DW 컨퍼런스 참가 결과보고

1. 출장 개요

가. 출장내용 : DW 국제 컨퍼런스 참석 및 자료수집

나. 행사주최 : DaWak2005

※DaWak2005 : 제7회 데이터 웨어하우징 및 지식발견 컨퍼런스

다. 장　　소 : 덴마크 코펜하겐

라. 기　　간 : 2005. 8. 20 ~ 8. 28(9일간)

※ 컨퍼런스 기간 : 2005. 8. 22 ~ 8. 26

마. 출장자(3명) : 이명호(전산개발과), 이영수(정보서비스과)
김기만(정보서비스과)

바. 회의 내용

- Data Warehouse 개념적・논리적・물리적 구조
- Data Warehouse 품질, 보안과 신뢰성 확보 방안
- Data Warehouse 구축・지원을 위한 Data Mining 기법
- 메타데이타 레파지토리 디자인과 유지보수
- Data Warehousing 통합과 다차원분석(OLAP)의 방법론 등

2. 세부 일정

월/일	시간	주제	내용
8/22 (1일차)	10:30 ~ 12:30	Data Warehouse 1	『트리 비교를 통한 DW 구조 변경 탐지』 외 3편
	14:00 ~ 15:30	Data Warehouse 2	『UML 2 활동 다이어그램을 BI 객체로의 확장』 외 2편
	16:00 ~ 18:00	Evaluating Data Warehouses and tools	『Business Intelligence 공개 툴 개관』 외 3편
8/23 (2일차)	10:30 ~ 12:30	Schema Transformations	『점진적 뷰 유지를 위한 스키마 변환 경로 사용』 외 3편
	14:00 ~ 15:30	Materialized Views	『DW에서 참조 통합 제약을 사용한 materialized 뷰의 병행 일관성 유지』 외 2편
	16:00 ~ 17:00	Aggregates	『집합체 리스트의 효율적인 저장·가공 방법』 외 2편
8/24 (3일차)	10:30 ~ 12:30	Data Warehouse Queries and Database processing Issues	『데이터 큐브의 질의응답 유연성』 외 2편
	14:00 ~ 15:30	Data Mining Algorithms and Techniques	『기호값 특성치를 위한 사례기반추론 및 신경망 혼합 시스템』 외 2편
	16:00 ~ 18:00	Data Mining	『유형화한 데이터 마이닝의 온라인 동시 갱신을 통한 증분적 데이터 마이닝 뷰』 외 3편
8/25 (4일차)	10:30 ~ 12:30	Association Rules	『인터벌 기반의 풍부한 시간적 연관규칙의 발견』 외3편
	14:00 ~ 15:30	Text Processing and Classification	『데이터 저장소 디멘션의 텍스트 속성을 효과적으로 압축하기』 외 2편
	16:00 ~ 17:00	Miscellaneous Applications	『분석에 의한 침입행위 탐지 및 사용자 명령어의 모형화』 외 1편
8/26 (5일차)	09:00 ~ 10:00	Security and Privacy Issues	『FMC: 사생활보호를 위한 온라인 분석법(OLAP)』 외 1편
	10:30 ~ 12:30	Patterns	『빈발성 패턴 검색 시퀀스를 최대 활용하는 법』 외 3편
	14:00 ~ 15:30	Cluster and Classification 1	『사생활 보호와 데이터를 공유를 위한 분류 규칙 은닉법』 외 2편
	16:00 ~ 17:30	Cluster and Classification 2	『Spectral Kernels 분류법』 외 3편

3. 주요 컨퍼런스 내용

Session 1 : Data Warehouse 1

A Tree Comparison Approach to Detect Changes in Data Warehouse Structure.**(트리 비교를 통한 DW 구조 변경 탐지)**

DW 구조의 버전 변경을 발견하고 표현하는데 대한 기술을 언급하고 있다. 다차원 자료구조의 특성을 표시하고, 노드 리네임의 탐지를 위한 모듈에 확장하는데 트리 변경 알고리즘을 선택하고 있다. 이 알고리즘의 결과는 이전 버전에서 나중 버전으로 일련적으로 변환되는 과정으로 구성되는 에디트스크립트(editscripts)라고 불린다. 이 절차는 DW 관리자가 변화를 등록하는데 도움을 준다. 이 논문은 Hyperion Essbass DW 다차원구조를 도입하고 이러한 버전들을 비교, 차이 리스트를 생성하는 개념을 프로토타입 형태로 수행한 것을 기술하고 있다.

