「11차 DaWak 및 제20차 DEXA 회의」
출장결과 보고

1. 출장 개요

○ **출 장 지** : 오스트리아 린츠 / the Johannes Kepler University

○ **출장목적** : 데이터웨어하우스, 데이터마이닝 및 지식창출에 관한 워크샵, 데이터베이스 및 전문 시스템 응용분야 참가 및 최신 국제동향을 파악하여 통계 DW 구축 유지·관리·개선에 필요한 다양한 정보 수집

○ **출장기간** : 2009년 8월 29일 ~ 2009년 9월 6일

○ **출 장 자** : 전산개발과 이지은 주무관

2. 컨퍼런스 개요

○ **행사명** : 제11차 DaWak(Data Warehousing and Knowledge Discovery) 및 제20차 DEXA(Database and Expert Systems Applications) International Conference

○ **행사개요**

- **특 징** : Database, Data Warehousing, 데이터 통합 및 데이터마이닝에 대한 각 국 대학교, 연구소 및 전문민간단체의 전문가들이 참석하여 세계동향 및 연구 결과를 발표하는 회의로서 상호 연구결과를 공유하고 의사소통 강화를 통하여 정보화 사회를 이끌어 가는 주도적인 전문가 집단 회의이며 세계 최대규모의 Database 관련 국제회의임

- **연 혁** : DaWak은 1999년 시작하였으며 DEXA는 1990년에 시작하여 매

년 열리는 국제회의로서 올해는 DaWak는 제11회이며, DEXA는 제20회 행사가 개최되었음

3. 내용 요약

○ 최신 동향 발표

- 정보통합에 대한 최신 동향, 오스트리아 연방 바이오뱅크의 데이터 관리 현황

○ 전문 분야 연구 발표

- 데이터웨어하우징 모델링, 물리적 디자인, 마이닝 패턴, 데이터 큐브, 데이터마이닝 어플리케이션, 데이터마이닝, OLAP 장점, 데이터 및 정보 통합 및 품질, 데이터 및 정보 경향, 데이터 및 정보 모델링, 데이터베이스 및 정보시스템 구조 및 수행, 정부 정보로의 접근

4. 주요 내용

○ 정보통합에 대한 최신 동향

- 발표자 : Laura M. Haas, IBM Almaden Research Center.

- 서 두 : 정보통합은 현대 비즈니스 정보학의 초석이나 광범위한 문제가 있다. 정보를 수집하고 통합하는 것 없이 새로운 어플리케이션을 만들기는 어려우며 정보 통합은 매우 다양한 형태로 이루어 지고 있다. 역사적으로 두개의 중요한 접근방법이 있으며 그것은 데이터 연합과 데이터웨어하우징이다. 오늘날 지저분하고 다양한 데이터가 증가하는 반면에 정보통합은 더욱 역동적으로 이루어지고 있음에 따라 우리는 새로운 접근이 필요하다. 동시에 정보 통합은 통합과정을 좀 더 다루기 쉽고 통합결과를 더욱 사용할 수 쉽게 만들기 위하여 통합된 결과에 영향을 줄 수 있는 어플리케이션과 분석을 더욱 단단하게 연결되어져야 한다.

