국외출장 결과 보고

-OCR 자료처리 신기법 습득과 자료수집-

Ⅰ 개요

1. 목적

○ 인구주택총조사 등 대규모 조사의 보다 신속·정확한 자료 처리를 위한 최신기법 습득 및 관련자료 수집

2. 방문기관 및 일정 등

○ 출장기관 : 뉴질랜드 통계청 (Statistics New Zealand)
- Auckland : 이미지 센터 (Imaging Centre)
- Christchurch : 총조사 자료처리 개발(Census Processing Developing)

○ 출장기간 : 2003. 12. 3 ~ 12. 11(9일간)

○ 출장자 : 정보처리과 5급 김규영, 6급 최인범
인구조사과 6급 정남수

3. 주요 수행 내용

○ 2001년 센서스시 OCR 자료처리 방법
- 사용장비, 입력방식(분산/집중), 조사표 설계 등

○ 2006년도 센서스 방향 파악
- OCR 자료처리 및 인터넷 조사 등

○ 기타, 자동코딩 방법

Ⅱ 방문기관 및 일정

□ Imaging Centre (오클랜드)

월/일	시간	설명자	주제
12월 5일	9:00~9:30	자료수집과장 Ray Freeman	이미지 센터 개요
	9:30~10:00	〃 Sean Keefe	이미지 기술 개요
	10:30~11:15	Susan Lloyd Angela Fong	이미지 실습
	11:15~12:30	Stuart Pitts Angela Fong	2006년 센서스의 시험조사(11월)방법 코드 단어사전 개발
	13:30~15:00	국제담당과장 Keith Sykes	질의 응답

□ Census Processing Developing (크라이스처치)

월/일	시간	설명자	주제
12월 8일	9:15~10:00	Neil Martin	2001년 센서스 개요
	10:15~12:00	Craig Lange	자동인식과 OCR
	13:30~15:00	Diane Trevella	인식데이터의 코딩
	15:30~16:30	Neil Martin	질의 답변
12월 9일	9:00~10:00	Ian Smith	인터넷 조사 개요
	10:15~12:00	〃	인터넷 조사와 RMS

Ⅲ 수행 내용

1. 총괄

○ ICR(Intelligent Character Recognition)/OCR 시스템은 1993년 개발 착수하여 1996년 센서스에 처음 도입

- 이때, 자료전환 등의 문제로 결과적으로 ICR 시스템에 의한 자료처리가 실패하여 큰 손실을 입었으나, 동일한 멤버로 계속 발전시켜 2001년 센서스에서는 큰 성공을 거둠

○ 2006년 센서스시에는 ICR 시스템을 사용할 뿐아니라 약 10%의 가구에 대해서는 인터넷 조사방법 실시 예정

○ 현재 뉴질랜드 통계청이 담당하는 22종의 통계조사중 6종의 통계 조사에 대해서는 ICR 시스템으로 자료처리하고 있음

- ICR 시스템 도입후 연간 NZ $900천(총 비용 NZ $ 2,000천)의 비용을 절약하고 있으며, 직원도 약 30명 감원

- 스캐너는 코닥 i840 (속도 : A4기준 140PPM) 1대와 중저속 1대 (코닥i260) 등 총2대를 사용하여 처리

2. 2001년 인구주택총조사 시스템

○ 외부용역에 의해 처리하였으며 Micro Editing은 내부에서 처리

- 용역비용 : NZ$ 350만불, 용역업체는 IBM이 수행

○ 스캐너(구형)는 중저속 2대 사용(스캐너:Kodak 9500)

○ 1주일 6일, 1일 2교대의 11시간 스캔 작업(스캔 소요일 : 48일)

- 오전팀 : 06:45 ~ 14:45(8시간)
- 오후팀 : 15:00 ~ 23:00(8시간)
  - 식사시간 30분씩 제외, 1시간당 10분 휴식, 아침 및 저녁 티타임 15분 등을 제외한 실작업시간은 11시간

○ 인원 : 80명(6개팀으로 구성)

○ 처리기간 : 3월말 ~ 9월말(6개월)

- 야간에는 Batch작업으로 자동인식 작업

- 1일 140,000매 처리(총 처리매수는 약 550만매)

○ 운반용 상자는 별도 제작하여 사용

○ 색인키에 대해서는 가구별로 Patch코드를 사용하여 처리하였으며, 동일가구가 여러 장의 조사표로 구성되어 섞일 경우 별도의 해결 방법이 없으므로 관리에 주의하였음

○ 인식엔진 : 독일제품으로 OCE Recostar(NZ $ 130,000)

- 엔진기능에 농도를 조절할수 있는 기능이 있어 더블마크나 잘못 표시한 경우 구분 가능
  (올바른 표시 :  , 잘못 표시한 경우 : )