Session 2 : Data Warehouse 2

Extending UML 2 Activity Diagrams with Business Intelligence* Objects. (UML 2 활동 다이어그램을 BI 객체로의 확장)

DW의 정보들은 비즈니스 프로세스에 의해 액세스된다. 오늘날 DW와 비즈니스 프로세스간의 관계를 명백하게 설정하는 개념 모델은 존재하지 않는다. 이 논문에서는 비즈니스 프로세스 모델링 다이어그램, 즉 UML 2 활동 다이어그램을 이 관계를 명백히 설정하도록 하는 UML 프로파일로 확장하고 있다. 이 모델은 샘플 비즈니스 프로세스로 테스트되고 있다.

* BI(Business Intelligence) : 기업 경영의 전반적인 의사결정과정에 필연적으로 수반되는 각종 지능화된 접근방법(방법론) 및 솔루션 그리고 이러한 것을 기반으로 하는 비즈니스와 비즈니스 솔루션

Session 3 : Evaluating Data Warehouses and Tools

A survey of Open Source Tools for Business Intelligence (Business Intelligence 공개 툴 개관)

BI 공개툴을 현업에서 적용하는 사례는 아직 일반적이지 못하다. 그래서 공개툴이 BI에 유용한지 검토하는 것과 툴들을 비교해 보는 것에 관심이 있다. 이 논문에서 많은 BI 공개툴에 대한 능력을 고려하고 있다. 이 논문에서는 3가지 ETL툴, 3개 OLAP서버 툴, 2개의 OLAP 클라이언트 툴, 4개의 DBMS 툴에 대해 다루고 있다. ETL툴은 DBMS만큼 현업에서 사용되고 있지는 못하다. OLAP 서버와 클라이언트 툴은 상업 솔루션만큼 강력하지는 못하지만 덜 부담이 되는 프로젝트에 유용할 것으로 보인다.

Session 4 : Schema Transformations

Using Schema Transformation Pathways for Incremental View Maintenance(점진적 뷰 유지를 위한 스키마 변환 경로 사용)

인터넷상에 이용할 수 있는 정보의 양과 다양성이 증가됨에 따라, 분산되고 이질적인 데이터 소스로부터 데이터를 통합할 필요가 있는 정보 시스템의 성장이 있어왔다. 통합 데이터를 점진적으로 유지하는 것은 데이터웨어하우징 연구에서 주요 문제 중에 하나이다. 이 논문은 스키마 변환 경로에 기반한 점진적 뷰 유지 접근을 제안하고 있다. 우리의 접근은 특별한 데이터 모델이나 질의어에 제한되지 않는다. 그리고 근본적인 스키마 변환의 결과에 기반한 어떤 데이터 변환이나 통합 프레임워크에 유용할 것이다.

Session 5 : Materialized Views

Parallel Consistency Maintenance of Materialized Views Using Referential Integrity Constraints in Data Warehouses(DW에서 참조 통합 제약을 사용한 materialized 뷰의 병행 일관성 유지)

DW는 분산된 데이터 소스로부터 추출된 온라인 분석 정보를 유지하는 materialized 뷰로 고려될 수 있다. 데이터 소스들이 변경될 때, materialized 뷰는 데이터 소스와 일관성을 유지하기 위하여 대응해서 유지되어야 한다. 만약 뷰가 몇 개의 소스 관계를 조인해서 정의된다면 한 소스에서의 업데이트는 일련의 조인 서브쿼리를 야기한다. 이렇게 뷰 유지는 많은 시간이 소요된다. 이 논문에서는 소스 관계상에서 참조 통합 제약을 사용함으로써 병행적으로 조인 서브질의어를 프로세싱 하는 뷰 유지 알고리즘을 제안한다. 몇 개 외래키를 가지고 있는 관계는 독립적으로 참조 관계와 조인될 수 있다. 제안된 알고리즘은 이러한 조인 작동을 병행적으로 수행하고, 그들의 결과를 합성한다. 병행 프로세싱을 함으로써 알고리즘은 효율적으로 materialized 뷰를 유지할 수 있다. 분석 비용 모델을 사용함으로써 제안된 알고리즘의 우수성을 입증할 수 있다.