- 내용 요약 : 통합문제, 기술, 데이터, 원천 및 분석의 다양성은 증가하고

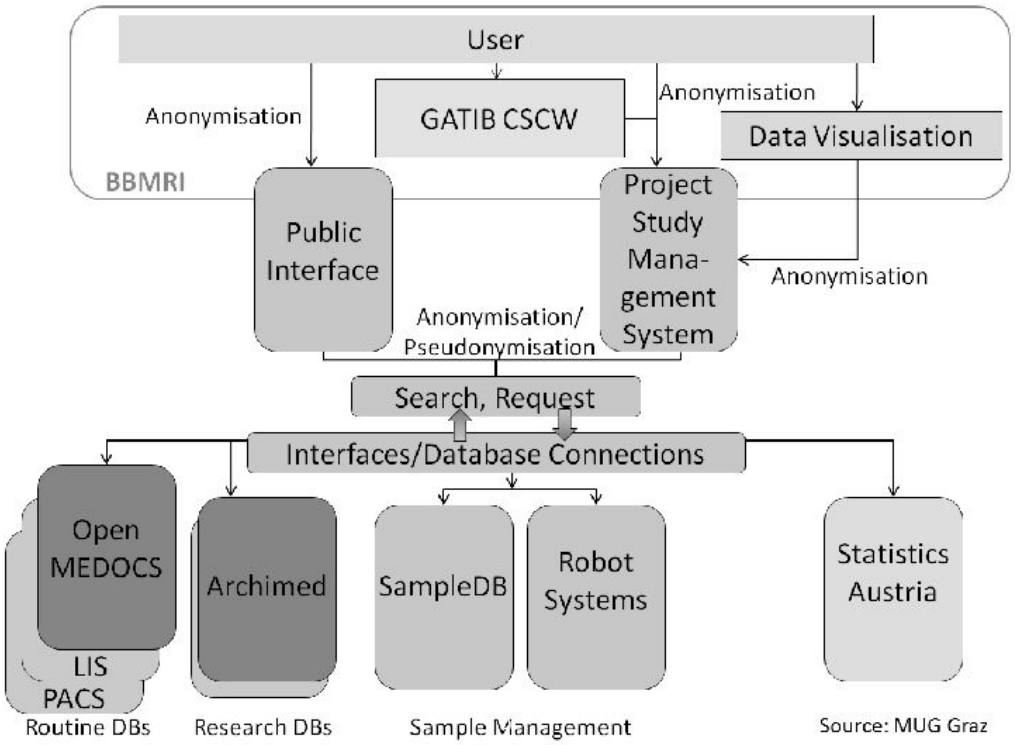
모든 것은 급속도로 변화하고 있는 등 정보 통합은 현재 광범위하고 어려운 난제에 직면해 있는 반면에 더욱 신속하게 통합으로부터 더 많은 가치를 이끌어 내기 위한 압력은 점점 더 요구되고 있다. 이러한 환경의 흐름은 새로운 연구과제로 발전되고 있다. 우리는 다양한 도전과 기회를 발견하고 기술하였다. 통합의 다양성은 전문화된 통합 솔루션과 통합을 위한 디자인 툴 발견을 위한 기회를 이끌어 내고 있으며 이것은 특별한 문제에 적당한 기술조합을 이끌 수 있을 것이다. 역동적인 환경은 통합에 대한 반복적인 접근과 실질적인 필요가 알려질 때까지 진행을 지연시키는 것에 관심을 발생시키고 있다. 자동적으로 영향력 있는 공동체에 의해 제공 되어진 메타데이터는 계속적으로 변화하는 세계에 대한 또 다른 접근이다. 통합과 분석이 분리되어 왔으나 상호 협동적인 접근이 개발되어지고 있다. 이것은 데이터 통합을 어떻게 어떤 통합을 언제, 어떤 분석을 고려해서 할 것인가를 위하여 더욱 계획된 선택을 추가하고 있다. 반면에 속도에 대한 필요는 통합을 정보통합과 분석에 대한 새로운 단계로 연구를 발전시키고 있다. 이런 복잡한 기호논리학에 대한 최종 결과에 대한 이해에 대한 요구는 어떻게 분석물을 공급할 것인가 어떻게 데이터와 통합과정을 시각화를 할 것이며 그것들의 기원을 어떻게 기록하고 표현할 것인가에 대한 새로운 연구로 이끌 것이다. 우리는 이러한 분야에서 끝없는 연구과제를 예상할 수 있으며 젊은 연구자들에게 이 분야 연구를 강력하게 추천한다.

