제13차 Data Management
국제 컨퍼런스 국외출장 결과보고

2007. 1.

통계정보국
전산개발과

목 차

13차 Data Management
국제 컨퍼런스 국외출장 결과보고

Ⅰ. 출장개요

1. 출장목적

□ 데이터웨어하우스 구축방법, 데이터모델링, 데이터베이스 관리 등에 관한 최신기술 습득 및 동향 파악

2. 출장기간 : 2006. 12. 12. ~ 12. 19. (8일간)

3. 출장장소 : 인도 델리

4. 출장자

□ 전산개발과 5급 한동철, 6급 황의태, 7급 전동현

5. 주요 회의내용

□ 데이터웨어하우스 구축방법

□ 데이텁이스 모니터링 및 튜닝기법

□ 메타데이터 정보관리

□ 병렬 및 분산데이터베이스 관리 등

Ⅱ. 컨퍼런스 소개

1. 명칭 : The COMAD International Conference

2. 컨퍼런스 내용 : Data Management 관련 내용

□ 인도 외 외국논문 참가율 : 30% (유럽, 미국, 아시아 등)

3. 연혁

□ 제 1차 COMAD 개최 (1989년) ~ 제 13차 COMAD (2006년)

4. 13차 COMAD 조직구성

구 분	성명 및 소속
General Chair	- S. K. Gupta (IIT Delhi)
Organizational Committee Chair	- R. K. Dutta (CSI, Delhi)
Program Committee Co-Chairs	- Laks V.S. Lakshmanan (University of British Columbia) - Anthony K. H. Tung (Natl. University of Singapore)
Application/Industrial Program Chair	- Rajeev Rastogi (Bell Lab, India) - Krishna Reddy (IIIT, Hyderabad)
Tutorials Chair	- P Sreenivasa Kumar (IIT Madras)
Panels Chair	- Soumen Chakrabarti (IIT Bombay)
Best Paper Award Chair	- Jayant Haritsa (IISc, Bangalore)
Publicity Chair	- Rajeev Gupta (IBM-IRL, India)
Local Arrangements Chair	- Naveen Kumar (University of Delhi)
Proceedings Editor	- Prasan Roy (IBM-IRL, India)
Demonstrations Chair	- Manish Bhide (IBM-IRL, India)

Ⅲ. 13차 컨퍼런스 내용(요약)

1. 핵심 논의사항

가. Peer-to-Peer(P2P) 시스템에서의 정보검색

Gerhard Weikum(독일, 막스-플랑크 정보학연구소)

P2P 연산 패러다임은 인터넷 공동체 (예, Gnutella, BitTorrent) 혹은 IP 전화(telephony)(예, Skype)에서 파일 공유등과 같은 전 세계적 애플리케이션의 확산 속에서 매우 성공적이었다. P2P 시스템은 완전히 분산형이어야 하며, 수행능력이나 이용가능성 혹은 공격의 취약성 측면에서 병목(bottleneck)상태가 될 수도 있는 중앙 집중형 구성성분 없이 작동될 수 있어야한다. P2P 시스템은 단지 몇 개의 노드(node)로부터 수백만 컴퓨터로 성장할 수 있도록 함으로써 어떤 제약도 없이 척도화 할 수 있어야한다. P2P 시스템은 기저를 형성하는 컴퓨터의 자율성을 강조한다. 그리고 사전예고 없이 노드가 네트워크에 결합하고 떠날 수 있도록 허용함으로써 빈번하게 발생하는 노드 실패, 급격히 변화하는 자료 및 로드(load) 특징 측면에서의 높은 역동성, 그리고 높은 동요(churn) 등에 대해 관대해야한다. P2P 시스템은 심지어 이기적이고, 남을 속이거나 악의가 있는 동료를 연결해줄지도 모르는 행실이 바르지 못한 동료에 대해서도 굳건해야한다. 이 같은 두드러진 특징 중 어느 것도 전 세계적 기획, 관리 혹은 통제를 요구해서는 안 된다. 그러므로 P2P 시스템은 온전히 자기 스스로 구성되어져야한다. 게다가 P2P 시스템은 척도화 가능성 및 자기 스스로의 구성력을 가능토록 하기위해서는 거대한 단일결정칩(monolithic) 시스템의 소프트웨어보다 훨씬 더 간편한 소프트웨어 구조를 갖고 있어야한다.

P2P 파일 공유 애플리케이션의 인상적인 성공에도 불구하고, 약술한 P2P 유토피아가 좀더 세련된 기능성을 갖춘 보다 진보한 형태의 포괄적 자료 관리 및 정보 검색을 위해서도 존속될 수 있는가에 대한 질문은 여전히 타당하게 되며, 인터넷상에서 간단하고 척도화 가능하며 또한 자기 스스로 구성되어지는 데이터 관리를 위한 P2P의 꿈 실현을 위한 설계 원리와 기본 원칙을 알아내고자한다. 이러한 논의의 동력이 되는 고급 애플리케이션으로는 P2P 방식으로 실행되는 구글 스타일 웹 검색과 시간 여행 검색 지원을 받는 분산형 웹 파일 보관, 그리고 학술 정보(예, 간행물, 프로젝트, 회의 및 보고서등) 및 그와 유사한 사회 공동체에 대한 P2P 공표-정기구독 기능성 등이 있다.

