□ 기술해설 □

신문과 방송 자료의 데이타베이스 시스템 구축

전자부품종합기술연구소 김종태* · 정혜윤*

● 목 차 ●

1. 서 론
2. 시스템 설계
2.1 텍스트 데이타베이스
2.2 이미지 데이타베이스
2.3 오디오 및 비디오 데이타베이스
2.4 방송자료의 데이타 모델
3. 시스템 구현
4. 현재 Status 및 향후 계획

1. 서 론

신문 기사의 데이타베이스는 비교적 오래전부터 구축되어 왔으나 신문 기사중 사진 데이타는 막대한 양의 저장 용량과 전용 하드웨어를 필요로 하여 널리 사용되지 못하였고 주로 텍스트 위주로 이루어져 왔다. 그러나 컴퓨터 하드웨어의 급속한 발달과 가격하락으로 인하여 고속 대용량의 처리가 가능해져 화상 형태의 사진 자료도 쉽게 다룰 수 있게 되었다. 국내 일간지 신문사의 경우 하루에 기사는 약 200건, 사진은 약 300장이 발생하며 장기간 저장할 경우 수백만건에 이른다. 이를 효과적으로 관리하기 위해서는 대용량을 다룰 수 있는 텍스트 및 이미지 데이타베이스가 구축되어야 한다.

방송사의 경우 비디오나 오디오의 방송 자료를 테이프에 수록하여 관련된 키워드나 서지정보를 부가하여 테이프 단위로 관리하고 있다. 사용자는 원하는 테이프를 찾기 위해서 먼저 키워드나 서지정보에 의한 검색을 통해 후보 테이프들을 찾은 다음 정확히 원하는 내용이 들어있는 테이프를 찾기 위해서는 후보 테이프들을 순차적으로 재생하면서 내용을 확인하는 과정을 거쳐야만 한다. 만일 원하는 장면이 없으면 위와 같은 작업을 여러번 반복하여야 하며, 어떤 테이프는 재생시 고가의 장비가 필요하므로 여러 사용자가 자원을 공유하는데 불편이 따른다. 또한 테이프의 반복적인 사용으로 원본이 손상될 우려가 있고 이미 대출된 테이프는 반납될 때까지 다른 사람이 이용할 수 없을 뿐 아니라 분실될 수 있으며 원하는 부분을 찾기위해 많은 시간과 노력이 요구되고 있다. 현재 KBS의 경우 보유하고 있는 테이프의 갯수는 비디오의 경우 약 29만개이고 오디오의 경우 약 38만개이다. 매월 평균 발생수는 비디오의 경우 2천개이고 오디오는 1천5백개 정도이다. 이러한 방대한 양의 자료 또한 효과적인 관리 및 검색을 위한 데이타베이스의 구축이 필요하다.

최근 컴퓨터 관련 기술의 눈부신 발달로 이러한 멀티미디어 데이타베이스의 구축이 현실화되고 있다. 고성능의 PC가 널리 보급되면서 정보시스템의 도입에 클라이언트-서버 형태의 시스템 구축이 일반화되고 있으며 그래픽 사용자 인터페이스의 발달로 편리한 검색 수단이 제공 되고 있다. 즉 클라이언트에서 기능의 일부를 수행하므로써 호스트의 부담을 줄일 수 있고, 사용자는 자료검색시 시스템의 안내를 받으면서 윈도우상에서 자료의 내용을 쉽게 확

*정 회 원

인할 수 있다. 또한 메모리 장치의 가격 하락으로 하드디스크도 큰 부담없이 활용할 수 있어 비정형 자료와 같은 방대한 양의 자료도 쉽게 저장할 수 있게 되었다. 네트웍의 기술도 크게 발달하여 광대역 통신이 점차 일반화되고 있어 비디오 자료도 PC에서의 검색이 가능해졌다.

신문기사는 시소러스 구축과 한글의 형태소 분석 기법이 발달하면서 컴퓨터가 자동으로 색인어 처리를 하여 데이타베이스 구축이 용이하게 되었다. 사진과 같은 이미지 데이타는 PC의 성능이 향상 되고 광저장 매체 및 이미지 압축과 복원 기법 등의 발달로 손쉽게 다룰 수 있다. 또한 컴퓨터에서 비디오 및 오디오의 자료 처리도 가능해져 아날로그 방식의 자료를 디지탈화하여 데이타베이스로 구축할 수 있다. 이러한 시스템이 구축되면 사용자는 책상위에 있는 PC상에서 손끝으로 원하는 신문 및 방송 자료를 쉽게 찾을 수 있으며 정확한 정보로의 접근이 용이해진다.

본 연구에서는 신문 및 방송 자료를 효과적으로 저장하고 검색, 관리할 수 있는 멀티미디어 데이타베이스 시스템인 일명 MMIS(Multimedia Information System)의 설계 및 구현에 관해 기술하고자 한다[1].

2. 시스템 설계

MMIS는 제 1단계로서 멀티미디어 데이타베이스를 구축하여 언론사 내부에 있는 기자나 PD 또는 자료가 필요한 인원을 대상으로 검색할 수 있는 시스템을 구축한다. 제 2단계는 시스템을 확장하여 일반인을 대상으로 정보서비스를 할 수 있는 시스템의 구현을 목표로 삼고 있다. 데이타베이스에 구축되는 자료의 형태는 첫 단계에서 신문기사, 사진, 라디오 방송과 관련된 음향, TV 방송과 관련된 뉴스, 다큐멘터리, 드라마, 스포츠 등 프로그램 성격에 따라 구분하여 구축된다. 비디오의 경우 제 1단계에서는 MPEG-1를 사용하여 영상의 질보다 검색의 속도를 주요 목표로 삼았으며 제 2단계에서 MPEG-2를 사용하여 영구적인 저장 및 송출용으로 쓸 수 있는 시스템을 목표로 추진할 예정이다. 그러므로 텍스트, 이미지, 오디오, 그리고 비디오의 데이타베이스를 하나의 시스템으로 통합 구축하고자 한다.