Session 6 : Aggregates

On Efficient Storing and Processing of Long Aggregate Lists (집합체 리스트의 효율적인 저장하고・가공하는 방법)

이 논문에서는 아주 긴 전체 리스트를 효율적으로 저장하고 가공하기 위하여 고안된 해결책, 즉 Materialized Aggregate List(유형화한 집합체 리스트)를 제시하고 있다. 하나의 전체 리스트에는 데이터베이스에 저장된 데이터로부터 계산된 여러 가지 집합체를 포함하고 있다. 여기서의 접근방법은 한번 생성된 전체 리스트를 장래에 활용하기 위해 유형화한다는 것이다. 리스트의 구조는 각 페이지로 나누어지는 테이블을 포함한다. 우리는 여기서 세 가지 다른 페이지 구성 알고리즘을 제시하는데, 이것은 리스트가 검색될 때 사용된다. 또한 우리는 테스트 결과를 보여주고, 구성 매개변수들의 최적 조합 등을 평가하는데 사용한다. 그 세 가지는 페이지 번호, 한 페이지의 크기, 유효 데이터베이스 연결을 말한다. Materialized Aggregate List(유형화한 집합체 리스트)는 aR-tree와

같은 다양한 indexing 구조의 각 집합체 레벨에 적용할 수 있다.

Session 7 : Data Warehouse Queries and Database processing Issues

Flexible Query Answering in Data Cubes
(데이터 큐브의 질의응답 유연성)

이 논문에서는 데이터웨어하우스에서 개략적인 질의응답을 도출하는 새로운 접근방법을 제시하고 있다. 이 접근방법은 다차원 데이터에 대한 러프 집합이론(rough set theory)의 적용에 기초한다. 데이터웨어하우스에 있는 데이터는 데이터의 신뢰 정도가 서로 다르고, 데이터 포맷이 다양한 다원적이고, 이질적인 데이터로부터 나오기 때문에 사용자들은 데이터웨어하우스의 환경에 관대하여야 하고, 실제 데이터와 조작된 데이터 사이의 유용한 정보 누출 및 불일치를 받아들이는 경향이 있다. 그러므로 여기에서의 작업 목적은 개략화 메카니즘을 통합하고, 조작인자를 데이터 큐브에 연관시킴으로써 OLAP 혹은 data mining 기법을 이용하여 탐사된 뷰(views)를 생성하는 것이다. 즉, OLAP 기법을 이용하여 데이터 개략화의 수용성을 통합하여 데이터 cube 탐사 및 분석을 위한 추가적인 수단을 제공한다.

Session 8 : Data Mining Algorithms and Techniques

Hybrid System of Case-Based Reasoning and Neural network for Symbolic Feature
(기호값 특성치를 위한 사례기반추론 및 신경망 혼합 시스템)

사례기반추론(Case-based reasoning)방법은 데이터마이닝에서 가장 자주 사용되는 도구이다. 이 기법은 여러 가지 문제를 해결하는데 유용한 것으로 밝혀져 있으나, 특성을 가중화하는 문제 등에 있어

서는 단점이 있는 것으로 알려져 있다. 이전 연구에서 사례기반추론과 신경망을 혼합한 혼성시스템(Hybrid system)을 제안하였다. 이 시스템은 특성가중치가 연습된 신경망으로부터 추출되고, 사례기반추론방법의 정보검색 정확성을 제고하는데 사용되었다. 그러나 이 시스템은 모든 특성들이 수치 값을 갖고 있는 영역에서 최상으로 작용된다. 특성이 기호 값인 경우에는 전형적으로 결부되는 특성치의 개수와 같은 아주 간단한 행렬에 의한 최근 nearest neighbor method에 의지하고 있다. 따라서 기호 값의 영역에서도 더욱 정교한 특성치 공간의 취급방법이 요구된다. 우리는 여기서 기호 값 특성치에 대하여 Value Difference Metric(VDM)을 사용하여, 사례기반추론과 신경망을 혼합한 또 다른 혼성시스템(Hybrid system)을 제안한다. 이 제안된 시스템은 기호 값 영역의 데이터집합에서 유효하다.

Session 9 : Data Mining

Incremental Data Mining Using Concurrent Online Refresh of Materialized Data Mining Views
(유형화한 데이터 마이닝의 온라인 동시 갱신을 통한 증분적 데이터 마이닝 뷰)

데이터마이닝은 쌍방향의 상호 작용과정이다. 사용자들은 일련의 유사한 데이터마이닝 질의문을 작성하고 각 경우에 따라 조작된 데이터 집합의 정의를 수정하거나, 데이터마이닝 알고리즘의 매개변수를 수정하여 사용한다. 이러한 모델은 이전 질의의 결과를 재사용하는 증분적 마이닝 알고리즘(Incremental mining algorithms)에 가장 적합하다. 증분적 마이닝 알고리즘은 유용한 이전 질의의 결과를 요구한다. 그러한 결과를 보전하는 한 가지 방법은 유형화한 데이터마이닝 뷰(Materialized data mining views)를 사용하는 것이다. 유형화한 데이터마이닝 뷰는 조작된 패턴을 저장하고, 근본적인 데이터가 변화함에 따라 그것들을 새롭게 갱신한다. 데이터마이닝과 지식발견은 가끔 데이터웨어하우스 환경에서 발생하기

