11. 텍스트를 위한 화일

❖ 역 리스트 화일

◆ 역 리스트 화일 구조

- 역 리스트 화일

  = 인덱스 화일 + 포스팅 화일 + 데이타 화일

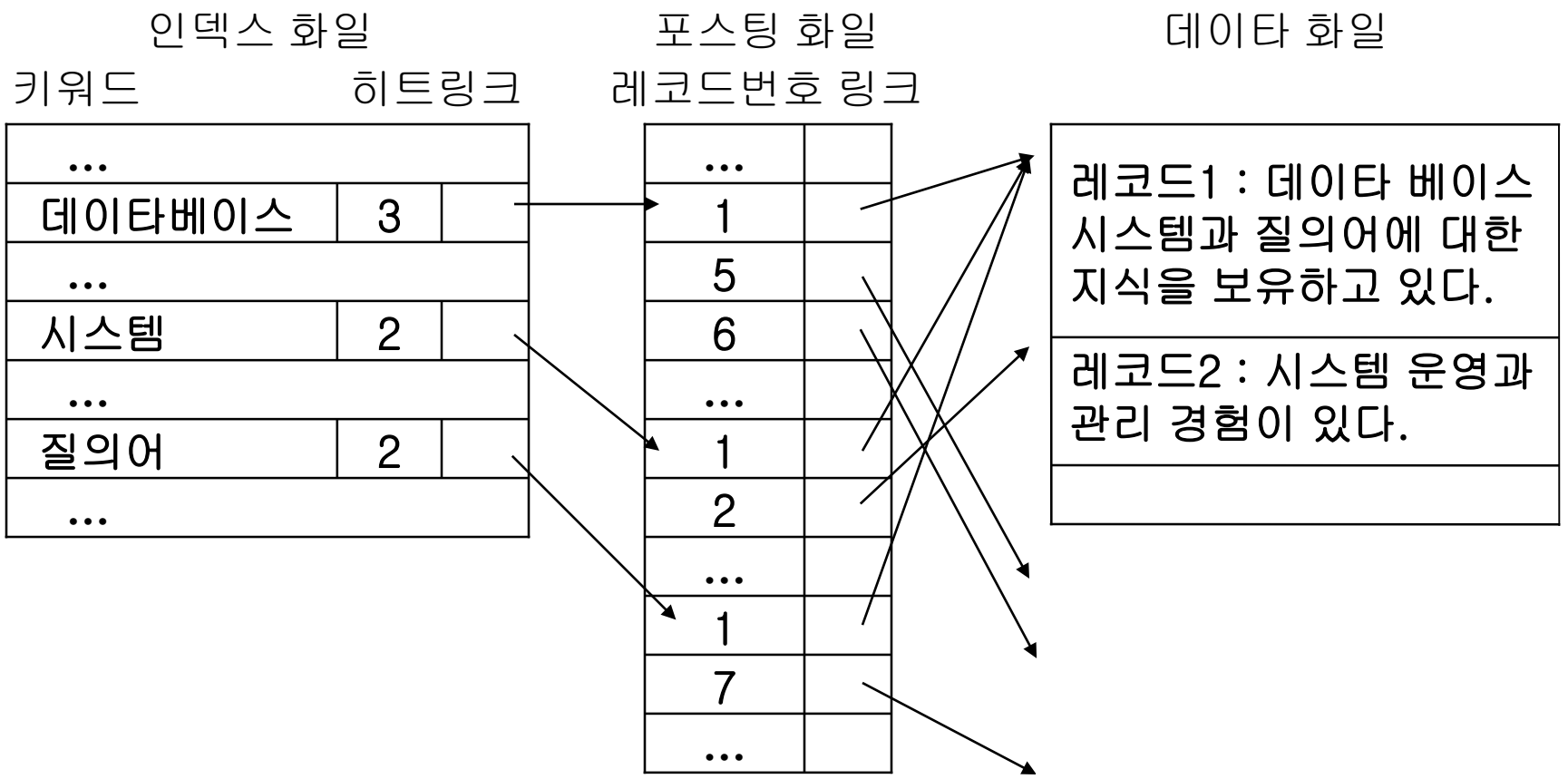
  - 인덱스 화일 :

    키워드 + 레코드 수(히트 수) + 포스팅에 대한 포인터
  - 포스팅 화일 :

    키워드를 포함한 데이타 레코드에 대한 포인터의 리스트

- 장점
  - 구현 용이
  - 속도가 빠름
  - 동의어 지원 용이
- 단점
  - 많은 기억 공간 요구
  - 인덱스 갱신 재구성시 많은 비용 요구

▶ 역 리스트 화일 구조의 예


인덱스 화일
키워드
히트링크
...
데이타베이스
3
...
시스템
2
...
질의어
2
...
포스팅 화일
레코드번호 링크
...
1
5
6
...
1
2
...
1
7
...
데이타 화일
레코드1 : 데이타 베이스 시스템과 질의어에 대한 지식을 보유하고 있다.
레코드2 : 시스템 운영과 관리 경험이 있다.

