ETLARS 시스템에서의 데이터베이스 설계 및 생성에 관한 연구

Database Design and Production in ETLARS System

김상도(Sang-Do Kim)*, 박계숙(Kay-Sook Park)*
김희섭(Hee-Seop Kim)*, 홍기채(Gi-Chai Hong)*
우동진(Dong-Chin Woo)**, 기민호(Min-Ho Key)***

□ 목 차 □

초 록

DB 구축 과정은 데이터 분석, DB 설계, 소스 데이터 입력, 데이터 변환 및 로드,DB 유지 보수 등의 일련의 과정을 포함한다. DB 구축을 위한 데이터 입력 및 DB 생성S/W의 개발 및 운영에 관한 기술은 컴퓨터 환경이나 정보 검색 시스템의 기능에 따라 다소 차이가 있을 수 있으나 대체로 유사하다고 볼 수 있다.

따라서 본 고에서는 한글 DB의 구축 및 운영 사례로서 국내 6,000 가입자('94년10월초 현재)를 대상으로 전자통신 분야의 전문(專門)적인 기술 정보 DB 12종을 서비스 하고 있는 한국전자통신연구소의 ETLARS 한글 정보검색 시스템에서의 데이터 입력및 DB 생성 S/W를 소개하고자 한다.

ABSTRACT

DB production process includes a series of data analysis, DB design, source data input, data conversion and load, and DB operation and maintenance. Software development and operation technique for data input and DB production is similar to each other, even though each computer environment or the function of information retrieval system is a little different.

The purpose of this paper is to introduce software for data input and DB production in ETLARS system, which is providing services to 6,000 users (as of Oct. 10, 1994) by Electronics and Telecommunications Research Institute, and to open know-how information in the production and operation of Hanguel DB.

* 한국전자통신연구소 정보유통개발실
** 한국전자통신연구소 전산망연구실
*** 한국전자통신연구소 기술정보센터
■논문제출일 : 1994년 10월 17일

1. 서 론

기술정보란 고도의 부가가치화된 지식으로 과학기술의 발전속도가 가속화됨에따라 데이터베이스(이하 DB로 칭함)는 기술정보의 신속하고 효율적인 유통매체로 각광받고 있다. 그러나 컴퓨터를 이용한 온라인 정보검색 시스템의 역사가 30년에 이르고, 한글 정보검색 시스템의 역사도 10년에 달하고 있으나(정영미,1986), 성공적으로 운영되고 있는 정보검색 시스템들에서의 S/W의 개발 및 운영에 관한 상세한 노하우(Know-how) 정보는 거의 발표되어 있지 않다.

최근들어 멀티미디어, 그래픽-사용자 인터페이스, 인공지능, 자연언어 처리 등의 신기술을 정보검색 시스템에 도입하기 위한 연구가 활발히 진행되고 있으나,대량의 기술정보를 운영하고 있는 데이터뱅크용 정보검색 시스템에 실제 적용하는 데는 아직도 많은 문제점이 있어 일부 기술만이 실용화되고 있을 뿐이다.

정보검색 시스템을 구성하는 소프트웨어(이하 S/W로 칭함)는 일반적으로 데이터 입력 및 DB 생성 S/W, 정보검색 S/W 그리고 시스템 관리 S/W로, 이들 S/W의 개발 및 운영에 관한 기술은 컴퓨터 환경이나 정보검색 시스템의 기능에 따라 다소 차이가 있을 수 있으나 대체로 유사하다고 볼 수 있다.

본 고에서는 한글 DB의 구축 및 운영 사례로서 국내 6,000 가입자('94년 10월초 현재)를 대상으로 전자통신 분야의 전문(專門)적인 기술정보 DB 12종을 서비스하고 있는 한국전자통신연구소의 ETLARS 한글 정보검색 시스템에서의 데이터 입력 및 DB 생성 S/W를 소개하고자 한다.

2. DB 구축 과정

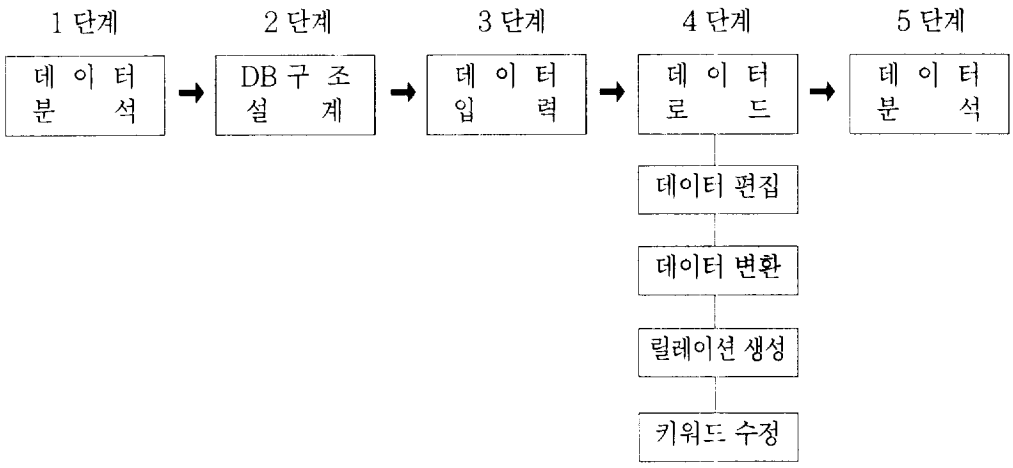
DB 구축이라 하면 데이터 분석, DB 구조 설계, 소스 데이터 입력, 소스 데이터 파일을 DB 구조로 변환시키기 까지의 데이터 변환/로드 과정, DB 유지.보수 등의 일련의 과정을 포함한다. 그러나 일반적으로 DB 구축이라 하면 소스 데이터 파일의 구축정도로 인식되고 있어 데이터의 입.출력형식(Format)에 관한 정보가 주로 공개되어 있으며, 이외에 DB 구조(파일 조직)에 관한 간략한 정보, 자동색인에 관한 이론적인 연구결과 등 만이 발표되고 있을 뿐, 실제 DB를 운영하는데 필요한 중요한 정보인 DB 생성을 위한 데이터 변환/로드 과정에서의 기술적인 노하우 정보는 거의 발표되어 있지 않다.