2.1 텍스트 데이타베이스

최근 신문제작은 납활자 대신에 CTS(Computerized Typesetting System)를 사용하여 기사의 입력부터 편집, 교정, 조판(Layout), 출력 등이 전산화되어 있다. 규모가 큰 신문사는 전공정이 전산화되어 있고 소규모 신문사는 일부만 되어 있다. DB구축을 위한 업무흐름은 그림 1에서 보는 바와 같다.

CTS에 저장되어 있는 기사의 문자정보와 조판정보는 먼저 전자문서의 표준인 SGML (Standard Generalized Markup Language) 포맷으로 변환된다. 문자정보는 형태소분석기를 통해 색인어 후보가 추출되고 후통제 과정에서 색인전문가에 의해 최종 색인어와 주제어가 부여된다. SGML로 변환된 기사는 추출된 색인어, 주제어, 서지사항 등과 함께 DB항목으로 텍스트 DBMS에 저장된다. 검색시에는 키워드를 사용하여 기사를 찾고 조판정보는 Rasterizer를 통해 화면상에 기사배치와 똑같은 그래픽형태로 변환되어 나타난다.

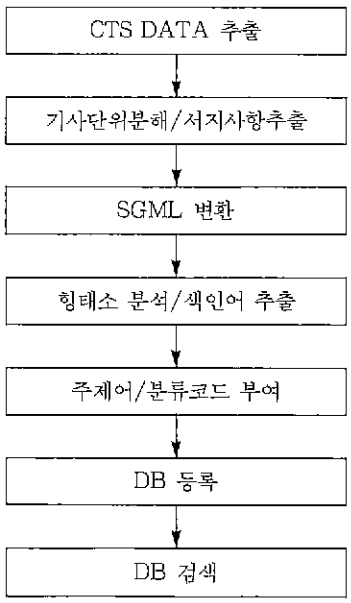

그림 1 기사 데이타베이스 구축을 위한 업무흐름도

형태소분석

대규모의 문헌정보나 기사정보를 검색하기 위해서는 키워드방식이 사용된다. 검색을 위한 키워드는 문헌이나 기사의 내용을 나타내는 주제어뿐만 아니라 문장중에 있는 단어를 망라하여야 한다. 즉 문장중에서 명사에 해당하는 단어를 추출하여 색인어로 사용할 필요가 있다. 명사를 추출하고 자동으로 색인하기 위해서는 형태소 분석기가 사용되는데 국내에서도 이에 대한 연구가 오래전부터 행하여져 왔다[2]. 본 연구에서는 KAIST에서 구현한 형태소분석기를 사용하였으며 그 처리순서는 다음과 같다.

1) 형태소 분리 : 명사와 조사, 동사와 어미 등의 형태소를 나누는 역할을 한다. 사전에 없는 단어에 대해서는 미등록어의 품사로 추정한다. 동사, 부사, 어미, 조사 등은 색인어 후보로서는 의미가 없으나 미등록어 추정에서 명사로 잘못 해석되는 것을 방지한다.

2) 중의성 처리 : 형태소 분리의 중의성을 해결하기 위해 간단한 Tagger를 사용하였다. 이 Tagger는 형태소의 종류와 분리 갯수 등에 대한 휴리스틱을 적용하여 여러 결과 중에서 하나만 선택한다.

3) 명사 추출 : 명사 추출은 형태소 분리의 결과에서 불필요한 형태소와 품사를 제거하여 명사만을 출력한다. 이때 사전에 없는 단어는 미등록어로 표시를 하며, 복합 명사는 단일 명사의 집합으로 분류한다.

4) 복합명사 처리 : 단일명사의 집합으로 분류된 복합명사는 단일 또는 합성된 명사 형태로 처리한다. 예를들면, A+B의 형식은 A, B, AB라는 명사를 만들고, A+B+C는 A, B, C, ABC로 처리하는데 AB나 BC의 복합명사는 만들지 않는다. 복합명사를 이루는 단어 중 하나라도 미등록어를 포함하면 여기에서 만들어진 복합명사는 미등록어로 취급한다.

5) 불용어 처리 : 이미 만들어진 색인어 후보중에서 불용어로 등록된 단어가 있으면 이를 제거한다.

이렇게 추출된 색인어 후보는 후통제 과정을 거치는데 색인전문가에 의해 색인어 후보를 최종 확정하거나, 이미 확정된 주제어를 사후 관리하거나, 사전에 신규 색인어나 불용어를 추가하는 기능을 수행한다.

사 전

여섯 종류의 사전이 구축되어 형태소 분석과 주제어 부여시 사용된다.

1) 일반 사전 : 형태소 분석을 지원하는 사전으로 체언, 용언, 조사, 어미 등으로 구성되었으며 단어와 해당하는 품사로 구성된다.

2) 고유명사 사전 : 인명과 지명, 사건과 사고 및 행사, 기업과 단체 및 대학명, 신문과 잡지에 포함된 상품명으로 국내외에서 유명도와 빈도수 등을 고려하여 선별된다. 대부분 성격상 새로운 명사가 발생할 소지가 많아 발생시마다 추가되어야 한다. 고유명사 전거화일(Authority File)은 빈번히 등장하는 고유명사 중에 여러가지 표기가 발생하는 용어들을 모아 대표어와 비대표어로 정리해 놓은 용어 집합이다.