나. 신뢰할 수 없는 서버에 대한 안전한 자료관리

Dennis Shasha(미국, 뉴욕대, Courant 수학연구소)

사용자가 평행수행제어(concurrency control), 회복, 검색처리 등을 원하기 때문에 DB내에서 정보를 사용자와 사용자 친구들이 공유하고 싶어 하지만 DB 관리자를 믿지 못한다고 가정해 보자. 당신은 자료가 관찰되지 않도록 보호하기를 원한다.(사생활보호) 당신은 권한 밖의 수정을 명백히 하기를 원한다.(일종의 안전) 당신은 서버로 하여금 모든 정확한 사용자에게 일관성 있는 그림을 전달하게 강요하기를 원하거나 발견되기를 원한다.(일종의 활기) 암호화 및 기호화는 처음 둘을 가능케 해준다. 활기는 또 다른 문제이다. 왜냐하면 DB 관리자는 DB를 "찍어서"(fork) 몇 개로 복사할 수 있으며, 당신 친구 중 몇 명이 당신의 최신 업데이트를 알지 못하게 할 수도 있으며, 당신이 당신 친구들의 최근 업데이트를 알지 못하게 할 수도 있기 때문이다. 이에 우리는 파일 시스템에 대해 이 같은 특징을 성취하기 위한 방법을 연구할 필요성이 있다.

다. 공표되어진 자료의 역동성 길들이는 방법

Krithi Ramamritham(인도, 봄베이, 인도 과학기술연구소)

(무선) 센서 네트워크에서 수집한 자료와 인터넷을 통해 전달되거나 흘러들어온 자료에는 우리 주변 세계의 급격한 변화와 예측 불가능한 변화가 반영되어 있다. 분명히 그러한 전달을 위한 서비스 품질 필요성은 정적인 자료의 경우보다 훨씬 더 강렬하다. 본 강의에서는 공표 자료의 역동성의 속성을 조사하게 되며 배포 시간이 서로 다른 정보를 배포하기 위한 현재의 인프라의 적합성을 연구하게 된다. 또한 역동적 자료와 그 같은 자료에 대한 검색에 있어서의 시간적 일관성을 유지하기 위한 신선한 접근법을 논의할 것이다. 우리가 주장하는 바는 역동적 자료에 대해 사용자 검색 실행의 경우, 변화 공표, 역동적이면서 합동적인 자료 처리, 내부 네트워크 필터, 총합검색 처리과정 등을 위한 테크닉의 주의 깊은 설계가 요구된다. 우리는 검색 결과와 관련해서 자료의 특징과 정확성 요건을 이용하면, 척도화 가능성을 개선하고 총비용을 감소시켜주는 능률적이면서 효과적인 해결책이 어떻게 도출되는지를 보여준다.

2. Research Session 1 : XML Processing

가. XML 자료에 대한 간편 분류(labeling)체계

Risi Thonangi(인포시스 과학기술사. SET 실험실)

본 논문은 역동적인 XML 나무형(tree) 노드(node)에 라벨을 할당하는 문제를 살펴본다. 라벨 할당은 노드간의 조상-자손 관계를 기호화하고 노드사이의 문서 순서를 부호화하기 위한 것이다. 그와 같은 분류는 능률적인 XML 검색을 촉진시킨다. 이를 위해 수많은 분류체계가 설계되었다. 이 분류체계는 크게 다음으로 분류된다. (1) 정적인(static)분류체계 (2) 역동적인 분류체계. 정적인 체계는 분류 라벨을 짧게 생성하는 반면에 수행능력은 업데이트 집약적 환경에서 떨어진다. 역동적인 체계는 업데이트 수행능력이 좋으나 라벨의 규모가 방대하다. 좋은 분류체계라면 간편 라벨을 생성해야하고 XML 나무형 구조에서 임의 업데이트가 있을 때 더 잘 수행해야 한다. 본 논문은 노드를 섹터(sector)별로 분류하는 섹터기반 분류체계(SL)라고 불리는 새로운 분류체계를 제시하고 있다. 제안한 SL체계를 분석하고, 이것이 정적인 체계보다 더 라벨을 적게 생성하며, 역동적인 체계만큼 좋은 수행능력을 갖고 있음을 보여줄 것이다.

나. 실제 XML 수집 자료에 대한 통계 분석

Irena Mlynkova(Charles대학, 체코)
Kamil Toman(Charles대학, 체코)
Jaroslav Pokorny(Charles대학, 체코)

최근 XML은 자료 표현을 위한 언어 중에서 주도적 역할을 맡았었다. 그러므로 XML자료 관리를 위한 관련 테크닉의 급격한 증가를 목격할 수 있었다. 그러나 대부분의 자료처리 테크닉은 병목현상이 발생하므로 시간이나 공간적 효율성이 악화된다. 주요원인으로는 실제 자료는 종종 훨씬 더 간편할지라도 그러한 테크닉은 XML 수집 자료를 너무 포괄적으로 간주하며 모든 가능한 특징과 관련짓기 때문이다. 비록 몇 가지 테크닉은 입력 자료를 제한하고 있으나 그 같은 제약은 종종 부자연스럽다.

본 논문은 기존의 XML 자료, 그 구조, 특히 실제의 복잡성을 분석한다. 20GB이상의 실제 XML 자료를 수집하였으며 자동분석기를 이용하였다. 유사한 주제에 대한 기존의 논문을 참조하면서 기존 논문의 내용을 확인하거나 반박하려고 하였으며 더불어 새로이 발견한 내용도 있다. 이러한 분석은 빈번히 나타나지만 종종 무시되었던 XML 항목과(혼합내용이나 반복내용 등), XML 체계 및 XML 실제 사례 사이의 관계에 집중하였다.