본 고에서는 ETLARS 시스템에서의 DB 구축과정을 (그림 1)에서와 같이 5단계로 나누어 설명한다.

3. 데이터 분석

DB 구축을 위해서는 먼저 DB를 구성할 데이터의 성격을 분석하여야 한다.데이터의 성격 분석이란 필드(입력항목)의 종류와 길이, 검색키로 사용할 필드의 지정 및 검색키의 추출방법(색인방식), 출력필드의 지정 및 출력방법 등 DB구축에 필요한 데이터 요소를 파악하고 이의 처리방법을 결정하는 것을 말한다(藤田節

〈그림 1〉 DB 구축 과정



子, 1992). 이러한 분석결과에 의거하여 각 필드에 태그번호(Tag No.)를 부여하고 필드별 데이터 기술요령을 정의하며 입.출력형태를 결정한다. 그리고 이 단계에서 정의한 사항들은 시스템 규격으로서 DB 구조를 설계하고 DB 변환/로드 프로그램을 작성할때 참고자료로 이용된다. (부록 "국내학술논문DB의 규격" 참조)

ETLARS 시스템에서는 검색효율을 증대시키기 위하여 검색키를 다양화 하였으며, 출력시 검색결과에 대한 독해가 용이하도록 출력형태를 설계하였다.

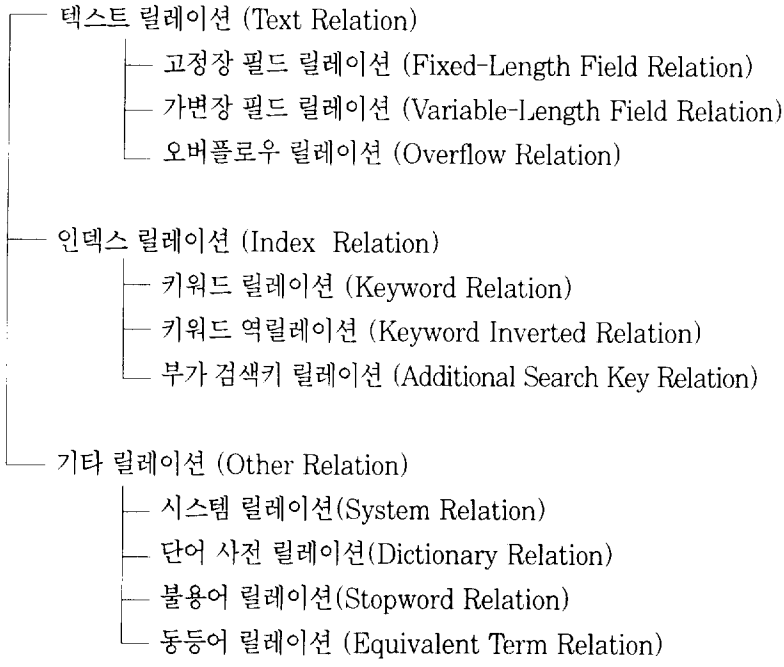
4. DB 구조 설계

ETLARS 시스템은 IBM 3090-200E/VF에서 운영되고 있으며, 관계형 DBMS인 DB2 Ver.2.2를 이용하여 각 DB를 독립적인 릴레이션(Relation) 형태로 구성하였다.

ETLARS 시스템을 구성하는 릴레이션들의 종류는 (그림 2)에서와 같이 입력한 소스 데이터가 모두 저장되어 있는 텍스트 릴레이션과 온라인 검색을 위한 검색키들(Search Keys)을 모아놓은 인덱스 릴레이션, 그리고 이외에 DB의 데이터 로드작업을 효율적으로 수행할 수 있도록 마련해 놓은 시스템 릴레이션, 단어사전 릴레이션, 불용어 릴레이션, 동등어 릴레이션이 있다. 텍스트 릴레이션과 인덱스 릴레이션은 데이터 로드결과 생성되어 각 DB마다 별개로 존재하는데 반해, 나머지 릴레이션들은 데이터 로드작업에 공동 활용될 수 있도록 사전에 생성해 놓아야 한다.

고정장 필드 릴레이션은 데이터의 길이가 고정되어 있는 필드를 저장하는 릴레이션으로서, 데이터 길이가 비교적 짧거나 필수적으로 입력되는 필드, 출력시 출력위치가 정해지는 필드 그리고 부가 검색키 필드(저자명, 출판년도, 자료종류 등과 같은 키워드(주제어) 이외의 검색키들로 검색시 AU=, PY=, DT=와 같은 접두코드를 수반함) 등으로 구성된다. 가

〈그림 2〉 릴레이션의 종류



변장 필드 릴레이션은 디스크 스페이스를 줄이기 위해 마련한 것으로 내용의 최대길이가 비교적 긴 필드(제목, 키워드, 초록, 본문 등)로 구성된다. 그리고 가변장 필드 릴레이션에서 정의한 필드의 길이를 초과하는 데이터를 저장하기 위해 오버플로우 릴레이션을 별도로 만들어 놓았다.

키워드 릴레이션은 가변장 필드 릴레이션의 키워드 필드와 제목 필드로부터 추출한 키워드들이 알파벳순으로 배열되어 있는 릴레이션으로서, DB 내에서 각 키워드를 포함하고 있는 레코드 수(자료 건수)도 함께 저장하고 있다. 키워드 역릴레이션은 키워드 릴레이션에 대한 역파일로서 포인터에 의해 상호연결되어 있으며 각 키워드를 포함하고 있는 자료들의 레코드 번호를 저장하고 있다. 부가 검색키 릴레이션은 텍스트 릴레이션으로부터 추출한 키워드 이외의 검색키들이 알파벳순으로 배열되어 있는 릴레이션으로서, 역릴레이션을 갖고 있지 않으므로 DB 내에서 각 검색키를 포함하고 있는 자료의 레코드 번호를 함께 저장하고 있다.