○ 오스트리아 연방 바이오뱅크의 데이터 관리 현황

- 발표자 : Johann Eder. University of Klagenfurtk, Austria

- 내용 요약 : 오스트리아 연방 바이오뱅크는 조직, 혈액 등 생물 소재 및 관련 데이터로서 쉽게 획득할 수 없는 개인 처방과 관련된 데이터로서 생활 스타일 및 유전적 요인에 대한 정보를 수집하고 저장하고 관리하여 약학과 유전자 기능과 의료관련 연구 수행 및 병에 대한 환경의 영향을 분석하기 위하여 필요한 자료를 관리하는 곳이다. 이곳 데이터를 이용하기 위해서 유저는 공용 인터페이스를 거쳐서 익명화된 자료를 이용할 수 있으며 관리자는 이 자료를 제공하기 위하여 영구 DB와 샘플DB 및 Robot 시스템을 이용하여 인터페이스/데이터베이스 연결 자료를 제공하

고 있는 현황을 발표하였다.

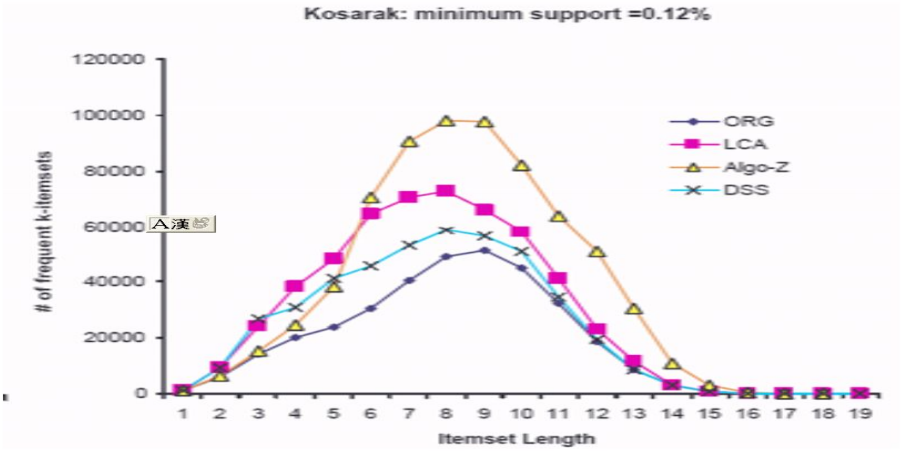


< 오스트리아 바이오 뱅크 데이터 서비스 설계도 >

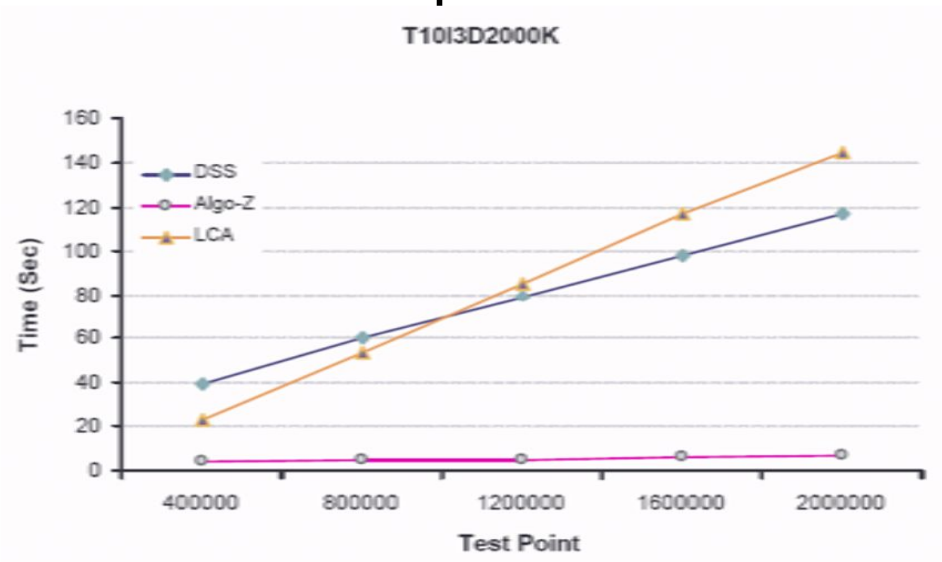
○ 전문 분야 발표 내용

- Which is Better for Frequent Pattern Mining(Approximate Counting or Sampling) (발표자 : Willie Ng and Manoranjan Dash) : 빈도 패턴 분석을 위하여 추정(LCA)과 샘플링(Algo-Z) 중 어떠한 방법이 보다 바람직한가를 비교하기 위하여 LCA와 Algo-Z를 상호 비교한 연구결과임. 이 결과에 따르면 LCA가 Algo-Z보다 좀 더 정확한 결과를 나타냄을 알수 있었으며 샘플링 알고리즘인 DSS는 비록 Algo-Z보다는 시간이 오래 걸리나 LCA보다는 적게 걸린다. 또한 DSS는 두가지 알고리즘인 LCA와 Algo-Z보다 정확도 측면에서 훨신 우월한 것으로 발표하였음