3) 주제어 사전(또는 시소러스) : 검색효율을 높이기 위해 사용하는 통제어휘집으로 일반명사와 고유명사 일부가 포함된다. 즉 고유명사중에서 일반명사와 같이 다루기에 자연스러운 고유명사를 포함한다. 일반 주제어는 주제에 따라 크게 46개 분야로 나누어지며 대주제 분류아래서 주제의 종속관계에 따라 상하위관계로 연결되는데 깊게는 6내지 7단계까지 구축된다. 또한 주제색인어의 동의어와 관련어도 포함된다.

4) 동의어 사전 : 기사 본문에 빈번히 등장하여 질의어로 쓰일 수 있는 표기가 다른 동의어, 약어, 외래어 등의 집합이며 주제어 사전의 동의어 관계는 모두 포함한다. 고유명사 전거화일도 포함한다.

5) 불용어 사전 : 색인용어로 부적합하다고 판단되는 단어를 말하며 문장중에 자주 등장한다. 이에 대한 뚜렷한 규정이 없어 경험적으로 판단한다. 의미가 뚜렷하지 않은 일반명사, 수사, 대명사, 동사, 형용사, 부사, 관형사, 감탄사 등이 이에 속한다.

6) 분류표색인 사전 : 주제분류표, 기사종별분류표, 국내 및 국외 지역분류표를 언급한다.

사전구축에는 상당한 인원과 노력을 요하는 작업으로 일부는 국내외에 나와있는 자료를 참조하였으며 DB구축과 더불어 첨가, 삭제 및

조율이 꾸준히 진행되어야 한다.

텍스트 저장

키워드를 사용하여 검색하고자 하는 경우 이를 효율적으로 색인 및 저장 관리할 수 있는 DBMS는 IRS(Information Retrieval System)를 사용하는데 관계형 DBMS에 비해 다음과 같은 잇점이 있다[3].

1) 대용량의 비정형 데이타를 효과적으로 처리한다.

2) DB 및 각종 사전관리에 다양한 기능이 제공된다.

3) 언어체계에 따른 효율적인 처리기능이 있다. 즉, 시소러스 및 분류어 처리, 형태소분석 및 자동색인기능, 동의어와 같은 사전 기능이 제공된다.

4) 다양한 텍스트 검색기능인 불리언, 인접어, 절단, 유사어 등의 검색기능을 제공한다. 반면에 유연성이 적고 다양한 개발이 어려우며 대용량의 정형데이타 처리에 적합하지 않다.

시스템 요구조건으로 대용량 데이타를 신속하게 검색할 수 있으며 하나의 신문기사로 구성된 본문, 제목, 사진설명, 지문제목, 사진 등의 복합문서에 대해서도 효과적으로 저장 및 검색할 수 있어야 한다. 이를 위해 본 연구에서는 기사와 사진의 DBMS로 BasisPlus를 선정하였다.

텍스트 DB의 주요 구성은 기사본문, 서지사항, 주제색인으로 이루어진다.기사 본문은 형태소 분석기에서 후보 색인어가 추출되고 후처리 과정에서 적합성 여부가 판정된다. 서지사항으로 기사 ID, 기사 제목, 게재연월일, 기고자명, 고정물명, 본문 등 관련 필드 항목을 수동으로 입력하거나 CTS 화일 등으로부터 정보가 있는 경우는 자동으로 입력하게 된다. 주제색인 정보로는 주제 키워드, 주제 분류코드, 지역분류코드가 있는데 반자동 또는 수동으로 입력된다. 주제 키워드는 자동색인 과정에서 주제어 사전을 참조하여 후보 주제어를 추천하면 색인 전문가가 최종 판정한다.

텍스트 검색

검색은 세 가지 기능인 브라우징(Browing), 검색조건 입력과 실행, 출력 기능을 제공한다. 먼저 브라우징 기능은 색인항목에 속하는 정보를 훑어보는 것으로 주제어 색인, 주제분류 색인, 단어 색인, 각종 서지정보필드 색인 등을 살펴볼 수 있다.

즉, 시소러스의 경우 주제어 사전으로부터 자모순보기 또는 계층관계 등을 훑어보면서 직접 검색항목을 선정할 수도 있다. 검색조건은 명령어 방식 또는 메뉴방식으로 입력된다. 명령어 방식은 불리언 표현, 범위지정, 검색결과의 조합 등의 조건이 적용된다. 메뉴방식은 GUI 도구인 메뉴, 박스, 버튼, 이이콘 등의 장점을 최대한 살려 이용자와 시스템간의 상호작용성이 높은 대화형 인터페이스를 일관성있게 유도해 나간다. 출력은 검색결과를 화면이나 프린터로 출력하는 기능으로 조판형태로도 볼 수 있다.

Rasterizer

입력변환기에서 CTS에 저장된 조판정보를 화면상에 적절한 형태로 배치되도록 조판데이타를 변환한다. 즉 신문지면상에 위치적으로 분리된 기사와 기사, 기사와 사진, 기사와 지문의 연결정보와 함께 적절한 형태로 변환된다. 최종적으로 실행결과에 대해 신문의 원본과 대조 및 확인 작업을 거친 후 표준문서양식인 SGML로 변환하여 DB에 저장된다. 기사 Rasterizer는 SGML형태로 변환된 조판정보를 화면이나 프린터에 출력하는 기능을 수행한다.

2.2 이미지 데이타베이스

신문사의 경우 하루 300장 정도의 사진자료가 발생하며 1/2정도는 신문에 실린다. 이미지 DB는 수백만장의 사진과 필름을 이미지형태로 저장, 관리 및 검색할 수 있도록 시스템이 구축되어야 한다. 또한 각 화상에 해당하는 정보를 효과적으로 표현하고 색인할 수 있는 방법이 고안되어야 한다. 본 연구에서는 주제키워드, 주제분류, 지역분류, 특성정보, 촬영정보, 사진설명 등의 사항을 상세 분류하여 필드항목을 설정하여 색인 및 검색할 수 있도록 하였다. 그림 2는 이미지 DB구축을 위한 업무 흐름도이다.