다. 의미 및 구조 기반 XML 유사성 : 통합접근법

Joe Tekli(프랑스, Bourgogne대학)
Richard Chbeir(프랑스, Bourgogne대학)
Kokou Yetongnon(프랑스, Bourgogne대학)

지난 십년간 XML은 정보관리를 위한 주요 수단으로서 점차로 중요해졌으며, 복잡한 자료의 표현에 필수적인 것이 되었다. XML 표준 사용이 전례 없을 정도로 증가하였기 때문에, XML 기반 문서 비교를 위해 효율적인 테크닉을 개발하는 일은 정보검색(IR) 연구에서 중요하게 되었다. 관련 문헌에서는 계층구조를 이루는 자료, 즉 XML문서를 비교하기 위한 다양한 알고리듬(algorithm)이 제안되었다. 그러나 우리가 아는 바로는 대부분의 경우 구조적 특성을 기초로 해서 문서를 다만 비교하는 것에 초점을 두고 있으며 이와 관련된 의미를 간과하고 있다. 우리의 방법은 내검 거리(edit distance) 알고리즘에서 IR의 의미 유사성 평가를 통합하고 있으며, XML기반 문서 비교에서 유사성 판단을 수정하고자한다. 기존의 작업과는 다르게 우리의 접근법은 원래의 내검 거리 운용비용 모델을 포함하면서 전통적인 내검 거리 연산에 XML 요소 / 속성 라벨의 의미 연관성을 도입한다.

3. Research Session 2 : Indexing and Similarity Search

가. 효과적 검색을 위해 생물학적 문서 저장소 색인에 관계를 이용하는 법

Lipika Dey(인도, IIT 델리)
Rohit Goyal(인도, IIT 델리)

생물의학 텍스트 문서에서 생물학적 관계를 추출할 때 자연언어 처리기법을 사용하는 규칙 기반 메커니즘을 제안한다. 규칙은 잠재적 관계라 할 수 있는 빈번하게 발생하는 패턴을 찾아내는 반면에, 통계적 분석을 통해 중요한 관계를 식별하였다. 테크닉에 대한 평가는 GENIA 군집자료(corpus)에서 얻은 MEDLINE 초록에 대해 실시되었다. 또한 그 결과 우리의 테크닉은 기타의 텍스트 마이닝(mining) 애플리케이션에도 잠재력이 있음이 드러났다. 예비 분석의 결과 이러한 관계에 의해 생물의학 문서를 색인화하면 좀 더 정확한 문서 검색이 촉진되는 것으로 드러났다.

나. 음악 DB에서의 능률적인 유사성 검색

Maria M. Ruxanda(덴마크, Aalborg대학)
Christian S. Jensen(덴마크, Aalborg대학)

오디오 음악은 점차 디지털 형태로 이용가능해지고 있으며, 개인에 의한 디지털 음악 수집은 계속 증가하고 있다. 본 논문은 그러한 수집 자료에서 음악을 검색하기 위한 효과적 수단이 필요하다는 점을 언급하면서, 내용 기반 유사성 검색을 위한 새로운 테크닉을 제안하고 있다. 각각의 음악은 고차원(high-dimensional) 특성 벡터의 시계열로 모형이 되어 있으며 역동적 시간 왜곡(dynamic time warping) (DTW)은 유사성 측정값으로 사용된다. 이를 위해서 본 논문은 시계열 길이 축소(time-series-length reduction)와 다차원 사례까지의 DTW 거리의 하계 (lower bounding)를 위한 테크닉으로 확대된다. 더구나 벡터 근사(approximation) 파일을 조정하여 시계열을 색인화하게 되며 DTW 거리에 대한 하계를 사용하게 된다. 이러한 기법을 사용할 때 본 논문은 좀 더 정확할지도 모르는 결과와는 약간 다르지만, 계산하기에 훨씬 더 비싼 검색 결과를 효과적으로 계산하기 위한 검색을 위한 기초 진실이 부족함을 이용한다. 특히 논문은 검색 확대와 더불어 시계열 길이 축소를 적극적으로 사용하게 되면 중대한 수행능력 향상이 초래되는 반면에 훌륭하면서 근사한 검색 결과를 제공하게 됨을 증명해 보인다.

다. P2P 시스템에서 품질이 보증된 유사성 검색지원

Qi Zhong(미국, 마이크로소프트사)
Iosif Lazaridis(미국, UC Irvine)
Mayur Deshpande(미국, UC Irvine)
Dhen Li(미국, UC Irvine)
Sharad Mehorotra(미국, UC Irvine)
Hal Stern(미국, UC Irvine)

P2P 시스템에서 유사어 검색을 지원하기 위한 방법을 연구하였는 바 그러한 검색은 P2P 네트워크에서 가장 관련성 있는 대상을 요구한다. 이때 관련성은 미리 정의되어진 유사성 기능에 기초한다. 또한 사용자는 가장 연관성이 높은 대상을 획득하는 것에 관심이 있다. 모든 대상을 검색하여 대량 네트워크로 정확한 답을 계산하는 것은 실용적이지 않다. 우리는 새로운 근사 답안 체계를 제안하며 이 경우 네트워크 중 일부만을 방문하여 답을 계산해낸다. 이용자에게는 점차로 세밀해지는 답안이 제시되

며 대답은 지금까지 보아온 것 중에서 최상의 대상으로 구성된다. 더불어 지속적으로 검색의 진행과정에 대한 피드백을 제공하여 품질보증을 꾀한다. 우리는 이러한 체계 내에서 품질보증을 결정하기 위한 통계적 기법을 개발하였다. 우리는 품질 추정량을 (estimator) 검색과정에 통합하는 메커니즘을 제안한다. 우리의 작업은 P2P 네트워크에서 자료에 접근하기 위한 새로운 방법으로서 유사검색의 시행을 가능케 하며 이러한 일이 어떻게 성취될 수 있는지를 보여준다.