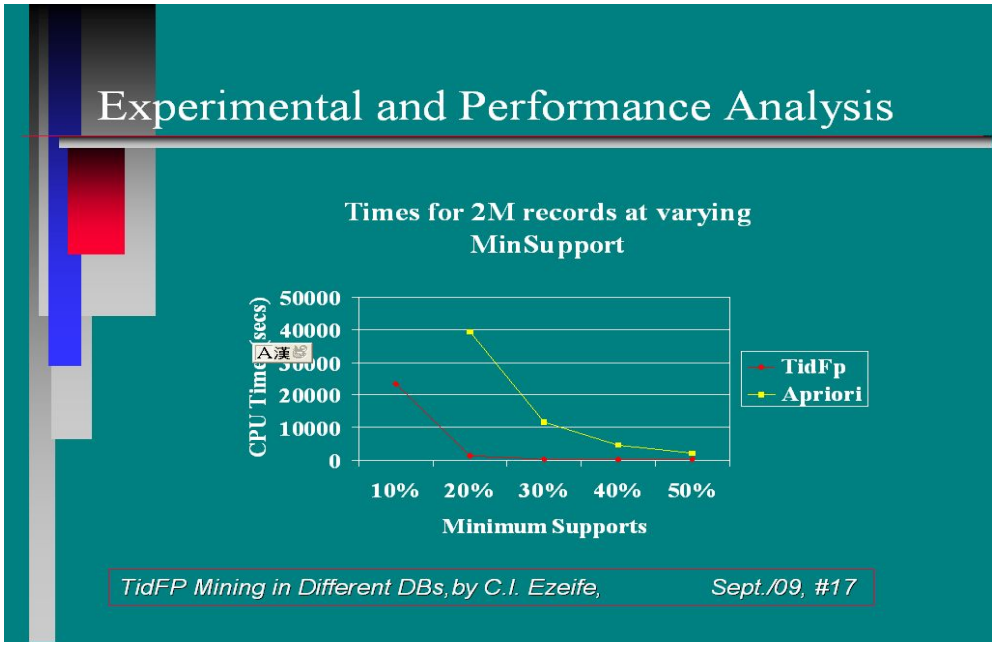
Detailed Accuracy Comparison: Kosarak



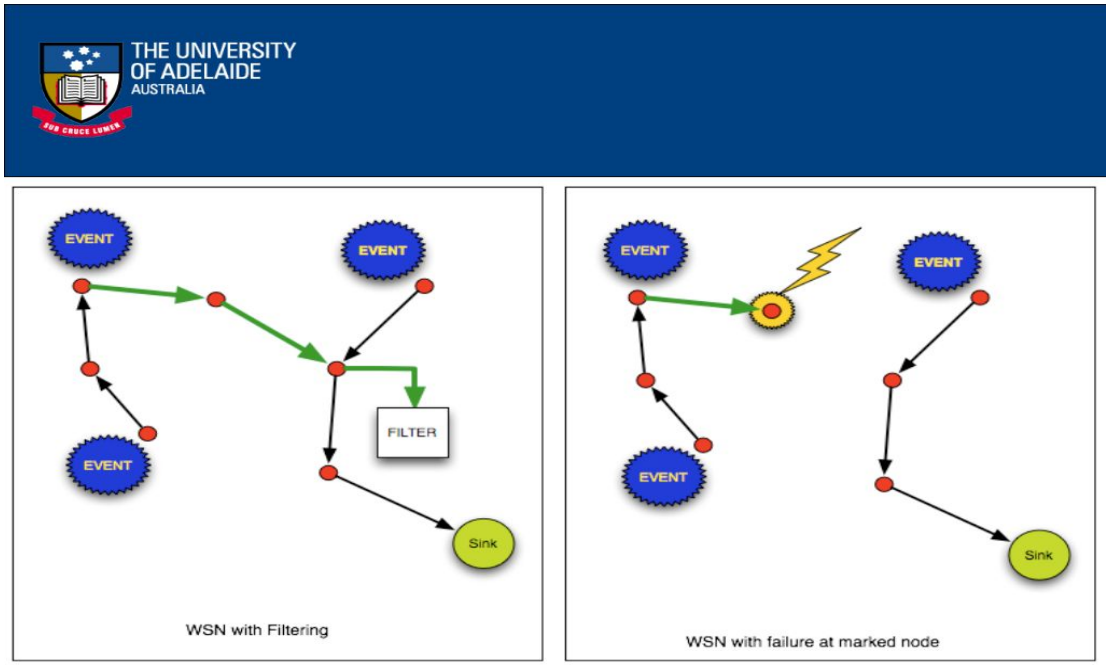
Time Comparison: T10



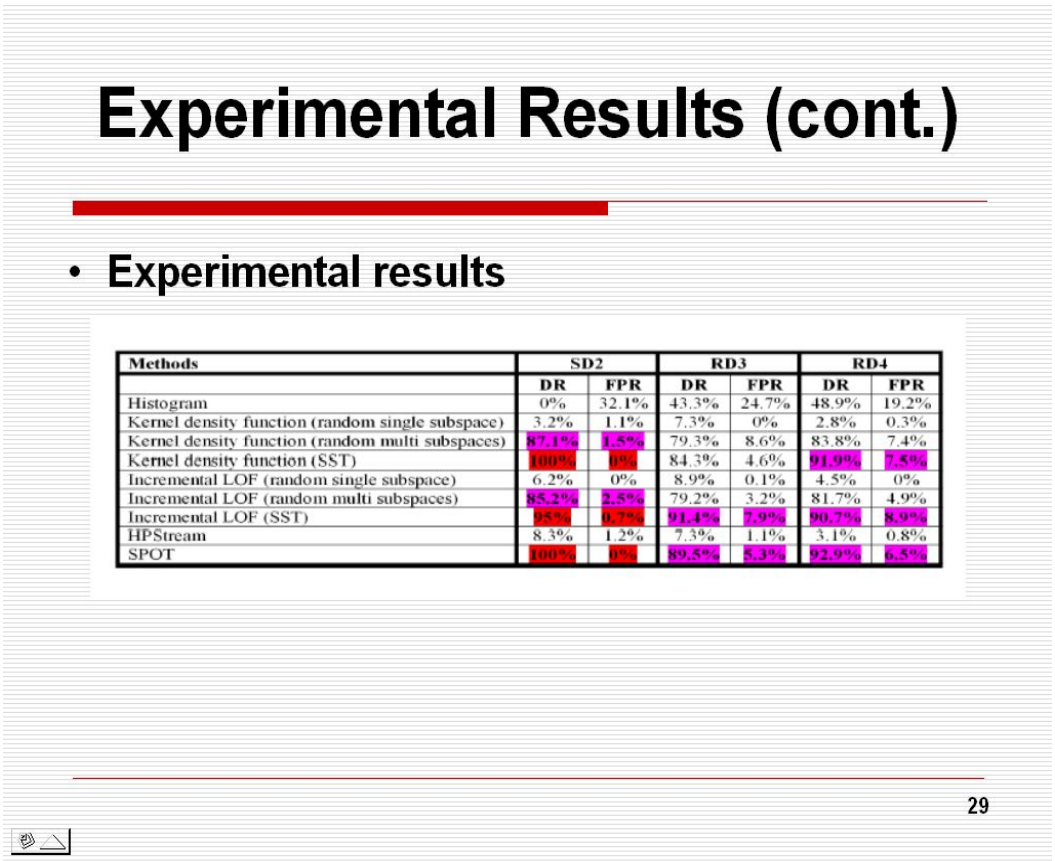
- TidFP : Mining Frequent Patterns in Different Databases with Transaction ID (발표자 : Christie I. Ezeife School of CS, Uinv. of Windsor Windsor, Canada) : 거래 아이디를 가지고 다른 데이터베이스에서 빈도 패턴을 마이닝하는 방법에 대한 설명으로서 이 방법을 사용했을 경우 Apriori 알고리즘보자 25더 효율적으로 수행되는 방법임을 발표하였으며 향후 보다 복잡한 정보 발견에 대한 마이닝을 위하여 다른 타입의 방법으로까지 확대됨을 미래의 과제로 제시하였음