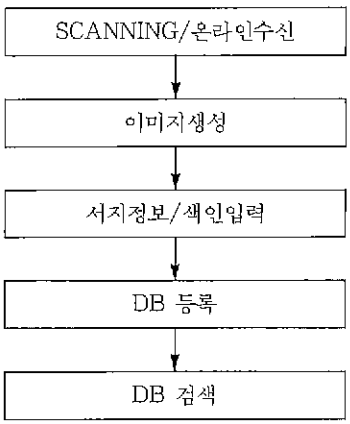

그림 2 이미지 데이타베이스 구축을 위한 업무흐름도

이미지 Quality

검색조건을 입력하여 원하는 화상을 찾을 때 조건에 맞는 화상이 여러 개 존재할 수 있다. 이때 해당하는 화상에 대해 빠른 시간안에 응답이 이루어져야 한다. 본 연구에서는 효율적인 검색을 위해 한장의 사진에 대해 해상도가 다른 세가지 종류의 이미지를 만들어 저장한다.

1) Thumbnail : 128 X 192, 256 Color
2) Preview : 512 X 768, 256 Color
3) Full-size : 2048 X 3072, True Color

Thumbnail은 빠른 응답을 위해 20개까지 하나의 화면에 표시되고 그 중에 순서를 바꾸거나 불필요한 사진은 제거하여 원하는 사진만으로 목록을 만들 수 있다. Preview는 한 화면에 한 화상만을 보아 자세히 확인할 수 있으며 고해상도로 출력할 때는 Full-size를 불러내어 프린트한다. Thumbnail이나 Preview는 소프트웨어로 복원할 수 있지만 Full-size는 전용 하드웨어가 필요하다.

이미지 입력

입력작업은 고속스캐너를 이용하여 입력전용 스테이션에서 이루어진다. 스캐닝 작업은 낱장 또는 일괄로 여러장 처리된다. 이때 스캐너로 입력된 이미지는 압축되어 입력스테이션의 하드디스크에 임시 저장된다. Full-size 이미지는 JPEG 보드상에서 압축된다. 임시저장된 화상은 입력 또는 편집 스테이션에서 한장씩 이미지의 상태를 확인하면서 삭제하거나 변경할 수 있다. 이때 색인 작업도 함께 이루어진다.

이미지 색인

사진과 같은 이미지를 표현하는 방법은 국내외에서도 연구단계에 있으며 대부분 초보적인 분류 기준을 적용하고 있다[4]. 본 시스템도 제한적이지만 경험을 바탕으로 다음과 같은 분류 기준을 만들어 색인에 사용하였다.

1) 주제특성에 의한 분류
   - 감각정보 : 감각적인 느낌
   - 인물정보 : 직업, 나이, 성별 등
   - 환경정보 : 계절, 날씨, 지리적 위치 등
   - 물체정보 : 동물, 식물, 건물 등
2) 주제어 분류 : 주제어 코드, 주제분류어
3) 지역분류 : 지역분류코드, 지역분류어
4) 신문에 게재된 사진설명에 대한 색인
5) 신문제목(캡션) 및 그에 대한 색인
6) 촬영정보 : 일시, 사람, 장소, 거리, 각도 등

분류기준에 대해서는 계속적인 개선이 이루어지겠지만 색인방법에 대한 이론적 연구가 선행되어야 한다.

이미지 저장 및 검색

이미지정보는 텍스트정보에 비해 대용량을 요구하므로 CD-ROM이나 광저장 장치가 필요하다. 또한 서버와 클라이언트사이에 이미지 데이타가 옮겨질 때 네트웍상의 트래픽을 고려하여 설계되어야 한다.

입력작업이 끝나면 네트웍을 통해 색인정보와 Thumbnail은 데이타베이스에 저장되고 Preview 및 Full-size는 2차 저장장치인 CD-ROM 디스크에 저장된다. 검색시 신속하게 확인할 수 있도록 Thumbnail은 데이타베이스에 직접 저장하지만 검색빈도가 적고 대용량을 요구하는 이미지는 속도는 다소 늦지만 경제성을 고려하여 CD-ROM 디스크용 Auto Changer를 이용하여 저장한다.

2.3 오디오 및 비디오 데이타베이스

방송자료실에서 취급하는 비디오와 오디오의 프로그램은 방송에 나간 방송본, 편집원본에

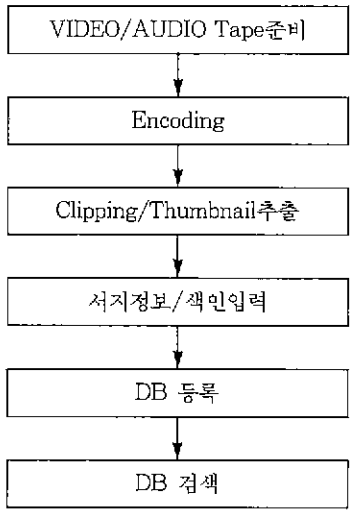

그림 3 오디오/비디오 데이타베이스 구축을 위한 업무흐름도

해당하는 취재본, 외국에서 입수된 외신본을 대상으로 다루고 있다. 본 시스템에서는 테이프에 녹화되어 있는 아날로그 데이타를 디지탈화하여 데이타베이스를 구축한다. 그림 3에서 보는 바와 같이 DB를 구축하기 위해서 엔코딩, 클립핑, 색인, DB 저장, DB 검색의 작업순서를 거치게 된다.