4. Research Session 3 : Potpourri

가. 허니팟(honeypot) DB의 망각성 보존

S.K. Gupta(인도, IIT 델리)
Anand Gupta(인도, NSIT 델리)
Renu Damor(인도, IIT 델리)
Vikram Goyal(인도, IIT 델리)
Sangeeta Sabharwal(인도, NSIT 델리)

네트워크 관련 연구자들은 허니팟 개념을 연구하였다. 우리는 DB 허니팟을 제안하였으며 그것의 구조를 [5]에 제시하였다. 허니팟을 사용하는 것은 실제 공격이 발생하기 전에 잠재된 침입자(attacker)를 찾아내기 위함이다. [6]에 주장한 바와 같이 사생활 보호 정책에 따라 의심이 가는 이용자에게 접근이 거부되리라고 기대된다. 그러나 변장을 하고 시스템에 진입하였을지도 모르는 의심이 가는 이용자를 찾아야한다. 우리는 의심이 가는 이용자에게 (접근 거부 대신) 종합 정보를 제공할 것을 제안하는 바이며, 이러한 종합 정보의 도움으로 관리자는 혐의를 확인할 수 있을 것이다. 본 논문은 허니팟의 그러한 몇 가지 특성을 제시한다. 즉 미끼를 던져서 혐의를 결정하고 이용자에게 투명성을 제시하는 점을 특징으로 들 수 있다. 또한 이용자가 그러한 허니팟을 망각하게 하기 위한 테크닉을 제시한다.

나. 스트림(stream) 자료의 핵심(kernel) 밀도 추정에 대하여

Christopher Heinz(독일, Marburg대학)
Bernhard Seeger(독일, Marburg대학)

다양한 실생활 애플리케이션은 변천하는 자료 흐름에 대한 분석에 크게 의존한다. 자료 흐름의 경직된 처리 필수요건 때문에, 데이터 마이닝(data mining)이라고 알려

진 일반적인 분석기법은 적용 불가능하다. 데이터 마이닝과 분석접근법의 기본적인 원리는 밀도추정이다. 이것은 연속적 자료 분포에 대해 잘 정의되어진 추정을 제공하며, 이 사실은 자료 흐름에의 적응을 바람직하게 해준다. 밀도추정을 위한 편리한 방법은 핵심을 이용한다. 그러나 이러한 방법의 연산상의 복잡성은 자료 흐름의 처리 필수요건과 충돌한다. 이번 작업에서 우리는 이러한 문제에 대한 새로운 접근법을 제시하며 이것은 선형(linear) 처리비용을 할당되어진 메모리의 불변량과 결합한다. 심지어 우리는 시스템 자원을 변화시키는 역동적 메모리 조정을 지원한다. 우리의 스트림 자료의 밀도추정량은 예전에 제안된 기법인 M-Kernel과도 관련이 있으나 본질적으로는 처리시간뿐만 아니라 정확도 측면에서 개선되어진 형태이다. 종합데이터 및 실생활 자료 흐름에 대한 실험연구의 결과, 추정 품질 및 자료 처리시간 측면에서는 M-Kernel보다 우수할 뿐만 아니라 우리의 접근법이 능률적이고 효용성이 있는 것이 입증되었다.

다. Genea: 온톨로지(ontology) 관계형 DB로 도식인지 지도화 (mapping) 하는 방법

Tim Kraska(독일, Muenster대학)
Uwe Rohm(호주, 시드니대학)

바이오정보학이나 보건과 같은 애플리케이션 영역에서뿐만 아니라 의미론적 웹상에서 자료를 서술하고 교환하는데 있어서 온톨로지는 중요한 메커니즘이 되었다. 이 논문은 RDBMS를 이용해서 온톨로지 사례 자료를 효과적으로 저장하고 검색하는 방법에 관한 문제를 다루게 된다. 우리의 접근법은 총칭적 지도화 규칙을 이용하여 온톨로지 도식인지 관계 표상을 자동적으로 생성한다. 포섭관계 등의 온톨로지 추론의 일부는 도식생성 중에 그리고 또한 실례자료의 로드타임에 실시되므로 검색과정이 더 빨라지게 된다. 우리는 지니(Genea)라고 불리는 OWL-지도화 툴로 우리의 접근법을 실시하였으며, 이것은 자동적으로 간결한 관계형 체계를 생성할 수 있으며 또한 OWL로 쓰인 주어진 온톨로지로부터 사례자료를 입력할 수 있다. 우리는 의미론적 웹 공동체로부터 도식망각 및 도식인지 RDF store와 비교하여 우리의 접근법을 양적으로 그리고 질적으로 평가한 결과를 보고할 것이다. 지니는 온톨로지 지도화를 위한 총칭적이면서 효과적인 툴인 것으로 밝혀졌다. 생성되어진 도식은 좀 더 많은 의미를 포착하기 때문에 기존의 접근법보다 더 빠른 검색을 허용한다.

5. Research Session 4 : Web and Distributed Data

가. 센서데이터에 대한 고장방지 검색

Iosif Lazaridis(미국, UC Irvine)
Qi Han(미국, Colorado 광산학교)
Sharad Mehrotra(미국, UC Irvine)
Nalini Venkatasubramamian(미국, UC Irvine)

