정보검색 개요

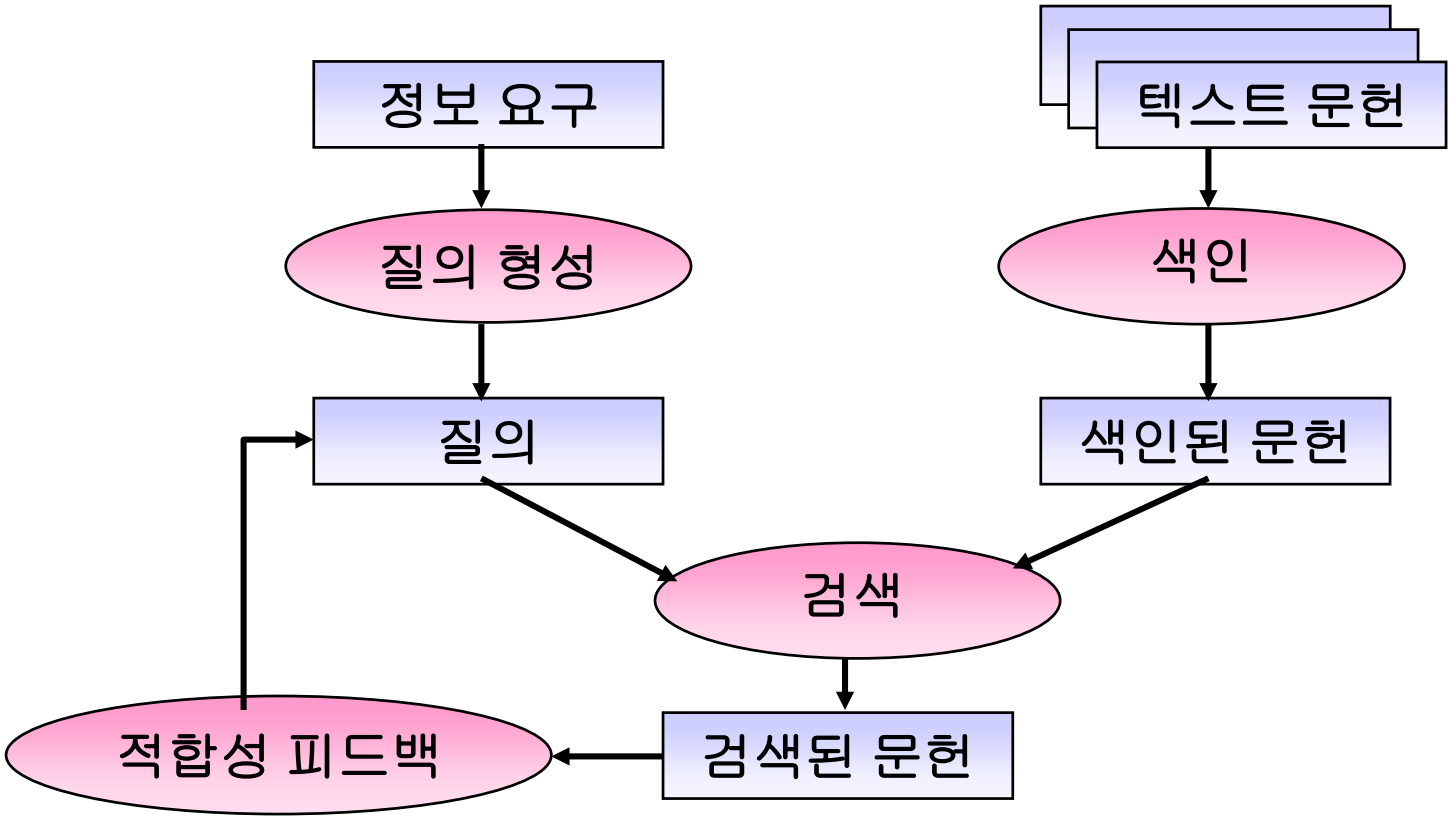
국민대학교 소프트웨어학부
강 승 식

검색 엔진의 필요성

- 정의
  - 사용자가 필요로 하는 정보 수집-저장
  - 효율적인 검색을 위한 색인
  - 검색 요구에 적합한 정보 검색 및 제공
- 고려사항
  - 인터넷 상의 문서의 수의 폭발적인 증가로 인해 검색 대상 문서의 수가 방대
  - 사용자 질의에 대한 빠른 응답시간 요구

정보검색 과정: 검색 엔진의 구조

- 구성도



색인 (indexing)