- Significance-Based Failure and Interference Detection in Data Streams (발표자 : Nick Falkner and Quan Z. (Michael) Sheng, School of Computer Science the University of Adelaide) : 무선인터넷 네트워크는 매우 유용하지만 불규칙한 데이터 흐름을 발생시키며 중요한 데이터가 다른 데이터들에 비하여 불규칙하게 도착하기 때문에 데이터를 분석하기 어렵다. 이를 해결하기 위하여 SEDIM을 이용한 연구결과를 발표하였음

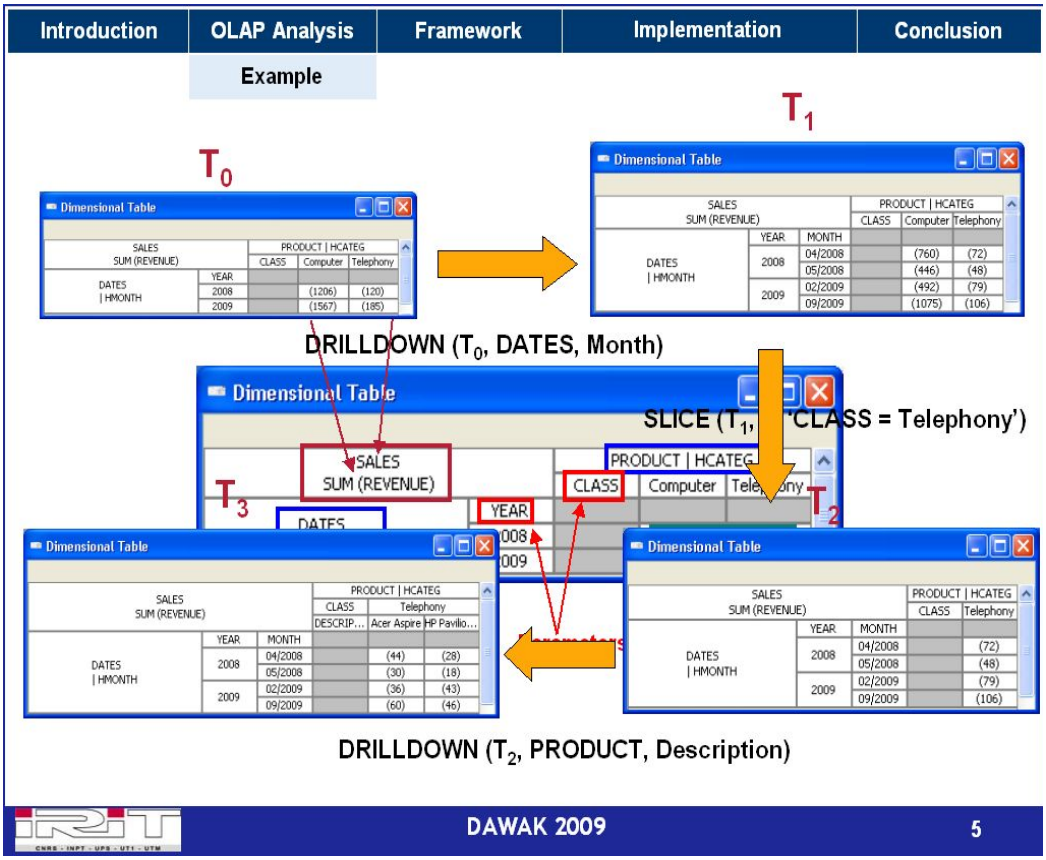


- Detecting Projected Outliers in High-dimensional Data Steams (발표자 : Ji Zhang, Qigang Gao, Hai Wang, Qing Liu, Kai Xu) : 거대한 데이터 흐름에 국외탐색기법은 잠재적으로 비정상적이고 불규칙하지만 유용한 패턴을 발견할 수 있게 하며 컴퓨터 및 네트워크 안전은 혼란을 일으키는 비정상적인 데트워크 흐름을 찾게 하며 신용카드 거짓 탐지기는 비정산적인 거래시잔, 거래장소 거래량에 대한 신용카드 거래를 찾을 수 있게 SPOT 프로그램에 대한 연구 발표였음



Methods	SD2		RD3		RD4	
	DR	FPR	DR	FPR	DR	FPR
Histogram	0%	32.1%	43.3%	24.7%	48.9%	19.2%
Kernel density function (random single subspace)	3.2%	1.1%	7.3%	0%	2.8%	0.3%
Kernel density function (random multi subspaces)	87.1%	1.5%	79.3%	8.6%	83.8%	7.4%
Kernel density function (SST)	100%	0%	84.3%	4.6%	91.9%	7.5%
Incremental LOF (random single subspace)	6.2%	0%	8.9%	0.1%	4.5%	0%