엔코딩(Encoding)

여기에서는 아날로그 테이프가 디지탈화 및 압축된다. 비디오와 오디오 테이프는 MPEG-1으로 각각 실시간 압축된다. 본 시스템의 주요 목적은 영상의 질보다 원하는 부분을 손쉽게 찾고자 하는데 주안점을 두었다. 이 작업에서는 테이프에 저장되어 있는 하나의 프로그램이 하나의 MPEG화일로 변환되어 하드디스크에 임시 저장된다. 이때 엔코딩에 필요한 매개변수가 미리 설정된다.

클립핑(Clipping)

클립은 프로그램의 내용에 따라 적당한 크기로 분할된 하나의 검색단위로 정의한다. 뉴스의 경우 하나의 주제에 대해 앵커가 내용을 소개하고 기자가 자세히 보도하는 통상 5분에서 10분정도의 영상을 하나의 클립으로 분할 할 수 있다. 클립을 생성하기 위해서 프로그램에 대한 관련정보를 사전에 수집분석하여 주제별로 또는 단락별로 구분한다. 엔코딩된 자료를 재생하면서 클립에 해당하는 부분을 표시하여 시작시간과 끝시간을 기록한다.

동화상 클립의 경우 내용전체를 함축적으로 표현하기 위해 클립중에서 대표적인 정지화상들을 추출하여 화면상에 나열한다. 즉 내용의 부분부분을 대표할 수 있는 Thumbnail을 모아 연결할 경우 동화상 전체를 본 것과 같은 효과를 얻도록 한다[5]. 이러한 효과는 검색시간을 절약하면서 원하는 클립을 찾을 수 있는 효율적인 방법이다. 클립편집기에서는 Thumbnail을 뽑아낼 수 있도록 해당 부분에서 느린 동작으로 프레임을 하나씩 화면상에 보여준다.

오디오 및 비디오 색인

프로그램관련 정보로는 제작진, 프로그램 편성, 프로그램 회별, 클립, 취재본, 외신본 등에 여러 필드가 주어지며 각 필드에 필요한 정보가 입력된다.

특히 비디오나 오디오 클립의 경우 Narration이나 대사에 해당하는 부분은 문자로 변환되어 형태소분석기를 거쳐 색인어가 추출된다. 텍스트 DB와 동일한 방법으로 형태소분석을 통해 색인어 후보가 추출되고 후통제 과정을 거쳐 최종 색인어가 선택된다. 또한 관련 서지정보도 함께 입력된다.

오디오 및 비디오 저장

시스템구축을 위해 채택한 데이타베이스 시스템은 UniSQL이다. UniSQL은 관계형 DBMS의 기능을 수용하면서 또한 객체지향적(Object-oriented) DBMS의 기능도 갖춘 시스템이다[6]. 기존의 관계형 데이타베이스 시스템은 고정된 데이타형태(숫자, 문자, 날짜 등)만을 지원하고 유기적으로 서로 내포되어 있는 데이타를 표현하기가 어려우며, 테이블의 행과 열에 표현할 수 있는 데이타는 오직 한가지 형태의 한가지 값만 갖도록 규정되어 있다. 이러한 제약은 데이타 모델링시 실세계의 객체들의 연관관계 및 이들의 구성요소를 논리적으로 구성하는데 많은 단점이 된다. UniSQL은 관계형

데이타베이스 시스템의 기능외에 다음과 같은 확장 기능을 통하여 실세계의 현상을 자유롭게 표현할 수 있게 한다.

1) 사용자가 임의의 데이타 형태를 정의할 수 있다.

2) 테이블의 행과 열에는 한개 이상의 오브젝트가 존재할 수 있다.

3) 테이블에는 행과 열을 조작하는 프로그램을 정의하여 등록할 수 있다.

4) 같은 종류의 오브젝트들을 클래스로 정의할 수 있고 클래스는 상위 클래스(Parent Class)로부터 모든 속성을 상속(Inheritance) 받을 수 있으며, 하위 클래스(Child Class)로 상속시킬 수 있다.

색인정보는 대부분 UniSQL에 저장되며 색인어관련 정보는 BasisPlus에 저장된다. 비디오 또는 오디오 화일은 대용량으로 비디오서버에 저장되며 각 화일의 포인터만 UniSQL에 색인정보와 함께 저장된다. 비디오서버는 동시에 여러 사용자에게 비디오가 전혀 끊어짐이없이 네트웍상에 제공하는 역할을 한다.

2.4 방송자료의 데이타 모델

그림 4는 본 연구에서 방송자료에 대한 데이타 모델의 클래스간 계층구조와 연관관계를 나타낸다.

본 시스템의 관리대상자료는 크게 라디오 방송자료(Audio)와 TV 방송자료(Video)로 나누어 진다. 라디오 방송자료는 가요, 팝, 클래식 등 음악자료를 제외한 음향(기록녹음)자료만을 대상으로 하며 방송본과 취재본 두종류가 있다. TV 방송자료는 방송본, 취재본, 외신본 세종류로 구분하여 관리한다. 방송본이란 TV나 라디오를 통해서 방송된 프로그램을 말하며 프로그램명, 방송일자, 회차 등의 고유정보를 가진다. 취재본은 특정 프로그램 제작을 위해 촬영 혹은 녹음하였거나 프로그램 제작의 기초자료로 활용하고자 입수하여 보관하는 자료를 말한다. 외신본은 외국 방송사에서 방송되었던 TV뉴스 프로그램으로 외국에서 입수하여 보관하는 자료이다.