- 정보자료(문서)의 내용을 표현하는 “색인어”로부터 색인어가 포함된 “문서”를 직접 검색하기 위한 작업

- 필요성
  - 정보자료의 탐색시간 최소화
  - 이용자에게 빠른 속도로 검색 결과 제공

- 색인 과정
  - 입력 문서들로부터 색인어 추출
  - 역파일 구조로 색인 정보 저장 (색인 DB 생성)
  - 일괄처리(전체문서를 한꺼번) 또는 점증 방식으로 수행

색인 과정

1. 각 문서에서 색인어 추출 → forward indexing
   - stopword 제거
   - stemming
   - 색인어의 출현 빈도(TF: Term Frequency) 계산

문서(docID)	색인어 및 가중치
001	(병렬,4), (시스템,3), (특성,1), (설계,2), (연구,3)
002	(정보,2), (검색,3), (시스템,3), (연구,3), (성과,1)
003	(프로그램,4), (시스템,3), (설계,4), (성능,1), (향상,1)
004	(병렬,2), (프로그램,2), (연구,6)

색인 과정

2. Inverted File 생성

- <문서,색인어 list> 대응관계를 <색인어, 문서 list> 형태로 재구성(색인어 순으로 sorting)
- 색인어별 문서 빈도(DF: Document Frequency) 계산
- TF-IDF를 고려하여 각 문서의 색인어 가중치 계산 (Weight = TF * IDF)

색인어	문서번호 및 가중치
검색	(002, 3)
병렬	(001, 2), (004, 1)
설계	(001, 1), (003, 2)
성과	(002, 1)
성능	(003, 1)
시스템	(001, 1), (002, 1), (003, 1)
연구	(001, 1), (002, 1), (004, 2)
정보	(002, 2)
특성	(001, 1)
프로그램	(003, 2), (004, 1)
향상	(003, 1)

Inverted File 저장 구조

Term Table

색인어	시작위치	문서개수
검색	0	1
병렬	1	2
설계	3	2
성과	5	1
성능	6	1
시스템	7	3
연구	10	3
정보	13	1
특성	14	1
프로그램	15	2
향상	17	1

Posting File(문서 list)

0	0023	0012	0041	0011	0032
5	0021	0031	0011	0021	0031
10	0011	0021	0042	0022	0011
15	0032	0041	0031		

"시스템"이 어떤 문서에 나타나는지를 알아보려면 출현문서 list의 7번째로부터 3개를 참조

"시스템"이라는 용어는 001, 002, 003 문서에 출현

색인 과정 (종합)

1. 각 문서들에 대해 순서대로 docID 부여 및 색인어 추출

   중복 색인어 제거 및 TF 계산

   (docID, <색인어, TF> <색인어, TF>, …)

2. 추출된 색인어들을 (색인어, docID, TF) 형태로 변환 및 색인어 순으로 sort & merge

   (색인어, <docID, TF> < docID, TF>, …)

3. 각 색인어들에 대한 DF 계산 및 Term Table 작성

   Term Table : TermID, DF

   Posting file : <docID, TF>

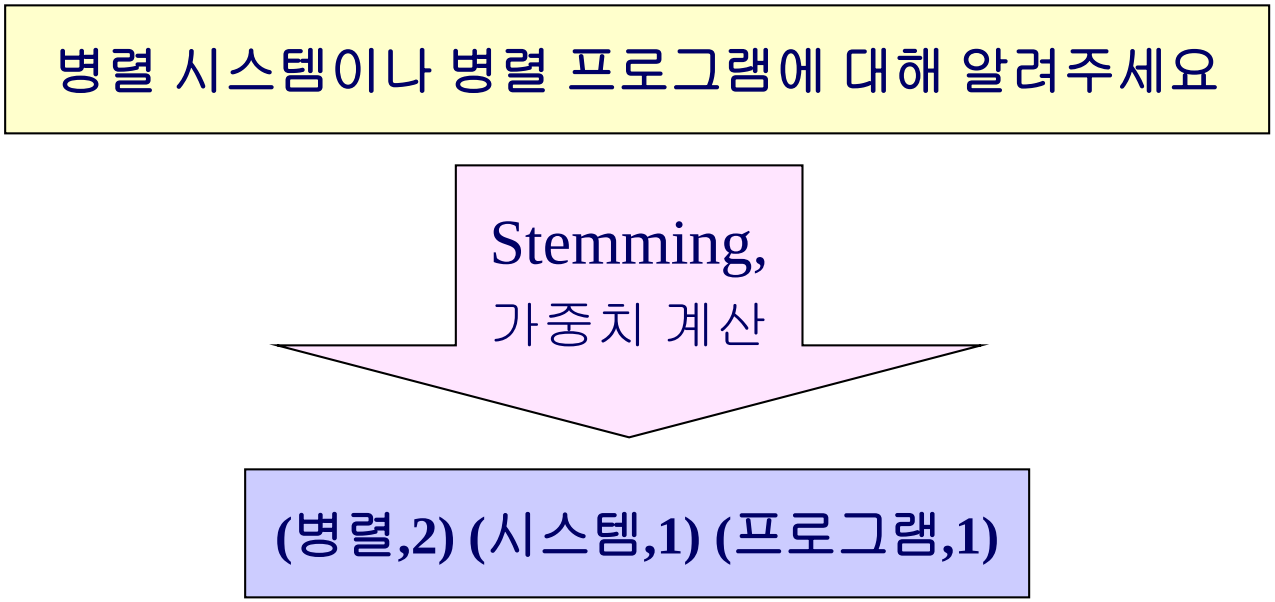
4. 실제 검색 엔진에서는 TF와 더불어 '위치정보'(색인어의 출현 위치)를 함께 저장