Incremental LOF (random multi subspaces)	85.2%	2.5%	79.2%	3.2%	81.7%	4.9%
Incremental LOF (SST)	95%	0.7%	91.4%	7.9%	90.7%	8.9%
HPStream	8.3%	1.2%	7.3%	1.1%	3.1%	0.8%
SPOT	100%	0%	89.5%	5.3%	92.9%	6.5%

- Preference-based Recommendations for OLAP Analysis (발표자 : Houssem Jerbi, Franck Ravat, Olivier Teste, Gilles Zurfluh) : 전형적인 OLAP 이용자는 대화식 탐식 방식을 사용하였으며 이용자 분석은 매우 지루하고 복잡한 것이었다. 현재의 OLAP 툴은 분석과정에서 이용자들에게 최소한의 가이드를 제공하고 있다. 이용자들이 데이터 탐색을 하는 이용자들에게 도움을 주소 이용자들이 이을 더 쉽도록 하는 방안을 소개하였음



- E-Government (발표자 : Annika Nietzio, Morten Goodwin Olsen, A Mikael Snaprud, Rudolph Brynn) : 정보화 국가가 장애를 가진 사람들에게 새로운 기회인가 아니면 새로운 장벽인가에 대한 연구발표로서 장애를 가진 사람들의 정보 접근을 용이하게 하기 위한 새로운 서비스가 마련되어야 하며 정보접근권과 접근 가능한 콘텐츠를 생산할 수 있는 능력에 대한 정부당국의 각성이 이루어져야 한다는 등의 내용이 발표 되었음

- E-Gov and Acces to Jusice in Brazil (발표자 : Andre Andrade, Luiz Antonio Joia) : 브라질은 E-Government 현황 및 정보접근성의 공정성에 대한 연구 발표로서 지난 30년동안의 브라질 국가의 정보 정의 및 2007년 브라질은 59%가 인터넷을 사용한 경험이 없었으나 2008년에는 50%이상이 인터넷을 사용한 경험이 있는 등 급속하게 향상되고 있는 정보 접근성 및 인터넷 성장, e-정부를 통한 정보접근성의 정의를 확대할 기회 등에 대한 소개였음.