도 한다. 데이터웨어하우스에 대하여 정의된 비교적 작은 데이터 마이닝 뷰가 있을 수 있다. 갱신 과정이 원본 데이터베이스에서 재발견 형태를 취한다면, 각 데이터마이닝 뷰를 개별적으로 갱신하는 것은 비용이 과다할 수 있다. 이 논문에서는 유형화한 데이터마이닝 뷰의 갱신과정에 관한 새로운 접근을 제안한다. 즉, 우리는 유형화된 데이터마이닝 뷰의 유지와 관련하여 데이터웨어하우스 갱신과정 통합을 위한 틀을 제안한다. 마지막으로 가공의 데이터 셋에 대하여 여러 가지 실험을 수행함으로써 접근방법의 실현 가능성을 증명한다.

Session 10 : Association Rules

Discovering Richer Temporal Association Rules from Interval-Based Data
(인터벌 기반의 풍부한 시간적 연관규칙의 발견)

시간적 연관 규칙(Temporal Association Rules) 데이터 마이닝은 대규모의 데이터에서 사건의 시간 의존적인 상관관계 혹은 패턴을 발견할 수 있도록 한다. 지금까지 시간적인 데이터 마이닝에 관한 연구는 시간 간격 보다는 어느 시점에 존재하는 사건에 초점이 맞추어져 있다. 정적인 규칙에 비교하여, 시점을 고려하는 데이터 마이닝은 의미론적으로 더욱 풍부한 규칙들을 제공한다. 그러나 지금까지는 시간적인 인터벌을 수용하는 것도 풍부한 규칙을 제공하고 있다. 이 논문에서는 자주 발생하는 시간적 패턴을 발견하고, 인터벌 기반의 시점 관련 규칙들을 생성하는 새로운 알고리즘의 윤곽을 제시한다.

Session 11 : Text Processing and Classification

Efficient Compression of Text Attributes of Data Warehouse Dimensions
(데이터 저장소 디멘션의 텍스트 속성의 효과적으로 압축방법)

본 논문은 기존의 텍스트 압축법을 이용하여 관계형 DB 관리시스템에 자료를 압축하는 방법에 관한 것이다. 이 방법은 널리 쓰이고 있으나 중형과 대형 디멘션 통계표를 데이터 저장소에 압축하는 경우 특히 편리하다. 사실 디멘션은 텍스트 속성을 많이 갖고 있어서 그 크기를 축소하면 검색 실행시간에 큰 영향을 준다. 이것은 디멘션과 결과표를 연결해주기 때문이다. 일반적으로 자료검색이 복잡하고 시간이 오래 걸리는 문제는 디멘션 텍스트 속성을 (또한 허위 사실 등 결과표에 들어있을 수 있는 텍스트 속성을) 압축함으로써 검색 응답시간을 단축할 수 있다. 본 논문이 제안하는 방법을 잘 알려진 TPC-H를 이용 평가한 결과, 대부분의 검색에 대해서 40%이상 속도가 향상된 것으로 드러났다.

Session 12 : Miscellaneous Applications

Intrusion Detection via Analysis and Modelling of User Commands(분석에 의한 침입행위 탐지 및 사용자 명령어의 모형화)

컴퓨터가 일상생활에서 큰 비중을 차지하게 된 이후로 시스템과 데이터를 원하지 않는 사람에게서 보호하는 일 뿐만 아니라 침입행위를 찾는 방법에 대한 관심도 최근 높아졌다. 본 논문은 사용자의 명령어 데이터를 분석하여 컴퓨터 시스템을 부적절하게 사용한 것을 찾아내는 것에 초점을 두었다. 이미 사용된 명령어의 구조를 살펴본 후 새 명령어를 테스트하는 모형도 생성하였다. 본 논문이 제안하는 방법을 사용해 본 후 분석도 하였는데 명령어에 초점을 두되 본 논문이 제안한 테크닉을 통해 시스템 콜과 같은 데이터 분석도 가능하다.