장애가 발생한 경우 센서의 의해 생성되어진 값에 대한 연속적 선정검색을 평가하는 문제를 살펴본다. 작은 센서는 고장 나기 쉬우며 유한의 에너지와 메모리를 가지며 상실된 매체로 의사소통한다. 그러므로 그것에 의해 생산된 투플(tuple)은 검색 노드에 도달하지 못하게 되어 결국 불완전하고 애매모호한 답이 도출된다. 왜냐하면 보고하지 않는(non-reporting) 센서 중 어느 것이라도 상실된 투플을 생성했을 수도 있기 때문이다. 우리는 연속선정 검색의 고장방지평가(FATE-CSQ) 프로토콜을 개발하였으며 이것은 효율적으로 사용자 요구에 의한 품질수준을 보증한다. 많은 장애가 발생하는 경우 이것은 성취될 수 없을는지도 모른다. 그러한 경우 검색시간이 제한된 상황에서 최상의 가능한 답을 목표로 삼게 된다. FATE-CSQ는 서로 다른 종류의 실패에 대해 탄력적이도록 설계되었다. 우리의 디자인은 피드백과 재전송에 기초한 서로 다른 고장방지 전략의 분석 모델에 기초하고 있다. 또한 사실적인 시뮬레이션과 여러 조건하에서 경쟁상태에 있는 프로토콜과 비교해서 FATE-CSQ가 훌륭한 수행능력을 보여준다는 점을 증명할 것이다.

나. EcoRep : 모바일 P2P 네트워크에서 효율적이며 역동적인 복제를 위한 경제 모델

Aniraban Mondal(일본, 도쿄대학)
Sanjay Madria(미국, 미주리-롤라대학)
Mosaru Kitsuregawa(일본, 도쿄대학)

모바일 P2P(M-P2P) 네트워크에서는 빈번하게 네트워크를 분할하게 되면 보통 낮은 자료 이용가능성이 유발된다. 그리하여 자료 복제가 필수사항이 된다. 본 논문은 M-P2P 네트워크에서 역동적인 복제 배치를 위한 새 경제모델, EcoRep을 제안한다. EcoRep은 자료항목의 상대적 중요도에 기초하여 복제를 실시하며 이것은 가상통화(currency)에 의한 자료항목의 가격에 의해 수량화된다. 자료항목의 가격은 접속 빈도수, 접속 이용자수, 존재하는 복제품수, (복제)일치 여부 및 접속에 필요한 평균 응답

시간에 의존한다. EcoRep은 자료항목에 대한 검색의 기원을 고려하여 공정한 복제의 배치를 실시한다. EcoRep은 검색을 요청한 이용자로 하여금 요청한 자료항목의 가격을 요청에 응한 이용자에게 지불할 것을 요구한다. 이렇게 하면 무임승차(free-riding)를 방지하며, 서비스 제공자가 되는 이용자에게 인센티브를 제공함으로써 이용자 참여를 촉진하게 된다. EcoRep은 또한 EcoRep은 복제 기준으로서 로딩, 에너지, 네트워크망의 형태(topology)등의 다른 문제도 고려한다. 우리가 실시한 수행능력 연구에 따르면 EcoRep은 검색 응답시간 향상과 M-P2P 네트워크에서 자료의 이용가능성 증진에 정말로 탁월하다.

6. Poster Session

가. P2P 분산 저장시스템내의 척도화 가능 복제품 관리방법

Jing Zhou(중국, 국방과학기술국립대학)
Yijie Wang(중국, 국방과학기술국립대학)
Sikun Li(중국, 국방과학기술국립대학)

P2P 분산 저장 시스템내의 수많은 복제품으로 인해 불일치와 로드 불균형이 악화된다. 그와 같은 자료 관리 문제점에 따라 분산형이면서 구조화되지 않은 P2P 네트워크에 기초한 척도화 가능 복제품 관리 방법이 제안되었다. 복제품은 단하나의 복제품 복제에 따라 서로 다른 계층과 클러스터로 분할되며 그 후 복제품은 사용자가 정의한 계층-코딩 규칙을 기초로 해서 부호화되고 관리된다. 그 후 복제품은 로컬내의 집중과 광범위한 지역 내의 P2P로 조직화된다. 조정 일치 비용은 정의되어진 번식시간 계획과 결합으로 크게 하락할 수 있다. 시뮬레이션 결과 이것은 효과적인 다수의 복제품관리 방법이며 훌륭한 척도 가능성이 성취되며 빈번한 업데이트가 실시되는 애플리케이션에 잘 적응하는 것으로 나타났다.

나. 데이터 및 항목집합을 암호화한 접두어 나무형 구조(prefix tree)

Ramkishore Battacharyya(인도, Jadavpur대학)

마이닝 알고리즘의 주요 기준은 공간-시간 필수조건에 문제를 제기하는 것에 대한 자료유입을 환영하는 것이다. 데이터가 치밀하고 효율적으로 배치되어 있지 않다면, 제한된 주기억 장치를 갖고 있는 알고리즘은 적절한 시간 내에 결과를 산출하지 못하게 된다. 본 논문은 제어 프로그램 실행시간이 거의 걸리지 않으면서, 매우 치밀하

고 기억장치에 효율적인 접두어 트리구조를 구축하게 해주는 데이터 암호화 기법을 제시한다. 표시할 때 공통 접두어가 있는 트랜잭션(transaction)을 함께 배열할 수 있는데 그렇게 하면 마이닝 알고리즘의 처리능력이 증가된다. 우리는 또한 암호화된 항목 집합을 체계적으로 관리하기 위한 새 자료구조, 이분지 검색 접두어 나무형(BSPT, Binary Search Prefix Tree)을 소개하고자한다. 실험 결과는 새로운 자료 구조 통합으로 알고리즘이 더 큰 정도로 척도화 가능해짐을 보여준다.