Encoding은 프로그램 단위별로 이루어진다. 예를 들어 뉴스 1회분은 하나의 MPEG화일로 생성된다. 생성된 MPEG화일은 색인 및 서지정보와 함께 등록되어 사용자는 서지사항 뿐만 아니라 MPEG화일의 내용을 화면상에서 확인할 수 있으므로 정확한 자료를 얻기가 용이해진다. 또한 사용자는 등록된 MPEG화일의 내용을 모니터링하면서 주제별 혹은 아이템별로 클립핑하여 하나이상의 클립을 생성할 수 있

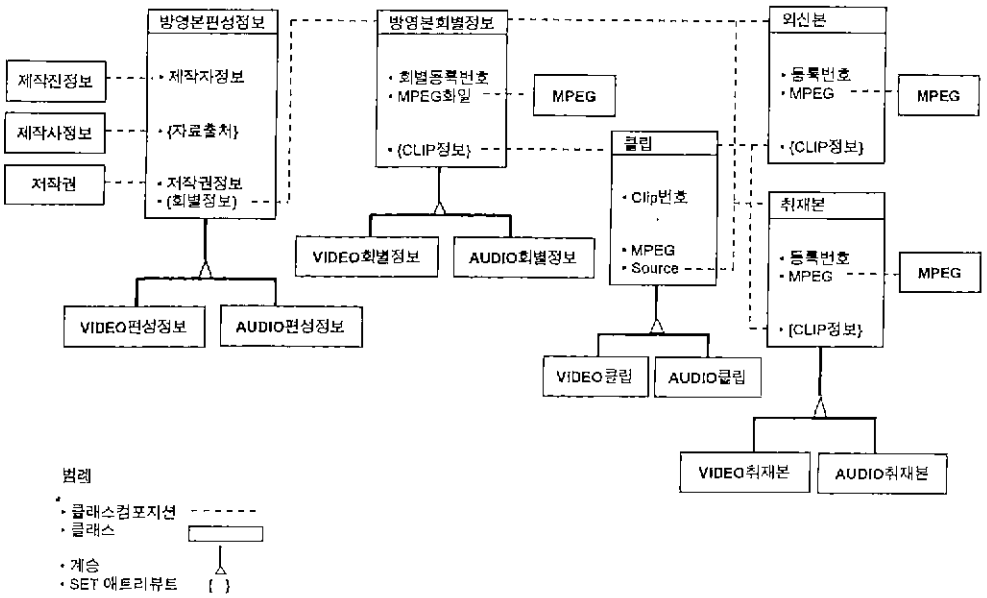

그림 4 방송자료의 데이타모델

다. 이때 MPEG화일은 물리적으로 분리되지 않으며 클립은 단지 시간정보(시작시간과 끝시간)로 구분되는 논리적 단위이다. 사용자는 자료 검색시 프로그램 단위별 혹은 클립 단위별로 검색할 수 있으며 클립 단위에서 클립이 속한 프로그램으로 프로그램 단위에서 현 프로그램에서 생성된 클립 단위로 자유롭게 브라우징할 수 있다.

클래스 및 클래스간의 계승과 연관관계에 대한 각각의 설명은 다음과 같다.

1) 방송본 편성정보 : 라디오나 TV방송을 위해서 새로운 프로그램이 편성되었을 때 발생하는 정보로 편성코드번호, 분류코드, 편성프로그램명, 원제명, 장르, 제작기획서, 제작자정보, 채널, 편성시간, 저작권 등이 있다.

그림 4에서 보는 바와 같이 애트리뷰트 제작자정보의 도메인은 제작진정보 클래스이다. 따라서 한 프로그램에 참여한 모든 제작자의 정보는 제작진정보에 등록되고 방영본편성정보 클래스에서 애트리뷰트-도메인간의 링크를 통해서 제작진정보를 직접 포인팅하는 구조이다. 만일 제작진중 새로운 직종을 추가해야할 경우 방영본편성정보 클래스는 변경할 필요없이 제작진정보클래스만 변경하면 되므로 스키마 변경이 용이해진다.

또한 어떤 편성프로그램은 다수개의 회별프로그램으로 구성될 수 있다. 그림에서 애트리뷰트 회별정보의 도메인은 방영본회별정보 클래스이고 {}로 표시되어 있다. {}표시는 애트리뷰트가 하나이상의 데이타값을 가질수 있음을 의미한다. 따라서 하나의 편성프로그램에서 나온 모든 회별정보를 직접 포인팅할 수 있으므로 관련 회별정보와의 연관성을 자연스럽게 표현할 수 있으며 불필요한 데이타의 중복저장이나 빈 공간을 갖는 일이 없어지고 질의문이 간단해진다.

2) Video편성정보, Audio편성정보 : 방송본은 라디오 방송본과 TV 방송본으로 나누어진다. 라디오 방송본인 경우 미디어형태는 오디오이고 TV방송본인 경우 비디오(오디오 포함)이다. Video편성정보 클래스는 TV방송본에 관한 정보를, Audio편성정보 클래스는 라디오 방송본에 관한 정보를 관리한다. 방송본의 필연적인 정보는 상위 클래스인 방송본편성정보에서 모두 계승받는다. 추후 라디오나 TV방송의 특성별로 추가해야 할 정보는 해당 클래스의 독립적인 변경을 통하여 관리할 수 있다.

3) 방송본 회별정보 : TV나 라디오 프로그램의 매회별 방송시 발생하는 정보로 회차, 부제명, 소제명, 방송일시, 방송길이, 전문, 초록, 큐시트 등이 있다. 실제 회별 프로그램의 내용은 하나의 MPEG화일로 변환되어 등록된다. 그림 4에서 애트리뷰트 MPEG화일의 도메인은 MPEG클래스이다. MPEG화일 자체에 관한 모든 정보는 MPEG클래스에서 관리하고 방송본회별정보 클래스에서는 관련 MPEG화일을 직접 포인팅하고 있다. 이러한 구조는 MPEG화일에 대한 새로운 애트리뷰트를 추가해야 할 경우 MPEG클래스만 독립적으로 확장할 수 있게 한다.