정보검색시에는 인덱스 릴레이션과 텍스트 릴레이션 만이 사용되며, 이들 릴레이션의 구조 및 연결관계는 (그림 3)과 같다. 즉 그림의 화살표 진행방향을 보면 ETLARS 시스템은 인덱스 릴레이션에서 찾고자 하는 자료의 건수와 이에 해당되는 자료의 레코드 번호를 알아낸 후 텍스트 릴레이션을 찾아가 자료의 실제 내용을 출력하게 된다. (* 릴레이션내 각 컬럼에 관한 상세한 내용은 부록의 "국내학술논문

(그림 3) 인덱스 릴레이션과 텍스트 릴레이션의 구조 및 연결관계

〈그림 3〉 인덱스 릴레이션과 텍스트 릴레이션의 구조 및 연결관계

DB의 릴레이션 구성" 부분에 수록하였음)

시스템 릴레이션은 데이터 필드의 종류 및 성격이 다양한 여러 DB들에 대한 색인작업 및 출력을 효율적으로 관리할 수 있도록 마련한 릴레이션으로서, 각 DB에 대하여 검색키 필드에 관한 정보(검색키 필드의 종류 및 해당 접두코드, 필드별 색인방식 및 동등어 처리여부 등), 출력형식에 관한 정보(출력 패러그라프 및 필드의 표목, 출력위치 등) 등을 저장하고 있다. 불용어 릴레이션과 단어사전 릴레이션은 키워드의 자동색인 작업에 사용된다.

5. 데이터 입력

한글인식 기술이 아직 완벽하지 않은 관계로 소스 데이터의 입력작업은 대부분 수작업에 의해 이루어지고 있다. ETLARS 시스템에서는 여러 입력자가 동시에 데이터 입력작업을 수행하는 등 입력자의 작업 편리성에 중점을 두어 주로 PC에서 데이터를 입력한 후 IBM 3270 Emulation S/W를 이용하여 IBM으로 화일전송하는 방법을 채택하였다. PC에 입력된 데이터 파일은 ASCII 파일로서 사전에 정의해 놓은 IBM의 Data Set으로 전송된다.

6. 데이터 로드

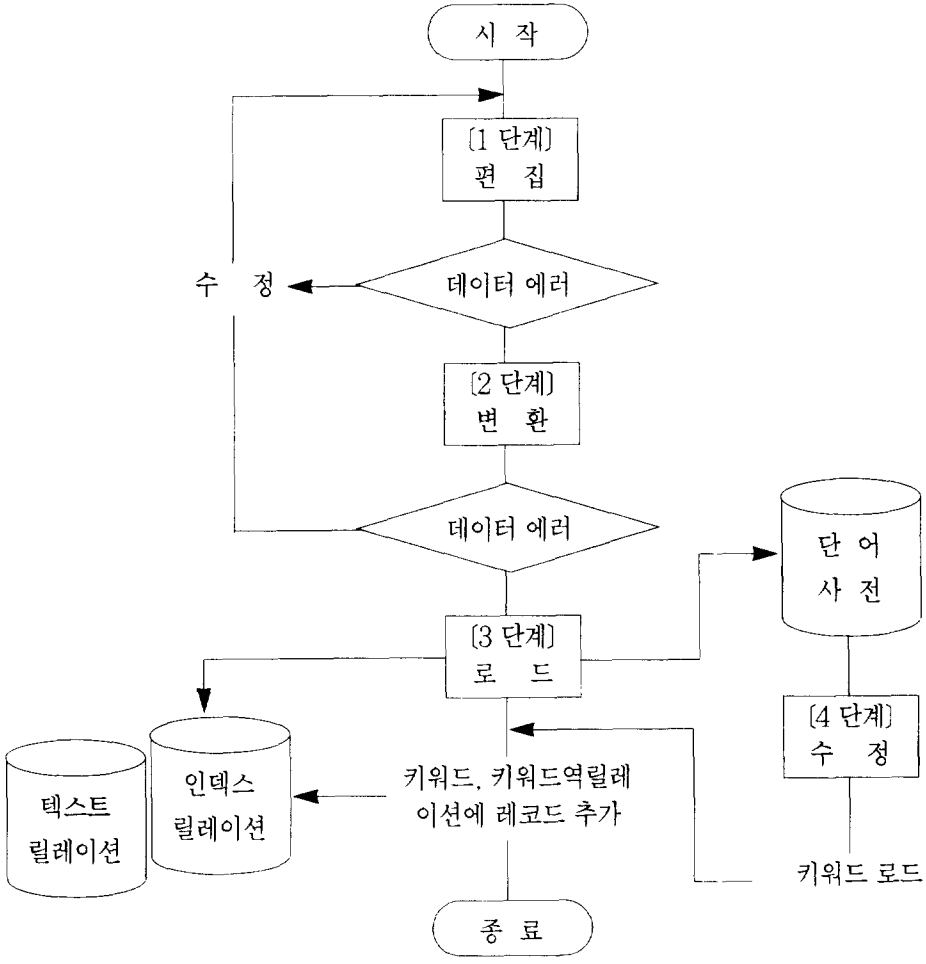
데이터 로드란 입력된 소스 데이터 파일을 ETLARS 시스템에서 활용할 수 있는 DB 구조 형태로 변환하는 것을 말한다. 다시 말해서 DB 구조설계 단계에서 설계한 형태대로 텍스트 릴레이션과 인덱스 릴레이션을 생성하는 과정으로서, DB2의 SQL 명령에 의하여 릴레이션을 구성하며 JCL(Job Control Language)을 이용하여 일괄(Batch)처리 한다.