- Finding Clothing that Fit through Cluster Analysis and Objective Interestingness Measures (발표자 : Isis Pena, Herna Viktor, Eric Paquet) : 여자 50% 남자 62%가 그들에게 적합한 옷을 찾을 수 없었다는 연구발표는 최신까지의 인체측정학에 대한 데이터 부족과 언제 의류를 설계해야 하는 지를 고려하기 위하여 필요한 데이터가 부족했기 때문이었음을 증명하는 발표로서 과거부터 인류측정학 측면에서 인류가 생체적인 변화와 의복 디자인 설계의 치수를 상호 비교함으로써 여자 50% 남자 62%가 자신에게 맞는 의복을 찾지 못하는 것에 대한 설명을 하였음

< Actual versus Ideal Measurements for Sizs Large Adult Males >

Chest	Waist	Hip	Neck	Stature
112.4	92.8	109.7	52.1	184.4
112.6	97.8	105	42.4	179.5
112.6	97.2	111.4	52	181.3
112.8	95.2	114	49.9	185.4
112.9	93.5	107.8	50.3	184.7
113	105.7	114.1	49.3	184.3
113.3	101.8	111.2	50.5	188.6
113.4	102.6	116.5	47	187.2
113.5	100.8	105.2	48.5	179.6
112	97	114	42	178

- A versatile Record Linkage Method by term Matching Model Using CRF (발표자 : Quang Minh Vu, Atsuhiro Takasu, Jun Adachi) : 데이터베이스 기록 및 효율적 기록물 관리를 위하여 CRF 방법을 이용하여 중복된 부분을 찾고 중복된 부분에서 유사한 특징 벡타를 계산하며 SVM 분리기를 이용하여 중복부분을 발견하는 것에 대한 연구 발표를 하였음

- Modeling Complex Relationships (발표자 : Mengchi Liu, Jie Hu Shool of Computer, Wuhan University, China) : 객체 지향과 같은 현재의 데이터 모델은 복잡하고 다양한 실세계의 상호 관계를 자연스럽거나 직접

적으로 표현하지 못하는 한계가 있으며 O-O 모델은 실세계의 정적인 측면에 대하여 주로 관심을 가지며 가장 특별한 측면에 대한 즉각적인 것에 목적을 주고 있다. 어떤 O-O모델은 몇 개의 특징을 보여주기도 하지만 복잡한 현실의 관계를 지원해 주지 못한다. Modeling Complex Relationships 에서는 역할관계 및 소설메카니즘을 이용하여 이들의 한계를 극복하고 현실을 잘 표현할 수 있는 모델에 대한 연구 결과를 발표하였음

- An Optimization Technique for Multiple Continuous Multiple Joins over Data (발표자 : Changwoo Byun, Hunjoo Lee, Youngha Ryu, Seog Park) : 다양한 센서로부터 사전에 계획된 시간간격으로 데이터를 수집하여 중앙프로세서에 수집된 데이터를 보내고 하나의 센서로부터 제한된 종류의 정보를 얻는 방법에 대한 연구를 발표하였음

5. 시사점

○ 데이터 통합 설계 시 급변하는 정보 환경 변화 고려

- 데이터 통합의 한 방법인 데이터웨어하우스 구축 및 유지시 인터넷 보급과 확대에 따른 다양한 데이터가 생성되거나 명확하지 않고 불분명한 데이터가 증가하는 등의 급변하는 정보 환경에 대응하여 설계되고 유지 보수 되어야 함

○ 통계전산프로그램 설계시 데이터 통합을 고려한 프로그램 설계 및 통합 조정 시스템 또는 관리자 필요

- 각 개별 통계 조사 별로 통계전산프로그램을 설계할 경우에도 데이터 통합에 필요한 정보를 함께 구축하는 등 데이터 통합을 고려한 프로그램이 설계되어야 하며 각 개별 통계전산프로그램을 통합하여 조정할 수 있는 시스템 또는 관리자가 요구됨