4) Video회별정보, Audio회별정보 : 방송본 회별정보의 필연적인항목은 상위클래스인 방송본회별정보에서 계승받고 라디오나 TV방송의 특성별로 확장해야할 경우를 고려하여 2개의 클래스로 나누어 모델링하였다. 라디오방송본은 Audio회별정보에서, TV방송본은 Video회별정보에서 관리한다.

5) 클립 : MPEG화일로 변환되어 등록된 방송자료를 모니터링하면서 주제별 혹은 아이템별로 구분하여 클립을 생성할 수 있다. 하나의 MPEG화일에서 다수개의 클립이 생성될 수 있다. 클립 클래스는 클립에 관한 정보를 관리하는 것으로 클립번호, 클립제목, 시작시간, 끝시간 등을 가진다. 클립과 원본과의 관계는 Source 애트리뷰트를 통하여 해당 원본을 직접 포인팅하는 구조로 모델링하였다.

6) Video클립정보, Audio클립정보 : 클립에 관한 필연적인 정보는 상위 클래스인 클립클래스에서 계승 받으며 Video클립의 특성상 필요한 색채구분, 촬영기법, 대표이미지, 오디오유무와 같은 정보를 가진다. Audio클립을 관리하기 위한 것으로 클립에 관한 필연적인 정보는 상위 클래스인 클립클래스에서 계승받으며 Audio의 특성상 부가되는 정보는 독립적인 클래스 확장을 통하여 구현할 수 있다.

7) 취재본 : 취재본에 관한 정보를 가지고 있

다. MPEG화일 및 생성된 클립과의 연관성은 방송본회별정보의 구현방법과 동일하다.

8) Video취재본, Audio취재본 : 취재본에 관한 필연적인 항목은 상위 클래스인 취재본클래스에서 계승받는다. 라디오 및 TV프로그램의 특성상 새로이 추가될 항목을 고려하여 라디오용 취채본은 Audio취재본에, TV용 취재본은 Video취재본에 저장한다.

9) 외신본 : 외신본에 관한 정보를 가지고 있다. 라디오 방송자료에는 외신본은 발생하지 않는다.

3. 시스템 구현

전체 시스템의 형상은 그림 5에서 보는 바와 같이 분산된 네트웍상에서 다음과 같이 4개의 시스템으로 나누어진다.

미디어 입력 시스템

각종 미디어 자료가 외부의 네트웍을 통해서 또는 입력용 단말기에서 입력작업을 통해서 입력이 된다. 기사의 경우는 CTS 컴퓨터에서 화일형태로 가져오고 사진의 경우는 스캐너나 CD-ROM 드라이브를 통해 입력된다. 오디오나

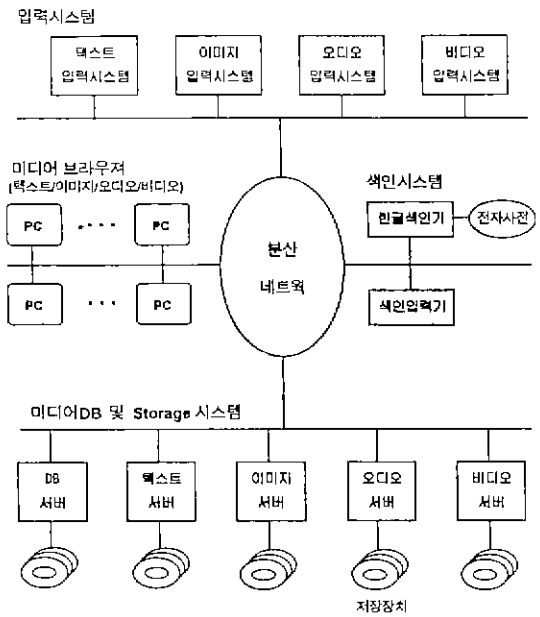

그림 5 MMIS Overall Configuation

비디오의 경우는 어낼로그 레코더에서 나오는 신호를 엔코딩하고 클립핑 추출작업을 통해 자료를 입력한다.

미디어 색인 시스템

입력된 자료는 하드디스크에 임시 저장된 후 DBMS에 저장되기 위해 색인과정을 거친다. 기사는 한글 형태소분석기를 통해 명사가 추출되어 색인어로 선택되고 각종 서지사항을 자동 또는 수동으로 입력하게 된다. 사진의 경우 스캔닝 확인과정을 거친후 서지항목과 사진설명에 대해서는 색인어를 추출하여 입력한다. 오디오와 비디오도 프로그램과 클립에 관련된 필드정보뿐만 아니라 클립에 담긴 Narration도 텍스트로 변환된 후 형태소분석을 통해 색인어가 추출된다.

미디어 데이타베이스와 Storage 시스템

검색기능 수행시 응답시간에 가장 중요한 영향을 미치는 요소는 데이타베이스 구성과 저장시스템 설계이다. DBMS와 Storage시스템은 서버에 장착되어 동작하는데 모든 색인자료는 DB에 저장되고 기사는 하드디스크에, 사진의 Thumbnail은 하드디스크에, Preview와 True-size는 CD Jukebox에 저장된다.

음향과 비디오의 경우 비디오서버의 하드디스크 어레이에 저장된다. 비디오서버는 동시에 다수의 사용자에게 끊어짐이 없는 비디오 또는 오디오를 제공하는 기기로서 클라이언트와 함께 네트웍상에 병목현상이 없도록 설계되어야 한다. 본 과제에서는 Starlight사의 Starworks를 도입하여 155 Mbps의 ATM 스위치를 통해 다수의 서버와 직접 연결되거나 ATM-To-Ethernet 스위치를 통해 기존에 Ethernet을 사용하는 서버 또는 클라이언트와 연결된다. 비교적 검색이 빈번히 행해질 자료는 1차 저장수단인 디스크어레이에 내용이 Striping되어 저장되고 그 이외의 자료는 2차 저장수단인 테이프 Jukebox 등을 사용하여 저장된다.