ETLARS 시스템에서의 데이터 로드 작업은 (그림 4)에서 보는 바와 같이 4단계를 거쳐 수행된다. 1단계에서는 소스 데이터의 형식을 통일시키는 편집작업을 수행하며, 2단계에서는 소스 데이터의 에러를 검사하고, 시스템 규격에 따라 부가 검색키 릴레이션을 구성할 데이터를 추출한다. 3단계에서는 먼저 키워드 릴레이션을 구성할 데이터를 추출한 후, 고정장 필드 릴레이션, 가변장 필드 릴레이션, 키워드 릴레이션, 부가 검색키 릴레이션에 데이터를 로드한다. 또한 키워드 역릴레이션을 구성할 데이터를 생성하여 로드하고 키워드와 동등어에 대응하는 레코드 수를 키워드 릴레이션에 추가 갱신한다. 마지막으로 4단계에서는 추출된 키워드 중에서 단어사전 릴레이션에 없는 키워드 후보에 대한 수정을 수작업으로 처리한 후 키워드로 인정된 후보 단어를 키워드 릴레이션에 추가 로드하고 단어사전 릴레이션에도 추가하여 단어사전을 확장한다. 만일 2단계에서 데이터 에러만 없다면 1단계에서부터 4단계까지의 전과정이 자동으로 처리될 수 있다.

6.1. 소스 데이터 편집 (1 단계)

소스 데이터의 편집 작업은 입력된 소스 데이터 파일의 데이터 형식이 통일되어 있지 않은 경우 이를 통일시키는 작업, 한글 데이터 좌우의 SOSI 코드(Shift Out Shift In Code : 영숫자 표준 부호이외의 문자를 표시하는데 사용하는 변환코드로서 화면상에는 스페이스로

〈그림 4〉 데이터 로드 흐름도



나타남)로 인한 한글단어 사이의 스페이스를 조절하는 한글 데이터 편집작업, 그리고 필요한 데이터 필드나 레코드를 생성하여 추가시키는 형식 재편집 작업을 포함한다.

첫단계는 소스 데이터의 입력 형식이 통일되지 않은채 여러 기관에서 작성한 데이터 파일의 태그 번호 및 데이터 내용 형식을 통일시키는 작업을 말한다. 두번째 단계의 한글 데이터 편집 작업이란 일반 IBM PC에서 작성한 한글이 IBM 메인프레임으로 전송되어 IBM 혼합코드로 저장되면 한글단어의 시작과 끝에 SOSI 코드가 포함되어 한글 단어와 단어사이가 3 Byte 떨어지게 되는데 이 불필요한 SOSI 코드를 아래와 같이 삭제하여 한글 단어

사이의 간격을 2 Byte로 줄여주는 작업을 말한다.

@ 편집전 :

〔한글〕〔데이터〕〔편집은〕 IBM PC 〔에서〕

@ 편집후 :

〔한글 데이터 편집은〕 IBM PC 〔에서〕

* 여기에서 "〔" 는 SO 코드로 16 진수값은 0E 이며, "〕"는 SI 코드로 16 진수값은 0F 이다.

마지막 단계에서는 한글 데이터의 편집 결과 레코드의 우측 마진(Right Margin)이 소스 데이터의 레코드 길이제한(72 바이트)을 초과하는 경우 우측 마진을 정렬시키고 연속 레코드를 생성시키며, 소스 데이터로부터 필요한 데이터를 추출하여 새로운 필드(예. 자료종류코드, 언어코드 등)를 생성하는 작업을 수행한다.

6.2. 데이터 변환 (2 단계)

데이터 변환 작업 단계에서는 먼저 소스 데이터의 에러를 검사하며 에러수정이 완료되면 DB별로 시스템에 의하여 연속적인 레코드 번호를 부여한다. 그 결과 소스 데이터 파일이 텍스트 릴레이션에 저장될 형태로 완성되면 시스템 규격에따라 부가 검색키 필드로부터 데이터를 추출하여 부가 검색키 릴레이션에 저장할 형태로 편집한다.

데이터 에러를 검사할 때는 입력 필수 필드와 중복된 필드를 검사하고, 부가 검색키 필드 데이터의 제한범위(문자모드, 길이 등)를 검사한다. 만약 검사결과 데이터 에러가 발생하면 다음 단계로 넘어갈 수 없으므로 일일히 수작업으로 소스 데이터를 수정한후 데이터 변환 단계를 재수행하여야 한다.

6.3. 릴레이션 생성 (3 단계)

릴레이션 생성은 DB 로드 프로그램의 핵심 단계로서, 이 단계에서는 고정장 필드 릴레이션, 가변장 필드 릴레이션, 부가 검색키 릴레이션, 키워드 릴레이션,그리고 키워드 역릴레이션이 모두 생성된다. 이때 본문이나 초록의 내용이 너무 길어 가변장 릴레이션의 필드길이를 초과할 경우는 초과한 내용을 오버플로우 릴레이션에 로드한다. 고정장 필드 릴레이션, 가변장 필드 릴레이션, 부가 검색키 릴레이션은 2 단계에서 변환된 데이터로부터 로드되나 키워드 릴레이션과 키워드 역릴레이션은 키워드 색인작업과 동등어 처리과정을 거쳐 생성된다.

6.3.1. 키워드 색인 작업

ETLARS 시스템에서는 검색시 자료에 대한 접근점(Access Point) 즉, 검색키를 증가시키기 위한 방안으로 구 색인방식(Phrase Indexing), 단어 색인방식(Word Indexing), 혼합 색인방식((Word & Phrase Indexing) 등 다양한 색인방식을 적용하고 있다. (DIALOG Information Services, Inc., 1985)

구 색인방식은 데이터 필드내 서브필드 구분자(Delimiter)인 세미콜론(;)을 기준으로 하여 검색키를 추출하는 방법으로 저자명, 출처명(잡지명), 자료번호 등의 필드에 적용하였으며, 단어 색인방식은 스페이스 또는 SOSI 코드를 기준으로 단어와 단어를 분리하여 검색키를 추출하는 방법으로 제목 필드에 적용하였다.또한 혼합 색인방식은 구 색인방식과 단어 색인방식을 혼용한 것으로 키워드 필드에 적용

하였다. 그러나 ETLARS 시스템에서는 검색시 한글 띄어쓰기의 모호성으로 인한 혼란을 막기 위하여 구 형태의 한글 키워드에 대하여 단어 사이의 스페이스를 제거한 복합어 키워드를 추가로 생성하는 변형된 혼합 색인방식을 적용하고 있다.

