 페이지 A의 페이지 랭크는 그 페이지를 인용하고 있는 다른 페이지 T1, T2, T3, .. 가 가진 페이지 랭크를 '정규화시킨(normalize)' 값의 합
- 영향력 있는 페이지가 인용해 줄수록 내 페이지의 페이지랭크가 올라감
- 소위 '불펌'이 만연하는 곳에서는 이 알고리즘은 제대로 기능하지 못함(원문의 링크를 걸지 않는 경우)



* Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, *56*(18), 3825-3833.

랭킹알고리즘: 페이지랭크(cont.)

$$PageRank\ of\ site = \sum \frac{PageRank\ of\ inbound\ link}{Number\ of\ links\ on\ that\ page}$$

OR

$$PR(u) = (1-d) + d \times \sum \frac{PR(v)}{N(v)}$$

- Damping factor: 어떤 마구잡이로 웹서핑을 하는 사람이 그 페이지에 만족을 못하고 다른 페이지로 가는 링크를 클릭할 확률
- damping factor가 1이면, 무한히 링크를 클릭한다는 뜻이고, 0이면 처음 방문한 페이지에서 무조건 멈추고 더 이상 클릭하지 않는다는 뜻
- 논문에서는 실험을 통해 주로 0.85의 값을 사용한다고 함(85%의 확률로 다른 페이지를 클릭해볼 것이라는 뜻. 이 경우 15%의 확률에 걸리는 순간 클릭을 멈추고 그 페이지를 살펴보게 됨)

THANK YOU ☺