▶ 역 리스트 화일의 탐색 방법

- **불리언 질의**
  - 논리 연산자('&', '|', '!')를 이용하여 표현
    Ex) '데이타베이스 & 질의어'

- **랭킹 질의**
  - 유사도 휴리스틱을 이용하여 일정한 개수의 근접 데이타 레코드들을 결과로 가져옴

❖ 시그니처 화일

- **시그니처**
  - 부정확한 필터
  - 데이타베이스의 내용을 부호화한 것

- **검색 과정**
  1) 데이타베이스의 내용을 부호화(시그니쳐)
  2) 시그너처가 조건 만족하는 것 우선 선택
  3) 2의 결과에서 실제 내용에 대한 검색

▶ 중첩 코딩 기법의 예

단어(키워드)	시그니처
데이타베이스	**0010 0100 0000 0001**
시스템	**0000 1000 1000 0010**
질의어	**0100 1000 0100 0000**
블록 시그니처	**0110 1100 1100 0011**

▶ 시그니처 화일을 이용한 검색

– 질의 : '시스템'이 포함된 텍스트를 검색하라

1) 질의 시그니처 생성 0000 1000 1000 0010

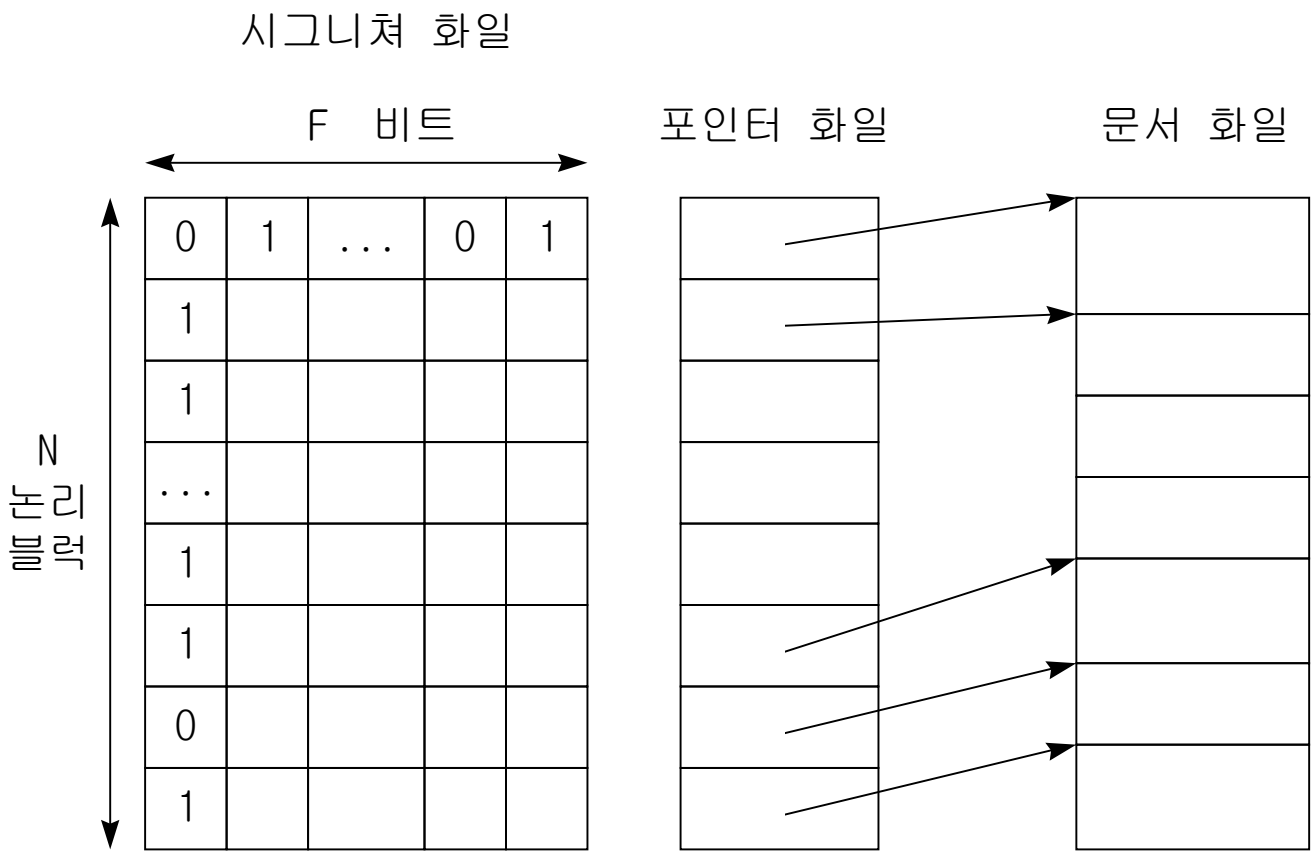
2) 질의 시그니처 0000 1000 1000 0010
AND) 블록 시그니처 0110 1100 1100 0011

0000 1000 1000 0010
(≡ 질의 시그니처)

3) 2와 같은 조건을 만족하는 텍스트에 대해 '시스템'이 실제로 포함되었는지 검사

▶ 시그니처 화일 구조 및 탐색 방법

(1) 순차 시그니처 화일



(2) 압축

- 시그니처 행렬 희소한 경우 압축
- 저장 공간 감소하므로 검색 시간 단축

(3) 수직 분할

- 열단위(bit slice)로 나누어 저장
- 1로 세트된 비트 슬라이스만 검색
- 삽입 시간은 증가

(4) 수평 분할

- 유사 시그니처 그룹 또는 시그니처 화일에 대한 인덱스 사용
- 전체 시그니처 화일 검색 불필요