예를 들어 키워드 필드에 "공장 자동화;산업용 로봇" 이라는 2개의 키워드가 부여된 경우, 아래와 같이 구 색인작업한 경우는 2개의 검색키, 단어 색인작업한 경우는 4개의 검색키, 혼합 색인작업한 경우는 6개의 검색키, 변형된 혼합색인작업한 경우는 8개의 검색키가 생성된다.

* 구 색인 결과 :
  공장 자동화 ; 산업용 로봇
* 단어 색인 결과 :
  공장 ; 자동화 ; 산업용 ; 로봇
* 혼합 색인 결과 :
  공장 자동화 ; 산업용 로봇 ; 공장 ; 자동화 ; 산업용 ; 로봇
* 변형된 혼합색인 결과 :
  공장 자동화 ; 산업용 로봇 ; 공장 ; 자동화 ; 산업용 ; 로봇 ; 공장자동화 ; 산업용로봇

또한 ETLARS 시스템에서는 제목 필드에 대한 단어 색인작업을 하는데 있어서 자동색인을 실시하고 있다. 즉, 제목으로부터 키워드로서 유효하지 않은 단어인 불용어를 제외한 나머지 단어들을 키워드로 추출하는 작업을 전적으로 프로그램에 의해 실시하고 있다. 영어 단어에 대한 자동색인은 용이하나 한글 단어의 경우는 불용어 뿐만 아니라 조사와 같은 토씨를 제거하는데 많은 어려움이 있으며 정확성도 100% 확신할 수 없는 실정이다.

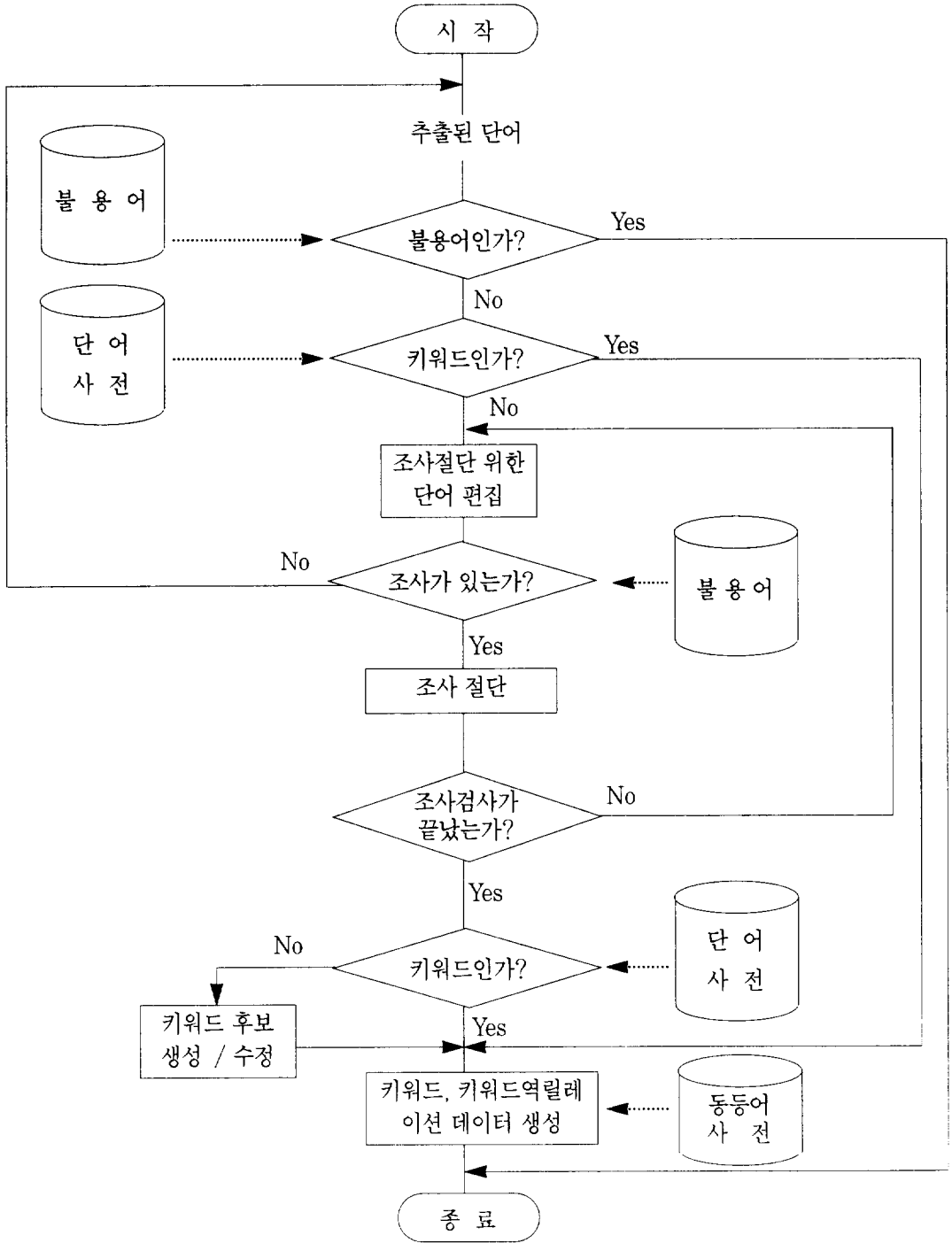
ETLARS 시스템에서는 한글 단어에 대한 자동색인 과정에서 불용어나 조사 등 불필요한 단어를 정확히 제거시킬 수 있도록 불용어 릴레이션과 단어사전 릴레이션을 두고 있다. 이들 릴레이션은 데이터 로드중 새로운 단어와 불용어를 수시로 추가하여 확장시켜 나갈 수 있도록 하고 있으며, 이들 릴레이션이 확장되어 정확한 자동색인 결과를 기대할 수 있을 경우는 초록이나 본문 필드에도 자동색인을 적용할 계획이다. (그림 5)는 ETLARS 시스템에서의 한글 자동색인 과정을 보여주고 있다.