다. XML 확장 OLAP 검색의 대수기반 최적화

Xuepeng Yin(덴마크, Aalborg대학)
Torben Bach Pedersen(덴마크, Aalborg대학)

오늘날의 OLAP 시스템에서 빠르게 변화하는 데이터를 물리적으로 큐브로 통합하는 일은 복잡하고 시간이 많이 걸린다. 우리가 내놓은 해결책, "OLAP-XML 연합 시스템"은 물리적으로 통합하지 않고도 OLAP 검색에서 XML 포맷으로 빠르게 변하는 데이터를 참고할 수 있게 해준다. 본 논문은 검색 옵티마이저(optimizer)를 포함해서 연합 시스템에 대해 전문성을 띠는 새로운 검색 최적화 테크닉을 소개하고 변형규칙을 계획해본다. 또한 물리적 통합과는 다르게 우리의 접근법은 빠르게 변하는 데이터를 OLAP 시스템에 통합하기 위한 실용적 해결책임을 보여주는 실험 결과를 제시할 것이다.

라. 인트라넷에서의 공생(symbiosis) : 문서 검색이 DB 정보에서 혜택을 보는 방법

Christoph Mangold(독일, 슈투트가르트대학)
Holger Schwarz(독일, 슈투트가르트대학)
Bernhard Mitschang(독일, 슈투트가르트대학)

기업의 정보 공간은 두 구역으로 분리되어 있다. 문서는 구조화되어 있지 않거나 부분적으로 구조화된 정보를 포함한다. 반면에 DB에는 구조화된 정보가 저장된다. 내용의 경우 두 가지 종류의 정보가 서로 상호보완적 관계에 있다. 그러나 보통 기업 정보 시스템은 다만 한 가지 부분에만 초점이 주어진다. 우리의 접근법은 기업의 DB를 이용해서 인트라넷 안에서의 문서 검색을 향상하는 것이다. 특히 문서의 문맥을(context) 서술하기 위해 DB 정보를 활용하며 이러한 문맥을 이용해서 텍스트 전체 검색을 신장하게 된다. 본 논문은 문서의 문맥을 모형화하고 계산하는 방법을 보여주

고 또한 실행상의(runtime) 수행능력 결과를 제시한다.

7. Demonstration Session

가. 평면적인(flat) 데이터 기록과 내포된(nested) 데이터 기록에서 웹 데이터를 비주얼 단서(visual clue) 기반으로 추출하는 법

Siddu P Algur(인도, Engg SDM대학)
P S Hiremath(인도, Gulbarga대학)

주어진 웹 페이지의 내포기록과 평면기록에서 구조화된 자료항목을 식별하고 추출하는 문제를 연구하였는 바 페이지 각각에는 구조화된 기록의 여러 집합이 포함되어 있는데 기존의 방법 대부분은 약간의 한계가 여전히 있어 자료항목 추출을 위한 좀 더 신선하고 효과적인 기법을 제안한다. 페이지가 주어지면 우리가 제안한 테크닉은 우선 비주얼 단서 정보를 기초로 해서 자료구역을 찾아내며 그 후 해당 자료 구역에서 각 기록을 추출하며 비주얼 정보를 기초로 해서 그것이 평면적 기록인지 내포된 기록인지를 알아낸다.(포함된 영역과 각 기록에 들어있는 자료항목의 수).

다음 단계는 이러한 기록에서 자료항목을 추출하여 DB로 이송하는 것이다. 일단 자료항목이 DB에 존재하게 되면 지식 발견을 시행할 수 있게 된다. 이 테크닉은 내포된 기록과 평면기록 둘 다에서 자료항목을 추출한다.

실험의 결과 우리가 제안하고 있는 테크닉은 기존의 테크닉보다 효율적이고 더 우수한 것으로 드러났다.

나. DB 자원에 유비쿼터스하게 접속하기 위한 검색 인터페이스

Subhash Bhalla(일본, Aizu대학)
Mashki Hasegawa(일본, Aizu대학)

대부분의 정보시스템은 정보를 구성할 때 DB 관리시스템(DBMS)에 의존하게 되며 그러한 정보시스템 접속은 사용자 측면에서 DB 검색 언어사용에 기초하게 되는데 이것은 사용자 기술(혹은 기술수준) 상의 문제를 유발시킨다.

예를 들어 병원의 의료담당 직원의 경우 DB 검색 언어학습에 시간을 할애할 수 없을 것이며 그 결과 의료담당 직원은 정보 접속을 위해 전문가와 프로그래머에게 의존하게 되는 경향이 있다.

검색 언어의 복잡성을 제거하고 유비쿼터스 접속을 위해서 웹 기반 정보접속 시스

템을 최종 사용자를 위해 제안하는 바 우리가 제안한 시스템은 부분적 기술을 갖춘 직원으로 하여금 잘 훈련받은 DB 검색 언어 프로그래머와 동일한 수준으로 검색을 할 수 있게 해줄 것으로 예상된다.

다. UnURL : URL로부터 관리감독을 받지 않은 학습

Deepak P(인도, IBM IRL)
Deepak Khemani(인도, IIT 마드라스)

웹 페이지는 URL로 식별되어 진다. 권위 있는 웹 페이지, 즉 특정 주제에 집중하고 있는 페이지의 경우 웹 마스터는 해당 페이지를 요약해주는 URL을 사용하는 경향이 있다.

URL 정보는 클러스터링(clustering)에 좋은데 이는 작고 유비쿼터스하므로, URL 정보크기에 기초한 테크닉을, 텍스트 내용까지도 사용하는 테크닉보다 더 빠르게 해주기 때문이다.

따라서 URL 정보만을 이용하는 시스템을 제시하는 바 이 시스템은 웹 검색 결과 집합의 클러스터링과 일반 웹 문서 군집자료(corpus)의 클러스터링과 총론적 URL 군집자료의 주제식별을 수행한다. UnURL이라고 부르는 이 연구방법은 우리가 알기로는 URL에 대한 관리감독을 받지 않으면서 기계적 학습 테크닉을 사용하는 최초의 시도이다.