- 알파벳 인식률을 높이기 위하여 사전 기능을 이용한 매치 방법 사용 (예 : Canaba 인식 → Canada로 전환)

* 매치율 : 국가명87%, 지역명80%, 인종명75%, 직업40%
평균 알파벳 인식률은 40 ~ 80% 수준임

→ 우리나라의 경우도 행정구역의 경우 한글인식을 시도할 필요는 있음

○ 인식후 자료검증 방법 : 4단계로 검증

- Carpets → Triples → Fields → Form

· Carpets : PC 화면에 91자(7자×13자)의 숫자를 띄어 Supersure(95% 이상의 신뢰도) 부분은 녹색으로 표시, Sure 부분은 노란색으로 표시하여 육안 확인후 교정
· Triples : 3개의 숫자 또는 문자를 이미지와 인식한 내용을 화면에 띄어 육안 확인후 교정
· Fields : 조사표상의 필드를 띄어 숫자 또는 문자 교정
· Form : 조사표 전체 폼을 띄어 불확실한 부분은 빨간 박스로 표시되어 육안 확인후 교정

○ 조사표 설계 : 별도의 조사표 설계 전문가에 의해 조사표의 형태, 인식박스 크기 등을 설계

- 하나의 문자 box 크기는 최소 6㎜×5㎜이고 자간 간격은 0.5㎜임

- Mark 표시란은 √가 상호 침범하지 않도록 간격을 두는 것이 좋으며, 가급적 √보다는 －를 선호

- 위의 모든 사항을 한 페이지내에 가급적 많은 조사항목을 배

치하여 스캔비용 및 처리기간등을 단축

○ 인식된 자료는 자동 Repair과정을 거침

- 문자 필드(직업, 종교, 언어) : 모두 자동 repair
- 숫자 필드 : 대체로 자동 repair
- Mark필드 : 자동 repair안됨

○ 작업은 중앙집중형으로 처리

- 미국 : 120대의 스캐너를 사용, 3곳에서 분산 처리
- 호주 : 12대의 스캐너를 사용, 중앙집중으로 처리

○ 스캔 순서는 잘 아는 지역부터 처리하여 경험을 얻은 후 가장 큰 지역을 처리(사전에 처리순서를 결정하였고, 그에 따라 조사표를 접수)

○ 조사항목은 거처 22항목, 인구 43항목으로 구성되어 있고, 조사표는 가구당 거처조사표 1매와 가구원별 조사표(1인 1매)로 구성

○ 2001년도 조사원수는 5,900명이며, 교육은 메니저→슈퍼바이저→조사원 교육의 3단계로 이루어짐

- 교육은 2일간 실시 (조사 개시전과 조사종료전)

- 조사원 수당은 가구 기준으로 지불하였으며 먼거리 조사자의 경우 별도의 수당 지급

3. 인터넷 조사시스템

○ 2001년부터 자체개발 착수하여 2003년 11월에 시험조사 실시

○ 기본적으로 인터넷으로 응답하겠다고 한 가구에만 사용

- 응답자에게 우편조사 또는 인터넷 조사중 선택권을 줌

- 2006년 센서스시에 약 10% 정도를 인터넷 조사대상으로 추정 (가구당 인터넷 보급률 현재 약 35.7%)

- 조사 2주일전에 인터넷 색인키와 PIN(Personal Identification Number)을 사전에 배부

  · 인터넷 ID는 총 11자리 인데 마지막 자리는 Check Digit임

  · PIN 번호는 12자리의 영문/숫자로 Random하게 생성

○ 접속에 필요한 PC사양

- Explorer 5.0 이상 또는 Netscape 6.0 이상
- JavaScript 지원가능
- 최적 화면 Resolution : 800 × 600

○ 자료보안 : 128 bit 암호화 기법사용

- 128 bit SSL(Secure Socket Layer) 사용

○ 인터넷 사용료는 개인이 부담

○ 자신의 가구별 PIN번호를 분실했을 때는 지정된 관련과로 연락하여 다른 번호를 재발급

○ 응답 도중에 Log-off가능하고 가구내의 다른 응답자가 응답하여도 무방

○ 응답 완료후 자료전송한 뒤에는 다시 자료를 불러 수정할 수 없고 관계자에게 연락하여 수정하거나, 메일로 수정할 내용을 송부하면 관계자가 수정

- 조사표 전송시 조사항목이 채워지지 않으면 전송이 않되도록 설계

- 조사항목별로 도움말 기능이 있어 도움말 클릭시 기입 방법 설명이 나타남

○ 조사사항은 응답자의 응답에 따라 필요항목만 Pop-up되는 Filtering Question방법으로 구성

- 조사대상자의 응답 실태는 통계청의 RMS시스템에 의해 실시간으로 관리

Ⅳ 수집자료

○ 별도 첨부