6.3.2. 동등어 처리

"컴퓨터"와 "Computer", "Computers"는 주제 개념은 동일하나 컴퓨터 내에서의 디지틀 값이 각기 다르므로 별개의 키워드로 존재한다. 따라서 "컴퓨터"에 관한 자료를 모두 검색하기 위해서는 "컴퓨터+COMPUTER+COMPUTERS"와 같이 세 단어를 조합한 검색문을 작성해야 한다. 그러나 일반 이용자가 검색시 이러한 관련 단어를 모두 생각해내어 검색문을 작성할 것으로 기대하기는 어렵다. ETLARS 시스템에서는 이용자가 이들 단어중 하나 만을 검색어로 사용해도 "컴퓨터"에 관한 자료가 모두 검색될 수 있도록 동등어 처리를 하였다.

동등어 처리방법은 데이터 로드시 미리 동등어 처리를 해놓는 방법과 검색시 동등어 처리를 하는 방법이 있으나 후자의 경우는 검색시간이 많이 걸릴 우려가 있으므로 ETLARS 시스템에서는 전자의 방법을 채택하였다. 즉, 키워드 색인단계에서 추출한 키워드를 동등어

〈그림 5〉 한글 자동색인 과정



사전을 이용하여 키워드를 확장하며, 이 확장된 키워드에 해당되는 레코드 수 다시 말해서, 동등어 그룹에 속하는 각 키워드를 포함하는 자료의 건수를 모두 합한 건수를 키워드 릴레이션에 추가한다.

(그림 3)의 키워드 릴레이션을 보면, "컴퓨터" 라는 키워드를 포함하고 있는 자료는 75건, "COMPUTER"는 86건, COMPUTERS는 10건으로 이들 키워드를 동등어 처리한 결과는 157건임을 알 수 있다.

ETLARS 시스템에서 구현한 동등어 처리법의 특징은, 첫째 키워드의 확장으로인해 키워드 릴레이션과 키워드 역릴레이션의 레코드 수가 크게 증가하지 않으며,둘째 동등어를 추가하고 삭제하기 쉽도록 구성하였다는 점이다. 즉, 동등어를 추가할 때는 단지 키워드 릴레이션 속의 동등어 번호와 동등어 레코드 수를 수정하기만 하면 되며, 마찬가지로 삭제할 때는 동등어 번호와 동등어 레코드 수를 "0"으로 변경하기만 하면 된다. 현재 동등어 사전에 등록된 동등어 그룹수는 864 그룹이며, 동등어 단어 수는 2,354개이나 검색효율을 향상시키기 위하여 동등어사전을 계속해서 확장시켜 나갈 계획이다.

6.3.3. 레코드 건수 생성

키워드 역릴레이션은 1차적으로는 키워드 번호순으로, 2차적으로는 동등어 번호순으로 배열되어 있으며, 키워드 릴레이션에 있어서 개개 키워드에 해당되는 키워드 레코드 수와 동등어 레코드 수는 키워드 역릴레이션을 통해 생성된다. 즉,키워드 역릴레이션에서 동일한 키워드 번호를 갖는 레코드 수와 동일한 동등어 번호를 갖는 레코드 수를 구해서 키워드 릴레이션의 레코드 수를 추가.갱신한다.키워드 역릴레이션에서 접두코드는 키워드가 추출된 출처필드를 나타내며, 문장번호와 단어번호는 키워드가 추출된 출처 필드에서의 단어의 발생 위치(순서)를 나타내는 정보로 검색시 인접연산자 검색을 위해 마련해 놓은 것이다.

6.4. 키워드 수정 (4 단계)

키워드 릴레이션을 구성할때 대부분 에러없이 한번에 구성하기는 어렵고 반드시 몇번의 수정.검증을 통한 보완작업이 필요하다. 에러는 데이터 작성시 발생한 입력오류와 키워드 색인작업 과정에서 발생하는 중간오류로 크게 나눠볼수 있으며, ETLARS 시스템에서는 이러한 에러로 인해 부정확한 키워드가 로드되는 것을 방지하기 위하여 키워드 후보에 대한 검증단계를 두고 있다.

키워드 색인과정에서 추출된 모든 키워드에는 〈표 1〉에 나타낸 것과 같이 키워드의 상태를 분류한 코드가 부여된다. 분류코드가 "1" 이거나 "4"인 단어는기존 단어사전 릴레이션에 존재하는 단어이므로 우선적으로 키워드 릴레이션에

〈표 1〉 키워드 후보 코드표

코 드	키워드 상태
1	단어사전 릴레이션에 존재
2	불용어 릴레이션에 존재
3	조사(불용어)가 절단된 후 불용어 릴레이션에 존재
4	조사가 절단된 후 단어사전 릴레이션에 존재
5	조사를 절단한 후에도 단어사전에 존재하지 않는 키워드 후보
6	단어사전에 존재하지 않는 키워드 후보
7	쓰이지 않는 데이터

로드되는데 반해, 분류코드가 "5" 이거나 "6"인 단어는 단어사전 릴레이션에 존재하지 않으므로 키워드 검증과정을 통해 키워드로 인정된 후에야 키워드 릴레이션에 추가 로드되게 된다.

키워드 후보는 IBM 시스템의 순차접근(SAM : Sequential Access Method)파일형태로 구성되어 있으며, 키워드 후보에 대한 검증과정에서 조사가 절단되기 이전의 키워드를 확인할 수 있게 하여 키워드 수정작업의 정확성과 다음 단계의 키워드 생성을 보다 용이하게 하고 있다. 예를 들어, 조사가 절단된 후에도 단어사전 릴레이션에 존재하지 않는 단어인 분류코드가 "5"인 키워드 후보를 단어사전 릴레이션에 추가하려면 분류코드를 "1"로, 불용어 릴레이션에 추가하려면 분류코드를 "2"로, 또는 키워드나 불용어로 사용하지 않으려면 분류코드를 "7"로 수정하면 된다. 그리고 분류코드 "5"나 "6"에 속하는 키워드 후보중에서 새로운 키워드로 인정된 경우는 후보 상태에서 벗어나 정식 키워드로서 키워드 릴레이션에 추가로드되며 키워드 역릴레이션 생성과정을 다시 거침으로서 데이터 로드작업이 모두 끝나게 된다.