라. OCHD: 허니팟 DB의 망각성 특징 보존

S.K. Gupta(인도, IIT 델리)
Renu Damor(인도, IIT 델리)
Anand Gupta(인도, NSIT 델리)
Vikram Goyal(인도, IIT 델리)

허니팟 DB의 (문맥 허니팟) 목적은 잠재적인 사생활 보호 위반자를 식별하기 위한 것이다. 의심이 가는 이용자의 DB와의 상호작용을 분석하여 의심을 확정하게 된다. 그러한 시스템은 일정한 특징을 만족시켜야 하는데 특징 중 하나가 망각성이다. 이와 같은 시스템의 성공여부는 의심이 가는 이용자가 망각한 채로 있느냐에 달려있다. 이번 연구에서는 OCHD(허니팟 DB의 망각성 특징)라고 약칭되어지는 그러한 시스템의 구조와 업무흐름을 살펴보았다.

마. 밀도 추정량을 통한 스트림 마이닝(stream mining) : 구체적 애플리케이션

Christopher Heinz(독일, Marburg대학)
Bernhard Seeger(독일, Marburg대학)

많은 실제 애플리케이션은 처리한 자료가 스트림으로 도달한다는 특징을 공유하는데 이러한 스트림의 과도기적인 속성과 변화 속성은 자료처리 및 분석 테크닉의 적용을 어렵게 만든다.

특히 스트림 마이닝은 데이터 스트림 시나리오 안에서 충족시켜야만 하는 경직된 자료처리 필수요건 때문에 어려운 과제인 것으로 입증되었다. 우리는 스트림 마이닝에 대해 핵심 밀도 추정을 이용할 것을 제안하는 바이다. 수학통계 영역에서 온 테크닉인 핵심 밀도 추정은 많은 마이닝 관련 주제와 애플리케이션에서 자리를 잡고 있다. 그러나 연산비용이 많이 드는 것 때문에 스트림에 직접 응용하는 것은 불가능하다. 이러한 점 때문에 기존의 연구에서 스트림에 대해 연속적으로 핵심밀도 추정량을 계산하는 정교한 인용법을 개발하였다. 이러한 추정량을 통해서 우리는 스트림 자료에 대해 다양한 마이닝 작업을 시행할 수 있다. 우리는 의료 환경에서 구체적인 애플리케이션을 배경으로 해서 일부를 설명하고자 한다. 좀 더 정확하게 얘기하자면 환자의 바이탈사인(vital sign)의 온라인 분석을 비주얼로 보여주는 의료 모니터 실시를 설명할 것이다. 모니터를 설명하는 것 외에도 또한 이와 같은 실시의 기저개념도 제시할 것이다.

바. 문서와 DB와의 만남 : 인트라넷 검색시스템

Christoph Mangold(독일, 슈투트가르트대학)
Holger Schwarz(독일, 슈투트가르트대학)

기업의 인트라넷에서는 정보가 문서와 DB안에 암호화된다. 논리적으로는 이 두 분야의 정보는 큰 차이를 보이는 시스템 수준에서 밀접하게 연결되어 있다.

논문은 기업의 인트라넷에서 문서를 검색하는 시스템을 제안하는데 이 시스템은 보통의 텍스트 검색의 연장선상에 있으며 이것은 문서의 내용을 고려할 뿐만 아니라 문서의 문맥을 결정하기 위해 기업의 DB도 이용한다.

사. 발견서비스 - EPCglobal 네트워크에서 RFID 추적가능성의 실현

Steve Beier(미국, IBM SVL)
Tyrone Grandison(미국, IBM Almaden)
Karin Kailing(미국, IBM Almaden)
Ralf Rantzau(미국, IBM SVL)

EPCglobal 컨소시엄은 기업내에서, 그리고 기업 사이에서 전자제품 코드 관련 정보공유를 가능케 하는 표준을 정의내렸다. 이것은 보통 상표가 붙어 있는 제품에 대한 제품정보 뿐 만 아니라 RFID reader의 event도 포함한다. EPCglobal 네트워크는 노드로 구성되어 있으며 각각의 노드는 복잡한 자료이용 및 공유정책을 갖고 있으며 네트워크에서 전체 값을 추출하기 위해서는 협력할 필요가 있다. 발견(discovery) 서비스는 아마도 이질적인 기업 체계에서 이 네트워크를 생성하고 채택하는데 있어서 중요한 단계이다.

우리의 작업은 발견 서비스 구축을 위해서 간단하면서 척도화 가능한 인프라의 초기 시행 제공에 초점을 두고 있다. 시연을 통해 EPCglobal 네트워크의 세 가지 요소 사이의 상호작용을 설명한다.: 발견 서비스, EPC 정보서비스(EPCIS) 그리고 EPCIC 검색을 위해 발견 서비스를 이용하는 애플리케이션 등. 이것은 복수의 EPCIS 서버가 호스트 서버이다. 애플리케이션은 식품 공급 체인 전체에 걸쳐 아보카도의 신선도를 추적하며 불만족스러운 제품의 수송로 변경을 허용한다. EPCglobal에 의한 발견서비스의 표준화 문제는 아직 해결되지 않고 있다.