검색

- 정의
  - 사용자 질의를 분석하여 질의 내용에 적합한 문서를 찾는 과정
- 검색 단계
  - 사용자 질의 분석 → 시스템 질의 형성
  - 시스템 질의와 각 문서들과의 유사도 계산
  - 유사도 순으로 각 문서를 정렬
- 검색 모델
  - 벡터공간 모델
  - 불린 모델

질의 분석

- 질의 분석
  - 사용자 질의를 분석하여 시스템 질의를 형성



검색 (불린 모델)

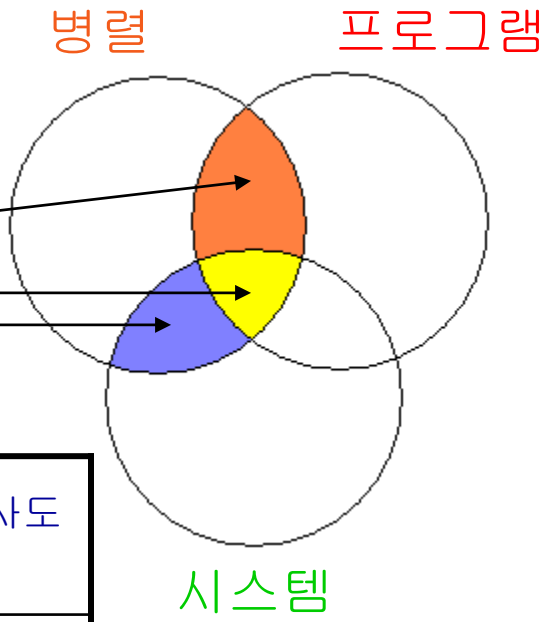
- 집합 이론에 기반
- 불린식으로 질의를 표현
  - 각 term들을 and, or, not의 연산자로 연결
- 모든 색인어들에 0 또는 1의 가중치 할당
  - 문서안에 해당 색인어가 있으면 1, 없으면 0
- 정확한 match를 통한 검색을 수행
  - 불린식 전체를 만족하는 문서만을 검색
  - 문서의 유사도는 1 또는 0
  - 검색된 문서를 ranking 하지 못함

$$sim(d_j, q) = \begin{cases} 1, & \text{불린식을 만족할 경우} \\ 0, & \text{만족하지 못할 경우} \end{cases}$$

검색 (불린 모델)

$$q = 병렬 \wedge (프로그램 \vee 시스템)$$

$$\vec{q}_{dnf} = (1,1,1) \vee (1,1,0) \vee (1,0,1)$$



문서	색인어				유사도
	병렬	프로그램	시스템	…	
001	1	0	1	…	1
002	0	0	1	…	0
003	0	1	1	…	0
004	1	1	0	…	1

검색 (벡터공간 모델)

- 질의와 문서를 모두 *t*차원의 벡터로 표현
  - *t*: 전체 문서집합 내에 존재하는 서로 다른 색인어의 수
  - 각 벡터의 원소에 w_i만큼의 가중치를 할당

$$\vec{q} = (w_{1q}, w_{2q}, \ldots w_{tq}) \qquad w_{iq} \geq 0$$

$$\vec{d}_j = (w_{1j}, w_{2j}, \ldots w_{tj}) \qquad w_{ij} \geq 0 \qquad (w_{ij} = tf \times idf)$$