7. DB의 유지 보수

DB의 유지보수란 텍스트 및 인덱스 릴레이션에 저장된 데이터가 최적의 상태를 유지하여 데이터 접근시 항상 최단 경로로 접근이 가능하도록 하고, 시스템 장애 등에 의해 데이터가 손상을 입었을 경우 데이터 복구가 용이하도록 사전에 릴레이션 데이터를 백업받아 두는 일련의 작업을 말한다.

검색 시스템의 검색 효율을 극대화 시키기 위해서는 텍스트 및 인덱스 릴레이션에 저장된 데이터가 주요 키(Primary Key)의 순서대로 배열되어 있어 검색시 최단 경로를 통한 데이터 접근이 가능하여야 한다. 특히 키워드 릴레이션은 데이터 로드 초기에는 데이터가 키 순으로 배열되어 있으나 데이터가 계속 추가됨에 따라추가되는 데이터는 부득이 오버플로우 영역에 저장되기 때문에 데이터의 키 순으로 배열되지 않게 되어 검색시 응답시간이 길어지게 된다. 이러한 경우 사전에 백업받아 놓은 데이터를 이용하여 인덱스 릴레이션을 재조직하여야 한다.

DB의 유지 보수 작업은 주로 DB2의 유틸리티 프로그램을 이용하여 배치작업으로 일괄 수행한다.

8. 결 어

90년대에 들어서 국내에서도 DB 산업에 대한 인식이 급증하고 있으나 문헌정보 DB는 아직 규모가 빈약할 뿐만 아니라 온라인 정보검색 시스템의 개발 기술도 미비한 실정이다. 그 결과 현재 국내에서 문헌정보 데이터뱅크용 한글 정보검색 시스템이 개발되어 성공적으로 운영되고 있는 사례는 극히 소수에 불과하다.

본 고에서는 한글 정보검색 시스템의 보급 확산을 도모하고자 10여년간의 연구개발 경력을 지닌 ETLARS 시스템에서의 DB 설계, 구축, 및 운영에 관한 내용을 게재하였다. ETLARS 시스템에서의 한글편집, 자동색인,

동등어 처리, 그리고 각 DB에 대한 데이터 변환/로드 과정의 표준화를 위한 시스템 규격 작성법 등에 관한 내용은 자체적인 경험에 의한 결과로 한글 정보검색 시스템의 발전에 조금이나마 도움이 되기를 기대한다.

참고문헌

藤田節子, 1992, データベース 設計入門, 日外アソシエーシ (株).정영미, 1986, 정보 검색론, 서울, 구미무역.

한국전자통신연구소, 1991, 1993, 정보통신 기술정보센터 운영, 연구보고서.

Date, C. J., 1982, An introduct ion to database systems, 3rd. ed., Reading, Addison-Wesley Publishing Company.DB2 application progra mming, 1991, 서울, 한국 IBM.

Fit ch, Carl et al., 1989, DB2 applicat ions development handbook, New York, McGraw-Hill Book Company.

Guide to DIALOG searching, 1985, Palo Alto, DIALOG Information Services, Inc.

Inmon, W. H., 1991 DB2 : maximizing performance of online production systems, Boston, OECD Information Sciences, Inc.,

〈부 록〉

○ 국내학술논문 DB의 규격

1. 출력견본

제 목 : 분할된 근거리 망 분산 데이터베이스 시스템에서 결합 연산 최적화

저 자 : 안명수 ; 박종수 ; 이동면 ; 박웅규 ; 김명환
(한국과학기술원 ; 성신여자대학교)

출 처 : 정보과학회 논문지 Vol. 19 No. 5 (1992.09.)
페이지 : 498-508

자료종류 : J 언어 : K

키 워드 : 분산 데이터베이스 ; Join operation; 비용 함수 ; Fragment join assignment; Graph partitioning; Heuristic algorithm

내 용 :

본 논문에서는 고속 근거리망에 구축된 분산 데이터베이스 시스템에서 분할된 두 릴레이션 사이의 결합 연산을 최적화하기 위한 효율적인 알고리즘을 제안한다. 제안된 알고리즘은 프래그먼트 (fragment) 결합 할당 방법을 사용하여 응답시간을 최소화 한다. 프래그먼트 결합 할당 문제와 평형된 그래프 분할 문제 사이의 유사성이 식별되었다. 두 문제가 모두 Np-hard 문제이므로 그래프 분할을 위해 잘 알려진 휴리스틱 알고리즘이 결합 할당을 위해 수정되어 응용되었다. 각 프래그먼트에 관한 의미정보와 복사본을 이용하여 불필요한 처리비용을 줄이고 병렬처리 능력을 향상시킨다. 모의 실험 결과 제안된 알고리즘이 기존의 알고리즘보다 좋은 성능을 보여주며 각 사이트의 부하도 평형됨이 밝혀졌다.

2. 입력 필드 정의

필 드 명	컬 럼 명	태그번호	압력필수	필드길이
@ 레코드번호(1)	RECNO1	000	o	F(8)
레코드번호(2)	RECNO2	001	o	F(15)
논문 제목	TITLE	100	o	V(240)
저 자 명	AUTHOR	200		V(160)
# 소속기관명	AFFILIAT	210		V(160)
잡 지 명	JNAME	300	o	V(160)
권 (Vol)	VOLUME	310		V(10)
호 (No)	NUMBER	320		V(10)
# 출판 년월	PUBDATE	330	o	V(12)
@ 출판 년도	PY	335	o	F(2)
페 이 지	PAGE	340	o	V(20)
회 의 명	CONFTITLE	360		V(160)
회의 장소	CONFLOCATE	370		V(50)
회의 날짜	CONFDATE	380		V(30)
@ 자료 종류	DT	400	o	F(1)
@ 언 어	LA	410	o	F(1)
ISSN 번호	ISSN	430		F(9)
키 워 드	KEYWORD	500		V(500)
내 용	CONTENT	600		V(31680)

(설 명)

-레코드 번호(2)는 다수의 입력자가 입력시 작업의 편의에 의해 부여한 번호이며, 레코드 번호(1)은 시스템에서 각 레코드에 순차적으로 부여한 번호임.