미디어 브라우저

클라이언트-서버 구조로 분산된 환경하에서 서버는 미디어에 따라 각각 다른 DB인데 클라

이언트는 여러 사용자가 동시에 각종미디어를 하나의 PC상에서 볼 수 있도록 구현하였다. 미디어 브라우져는 GUI를 이용하여 사용하기 쉽도록 하였으며 브라우징, 검색 및 출력하는 기능이 제공된다.

기사 브라우징은 특정 DB정보를 편리하게 볼 수 있도록 지원하며 네비게이션을 통해 원하는 데이타에 쉽게 접근할 수 있다. 브라우징 대상으로는 시소러스, 주제색인필드, 검색이 가능한 서지정보 필드, 기사제목 등이다. 검색기능으로는 불리언, 범위지정, Sorting, Wild Character, 조합검색 등으로 SQL 형태로 변환하여 DBMS에서 수행된 결과를 얻는다. 수행된 결과나 브라우징된 결과는 화면이나 프린터로 출력하거나 클립보드에 복사된다. 특히 기사의 레이아웃은 기사 Rasterizer를 통해 화면상에 보여진다.

이미지 브라우져는 사진정보가 저장되어 있는 이미지 DB의 자료를 사진의 주제분류, 감각, 인물, 환경 정보 등과 함께 Thumbnail, Preview이미지를 통해 쉽게 브라우징, 검색, 출력하는 기능을 제공한다. Full-size이미지는 JPEG 보드가 이용하여 볼 수 있다.

오디오 및 비디오 브라우져는 방송에 관련된 프로그램을 참조 관계로 모델링하였는데 실제 클립자체는 클래스인 방영본, 취재본, 외신본 등에서 생성되므로 원 프로그램의 일부분이다. 브라우징 기능은 참조 관계를 이루는 객체 항목들의 내비게이션을 통해 필드 값을 볼 수 있고 클립을 재생하거나 대표화상을 살펴보면서 원하는 클립을 찾을 수 있다. 또는 키워드를 사용하여 관련 객체의 항목을 검색할 수 있고 그 결과로 부터 브라우징 기능을 수행할 수 있다.

4. 현재 Status 및 향후 계획

본 연구에는 사용자와 개발자가 함께 참여하였으며 사용자인 언론사는 원하는 요구사항을 돌출하고 개발자는 그 요구에 알맞게 개발하여 실용적으로 사용할 수 있는 시스템을 목표로 진행되었다. 제 1단계는 통상산업부의 중기거점과제로 1993년부터 3년과제로 진행이 되고 있으며 전자부품종합기술연구소가 주관하여 언론사인 조선일보, 한국경제, KBS, MBC가 참여하고 있고 솔빛조선미디어, 휴먼컴퓨터, KAIST 등 총 11개 기관이 컨소시움을 이루어 공동 개발하고 있다.

현재 신문과 사진에 대한 개발이 완료되어 DB를 입력하고 있으며 대량의 DB 구축과 동시에 테스트 및 성능개선이 이루어질 예정이다. 방송자료는 DB설계가 끝나고 구현 중에 있으며 곧이어 대량의 DB를 구축할 예정이다. 현재는 미디어간에 연관관계가 거의 적용되지 못하고 있지만 향후 하이퍼링크의 기법을 적용하여 미디어간의 브라우징을 자유롭게 할 것이다. 제 2단계에서는 일반인을 대상으로 언론사가 보유하고 있는 대용량의 자료를 검색할 수 있는 시스템을 구축해 나갈 예정이다. 앞으로 활발히 전개되고 있는 초고속정보통신시대에 유용한 정보를 제공하는 역할을 수행해 나갈 것이다.

참고문헌

[1] 김종태외 다수, "멀티미디어 종합 정보 처리 시스템 개발," 통상산업부 연구보고서, 1995.

[2] 조성호외 다수, "한글 자동색인 및 키워드 검색 시스템 개발," 정보통신부 연구보고서, 1995.

[3] Frakes, W.B. and Baeza-Yates, R., "Information Retrieval Data Structures & Algorithms," Prentice Hall, 1992.

[4] Tatsuo, K., "Bibliographic Description of Photo Databases : The Case Study of Expo 90, The Photo Museum," 정보관리, Vol.33 No.1, Japan, April 1990.

[5] Smoliar, S. W. and Zhang H., "Content-Based Video Indexing and Retrieval," IEEE Multimedia, Vol. 1, No. 2, pp. 62-72, 1994.

[6] UniSQL Inc., "UniSQL/X Database Management System User's Manual," May 1995.

김 종 태

1977 서울대학교 공과대학 전기공학과, 학사
1985 University of Florida 전자공학과, 석사
1990 University of Florida 전자공학과, 박사
1977~83 국방과학연구소 연구원
1990~93 삼성전자 SI사업부 책임연구원
1993~현재 전자부품종합기술연구소 수석연구원
관심분야 : 인공지능, 멀티미디어 정보시스템, 정보단말기

정 혜 윤

1991 한양대학교 공학대학 전자계산학과, 학사
1993 한양대학교 대학원 전자계산학과, 석사
1993~현재 전자부품종합기술연구소 전임연구원
관심분야 : 데이타베이스, 멀티미디어 정보시스템
신문과 방송 자료의 데이타베이스 시스템 구축

●'96 컴퓨터시스템 동계 학술대회●

- 일 자 : 미정(1996년 2월 예정)
- 장 소 : 추후결정
- 내 용 : 연구 발표 및 토의 등
- 주 관 : 컴퓨터시스템연구회
- 문 의 : 중앙대 컴퓨터공학과 김성조 교수
  T. 02-820-5307