- 질의 벡터와 각 문서 벡터의 cosine 유사도 계산
  - 유사도 순으로 검색된 문서를 정렬

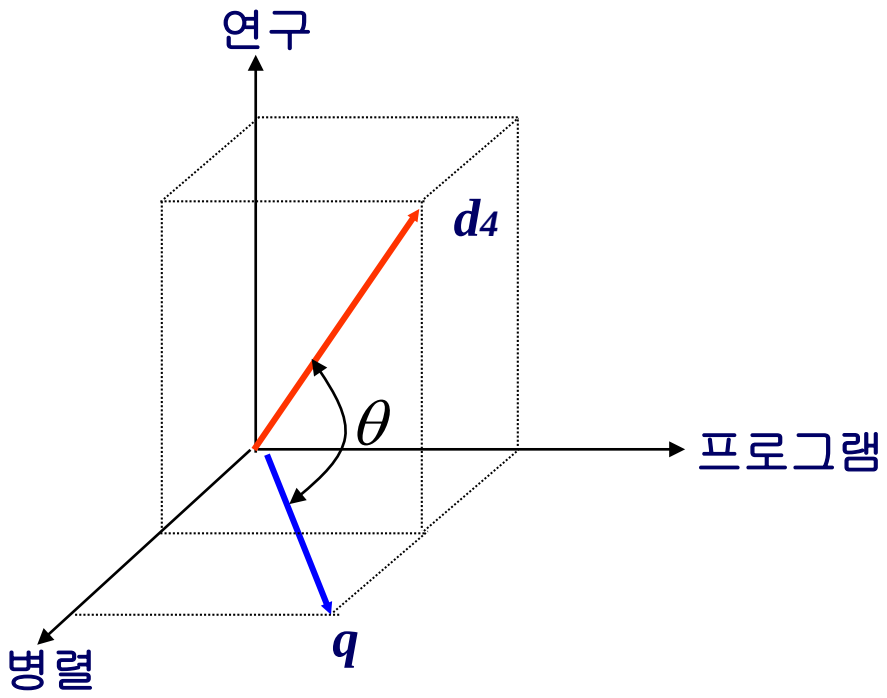
$$sim(d_i, q) = \cos\theta = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|}$$

검색 (벡터공간 모델)

<문서 004>
병렬, 프로그램, 연구

<질의 *q*>
병렬, 프로그램

	가중치		
	병렬	프로그램	연구
질의 q	2	1	0
문서 d_4	1	1	2



$$\vec{q} = (2,1,0)$$
$$\vec{d}_4 = (1,1,2)$$

$$\text{유사도} = sim(\vec{d}_4, \vec{q}) = \frac{\vec{d}_4 \cdot \vec{q}}{|\vec{d}_4| \times |\vec{q}|} = \frac{3}{\sqrt{6} \times \sqrt{5}}$$

적합성 피드백

- 정의
  - 적합 또는 부적합 하다고 판단된 문서들 중에서 중요한 단어들을 추출하고, 이를 이용하여 질의를 재구성하여 다시 검색하는 방법
- 동기
  - 사용자의 정보요구를 정확한 질의로 표현하기 어려움
  - 질의어와 동일 개념의 용어라도 다른 유사 용어로 색인된 문서를 검색할 수 없음 (예. 질의어:*선생님* - 색인어:*교사*)
  - 검색된 문서의 정확도를 높이기 위해 사용
- 질의 재구성 방법
  - 질의 확장
  - 가중치 재부여

사용자 적합성 피드백

- 벡터공간 피드백

<질의 q>
병렬(0.5) 프로그램(0.5)

검색된 문서	문서 벡터
	용어(가중치)
1	병렬(0.3) 프로그램(0.2) 시스템(0.5) …
2	병렬(0.7) 프로그램(0.0) … 처리(0.5) …
…	…
99	병렬(0.1) 프로그램(0.1) 시스템(0.9) …

사용자가 진정으로 원하는 문서

<수정된 질의 q_m>
병렬(0.4) 프로그램(0.4) 시스템(0.2)

검색된 문서	문서 벡터
	용어(가중치)
1	병렬(0.3) 프로그램(0.2) 시스템(0.5) …
2	병렬(0.2) 프로그램(0.1) 시스템(0.9) …
…	…

자연어 검색 시스템

- 목표

검색어 : 대한민국 15대 대통령 선거에 당선된 사람은 누구인가?

결과 : 대한민국 15대 대통령 선거에 당선된 사람은 김대중 대통령이다.

- 현실

검색어 : **토끼**는 무엇을 먹고 사는가?

결과 1 : **토끼**와 거북이가 달리기 시합을 하면 **토끼**가 빠르고……

결과 2 : (**토끼**전) **토끼**의 간을 먹으면 낫는 병이 있습니다.

결과 3 : 철수네 **토끼**가 쥐약을 먹고 죽었다.

결과 4 : **토끼**는 당근을 먹으면서 산다.