-컬럼(Column)명 : 텍스트 릴레이션에서의 필드명

-@ 표시 : 데이터의 정확성을 보장하기 위해 시스템이 데이터를 생성하여 입력함.

-# 표시 : 출력형태로 데이터를 미리 편집해 놓음.

3. 검색키 정의

검 색 키	접두코드	태그번호	색인방식	문자수	비 고
* 주 제 어	없음	500	M		○동등어 처리
		100	W		
* 저 자 명	AU=	200	P		
회 의 명	CT=	360	P	64자	
자료종류	DT=	400	P	1자	
잡 지 명	JN=	300	P	64자	
언 어	LA=	410	P	1자	
출판년도	PY=	335	P	2자	○ 'YY' 형태
제목내단어	TI=	100	W		

(설 명)

－색인방식

· W : 단어 색인방식

· P : 구 색인방식

· M : 혼합 색인방식

－ * 표시 : 복수의 데이터(Multi-key) 가능

4. 출력 필드 정의

입력 필드명	태그번호	출 력 패러그라프명	출력표목	기 타
레코드번호(1)	000	레코드		
레코드번호(2)	001			○출력안함
논문 제목	100	제 목		
저 자 명	200	저 자		
소속기관명	210			○줄바꿈
잡 지 명	300	출 처		○ 줄바꿈
권	310		Vol.	
호	320		No.	
출판 년월	330			
페 이 지	340		페이지 :	
회 의 명	360		회의명 :	
회의 장소	370		장소 :	
회의 날짜	380		날짜 :	
자료 종류	400	자료종류		
언 어	410	언 어		○위치지정
ISSN 번호	430			○출력안함
키 워 드	500	키워드		
내 용	600	내 용		○1줄후 출력

(설 명)

– 패러그라프는 출력의 단위가 되는 데이터 그룹을 말하며, 패러그라프 및 이에 속하는 필드들에 대한 이해를 돕기 위해 표목을 부여함.

○ 국내학술논문 DB의 릴레이션 구성

릴레이션 종류	릴레이션 명	Table Space 명
고정장 필드 릴레이션	TINFO42.E3012FIXED	E3012DB.E3012003
가변장 필드 릴레이션	TINFO42.E3012VARI	E3012DB.E3012004
가변장 필드 오버플로릴레이션	TINFO42.E3012OVARI	E3012DB.E3012004
키워드 릴레이션	TINFO42.E3012KW	E3012DB.E3012001
키워드 역 릴레이션	TINFO42.E3012KWINV	E3012DB.E3012001
접두코드 릴레이션	TINFO42.E3012SKEY	E3012DB.E3012005

■ 텍스트 릴레이션 ■

◇ 고정장 필드 릴레이션

```
CREATE TABLE TINFO42.E3012FIXED
    ( RECNO1        CHAR(8)          NOT NULL,
    RECNO2          CHAR(15)         NOT NULL,
    VOLUME          CHAR(10),
    NUMBER          CHAR(10),
    PUBDATE         CHAR(12),
    PAGE            CHAR(20),
    CONFDATE        CHAR(30),
    DT              CHAR(1),
    LA              CHAR(1),
    ISSN            CHAR(9) )
  IN E3012DB.E3012003;
```

◇ 가변장 필드 릴레이션

```
CREATE TABLE TINFO42.E3012VARI
    ( RECNO1        CHAR(8)          NOT NULL,
    TITLE           VARCHAR(240),
    AUTHOR          VAR CHAR(160),
    AFFILIAT        VAR CHAR(160),
    JNAME           VAR CHAR(160),
    CONFTITLE       VAR CHAR(160),
    CONFLOCATE      VAR CHAR(50),
    KEYWORD         VAR CHAR(500),
    OVERFLAG        CHAR(1)          NOT NULL,
    CONTENT         VAR CHAR(2448) )
  IN E3012DB.E3012004;
```

◇ 가변장 필드 오버플로 릴레이션

```
CREATE TABLE TINFO42.E3012OVARI
    ( RECNO1        CHAR(8)          NOT NULL,
    FLDVAR          CHAR(18)         NOT NULL,
    OVERTEXTNO SMALLINT              NOT NULL,
    OVERTEXT        VARCHAR(3960)    NOT NULL )
  IN E3012DB.E3012004;
```

■ 인덱스 릴레이션 ■

◇ 키워드 릴레이션

```
CREATE TABLE TINFO42.E3012KW
      ( KEYWORD        CHAR(64)        NOT NULL,
      KWNO INTEGER                     NOT NULL,
      KWRECCNT INTEGER                 NOT NULL,
      KWOCCURCNT INTEGER               NOT NULL,
      SYNNO INTEGER,
      SYNRECCNT INTEGER,
      PRIMARY KEY (KEYWORD) )
   IN E3012DB.E3012001;
```

◇ 키워드 역릴레이션

```
CREATE TABLE TINFO42.E3012KWINV
      ( KWNO            INTEGER        NOT NULL,
      SYNNO             INTEGER,
      RECNO1            CHAR(8)        NOT NULL,
      PREFIX            CHAR(2)        NOT NULL,
      SENTNO            SMALLINT,
      WORDNO            SMALLINT )
   IN E3012DB.E3012001;
```

◇ 접두코드 릴레이션

```
CREATE TABLE TINFO42.E3012SKEY
      ( RECNO1          CHAR(8)        NOT NULL,
      AU                CHAR(64)
      CT                CHAR(64)
      DT                CHAR(1)        NOT NULL,
      JN                CHAR(64)       NOT NULL,
      LA                CHAR(1)        NOT NULL,
      PY                CHAR(2)        NOT NULL WITH DEFAULT )
   IN E3012DB.E3012006;
```