Session 13 : Security and Privacy Issues

FMC: An Approach for Privacy Preserving OLAP
(FMC: 사생활보호를 위한 온라인 분석법 (OLAP))

철저한 분석을 제공하면서 개인정보를 보호하는 문제는 OLAP의 중요한 문제 중 하나다. 한 가지 도전과제로는 민감하지 않은 데이터 집합으로부터 민감한 값을 도출하지 않도록 하는 문제가 있다. 본 논문은 민감한 정보를 제외한 추가 데이터를 은폐함으로써 추론 문제를 해결하기 위한 새로운 FMC를 살펴보고자한다. 또한 이 추가 정보는 필요충분조건을 갖추고 있음을 입증하고자 한다. 따라서 본 논문의 접근법은 사용자에게는 가능한 더 많은 정보를 제공 할 수 있을 뿐만 아니라 보안도 유지할 수 있다. 이러한 전략이 OLAP 시스템의 사용에는 영향을 주지 않는다. 본 논문은 FMC의 효율성과 타당성을 입증하기 위한 체계적 분석 및 실험에 의한 비교도 제공한다.

Session 14 : Patterns

Optimizing a Sequence of Frequent Pattern Queries(빈발성 패턴 검색 시퀀스를 최대 활용하는 법)

빈발성 패턴을 찾아내는 일은 여러 가지로 응용될 수 있는 중요한 데이터 검색의 문제이다. 빈발성 패턴 검색은 고급 검색으로 간주된다. 이 경우 사용자는 정해진 제약 모형을 이용하여 소스 데이터 셋과 패턴에 대한 제약을 명시한다. 빈발성 패턴 검색을 능률적으로 하기위한 많은 연구가 최근 행해졌으며 대개 제약 처리 문제와 기존 검색의 결과를 재사용하는 문제에 초점이 있다. 본 논문은 일괄처리시스템으로 빈발성 패턴 검색 시퀀스 최대 활용법을 다루고자한다. 우리의 해법은 기존 검색 시퀀스를 알아내면 시스템이 검색 일정을 짜거나 조정하여 먼저 실행된 검색의 결과를 이용할 수도 있다는 사실을 이용하고 있다. 우선 간단히 검색 일정을 짜보는 것을 시작으로 일괄 검색을 달리 변형하는 방법도 살펴보았다.

Session 15 : Cluster and Classification 1

Hiding Classification Rules for Data Sharing with Privacy Preservation (사생활 보호와 데이터를 공유를 위한 분류 규칙 은닉법)

본 논문은 분류 데이터 검색 연산으로부터 민감한 분류 규칙을 숨기는 법을 제안하고자 한다. 우리의 접근법은 데이터 제공자가 점검 후 동의한 분류 규칙을 따르되 데이터 공유를 위한 발표 때 데이터를 재구성하는 방법이다. 여타의 발견적 수정 모형과는 달리 우리의 방법은 우선 정해진 데이터 셋을 분류한다. 그 후 감춰야 할 민감한 분류 규칙을 정하기 위해 데이터 제공자에게 분류 규칙을 보여준다. 그 다음 민감하지 않은 분류 규칙으로만 구성된 새로운 모양을 작성한다. 마지막으로 새 데이터 셋으로 재구성한다. 민감한 분류 규칙이 재구성된 데이터 셋에서 완벽하게 숨겨져 있는 실험이 제시되어 있다. 한편 민감하지 않은 규칙도 아무 문제없이 볼 수 있다. 또한 우리가 제안하는 방법은 재구성된 데이터 셋의 활용도를 높은 상태로 유지할 수 있다.

Session 16 : Cluster and Classification 2

Spectral Kernels for Classification (Spectral Kernels 분류법)

Spectral methods은 관리자가 필요 없는 테크닉이며 정보 검색 시 LSI나 웹 검색엔진의 HITS 및 PageRank 등의 자료 검색에서 성공적으로 사용되었다. 또한 기계 학습의 분광형 클러스터링에도 사용되었다. 이 같은 성공의 본질은 관리자가 없는 학습에서 필요한 많은 양의 데이터 고유의 의미를 포착해주는 분광형 정보 덕분이다. 본 논문은 관리자가 있는 학습 (예를 들면 분류)에서 **Spectral methods**이 사용될 수 있는지를 알아내고자한다. 이 문제의 답을 찾기 위한 본 연구는 새로운 핵을 발견하였는데 여기서는 분광형 덩어리 정보를 원활하게 이용하여 분류 작업 중 새로 들어오는 데

이터에 쉽게 확대 이용할 수 있었다. 실험 결과 우리가 제안하는 **Spectral Kernel** 분류법은 정확도를 유지한 채 분류 작업 속도를 단축할 수 있음이 입증되었다.