아. 의미와 구조에 기반을 둔 XML유사성: XS^3 원형(prototype)

Joe Tekli(프랑스, Bourgogne대학)
Richard Chbeir(프랑스, Bourgogne대학)
Kokou Yetongnon(프랑스, Bourgogne대학)

XML 기반 데이터의 이용가능성이 점차 증가하고 있기 때문에 XML 문서를 비교하기 위한 효율적인 접근법은 정보 검색에서 중요해지고 있다. 그러한 XML 문서비교는 버전 제어(서로 다른 문서 버전사이의 변화를 찾아내어 평가하고 검색하기), 변화관리 및 자료 저장(시간 검색지원 및 색인 유지보수) [3, 4, 5], XML 검색시스템(최상의 결과 검색을 위해 유사성에 따라 결과를 찾아내고 순위매기기) [10, 12], 그리고 XML DB의 DTD 집합에 맞서 웹에서 수집한 XML 문서의 분류/클러스터링(마치 전통적 DBMS에서 능률적인 저장, 검색, 보호 및 색인 편의제공을 위해 도식이 필요한 것처럼 DTD 및 XML 저장소의 경우에서도 똑같다.)[1, 2, 8]등에서도 애플리케이션이 있다. 여기서 우리는 내검 거리구조 유사성 알고리듬에서 IR 의미 유사성 평가를 통

합할 수 있는 XML 비교 원형 XS^3(XML 구조 및 의미 유사성)을 제시한다. 또한 이질적인 XML 기반 문서를 비교할 때 유사성 판단을 수정하려고 노력한다.

자. OLAP-XML 연합시스템

Xuepeng Yin(덴마크, Aalborg대학)
Torben Bach Pedersen(덴마크, Aalborg대학)

"OLAP-XML 연합스시템"은 XML 포맷으로 이용가능한 외부자료를 가상 디멘션으로 사용할 수 있도록 해주며현재의 OLAP 시스템안에 있는 복잡하고 시간이 많이 걸리는 OLAP와 외부자료의 물리적 통합과는 다르게, 우리의 시스템은 OLAP검색이 빠르게 변하는 외부자료를 참고할 수 있도록 해준다.

차. BlogHarvest: 블로그 마이닝과 검색 틀

Mukul Joshi(인도, 소프트웨어실험연구소)
Nikhil Belsare(인도, 테크 Mahindra유한회사)

블로그는 온라인상의 일기라는 개념을 뛰어넘어서 복잡한 사회 구조물로 진화하였다. 블로그용 소프트웨어는 이용자로 하여금 미리 정의되어진 도식에 따라 아무런 제약 없이 어떤 주제에 대해서라도 의견을 발표할 수 있도록 해준다. 블로그 사이의 연결을 분석한 결과 블로그 영역을 형성하는 공동체는 임의과정이 아니라 블로거들을 결합하는 공유된 관심의 결과임을 알 수 있었다.

이용자의 관심과 블로그로부터의 사회적연결에 대해 학습하고 분석하고 이용하려면, 블로거들에게 블로그 영역에서 유용한 검색기능을 제공해야 하며 블로그 서비스 제공자에게 광고와 같은 수입 창출 기회를 제공해야 한다. 본 논문에서 우리는 BlogHarvest를 시연할 것이다. BlogHarvest는 블로그 마이닝과 검색 틀이며 블러거의 관심을 추출하여 유사한 주제를 가진 블러그를 찾아내어 추전하고 블로그 지향 검색 기능을 제공하는 것이다. BlogHarvest는 이러한 특징을 제공하기 위해 분류, 연결 및 주제 유사성에 기반을 둔 클러스터링과 POS tagging에 기반을 둔 오피니언 마이닝을 이용한다. 보통의 검색 순위와 더불어 검색에 대한 관련 블로그를 제공하기 위해 새로운 검색 인터페이스가 구축된다. POS tag에서 발견된 연합규칙을 사용하여, 목표로 하는 결과를 획득하기 위해서 검색 확장을 제공하기 위해 검색 문맥을 획득하게 된다. 블로그 영역을 살펴보고 블로그 post를 추출 및 색인하고 연결 메타데이터를 살피는 등 우리는 우리의 알고리듬을 조율하기 위해 약 5만개의 블로그를 분석하였다.

8. Application and Industrial Session

가. 분산 제약 위반의 효율적인 탐지

Shipra Agrawal(미국, 스탠포드대학)
Supratim Deb(인도, Bell 실험연구소)
K.V.M Naidu(인도, Bell 실험연구소)
Manish Arya(인도, Bell 실험연구소)

많은 분산환경에서, 모니터용 소프트웨어의 주요 기능은 변칙성, 즉 시스템의 행위가 기준에서 상당하게 일탈한 사례를 찾아내는 것이다. 그러한 비정상적인 행위를 찾아내기 위한 기존의 접근법은 항상, 심지어 정상적으로 작동되는 시간에도 시스템 상태를 기록한다. 그러므로 총통신비의 낭비가 초래된다.

이에 변칙성 간파 문제를 위한 통신면에서 효율적인 도식을 제안하는 바 분산 시스템 변수에 대해 정의되어진 포괄적인 제약 위반을 찾아내는 것을 모형화하였다.

이러한 접근법은 포괄적인 제약을, 각 사이트에서 효율적으로 점검할 수 있는 로컬 제약으로 분해함으로써, 포괄적인 시스템 상태를 지속적으로 추적한 필요를 제거해준다. 로컬 제약이 위반된 경우에만 우리는 좀더 비용이 많이 드는 포괄적인 제약 점검을 실시한다. 로컬 제약 선정의 문제를, 개별 시스템 변수의 도수분포를 고려한 최적화 문제로 공식화하였는데 이는 통신비용을 최소화하기 위함이다.

문제가 NP-hard (어렵다) 하다는 것을 보여준 후, 입증 가능한 근사 최적(메시지의 수 측면에서) 로컬 제약을 계산하기 위한 근사 알고리듬을 제안한다. 실생활 네트워크 교통 자료 집합을 가지고 실시한 실험에서, 우리는 포괄적인 제약 위반을 찾아내기 위한 우리의 테크닉은 기존의 데이터 분포-논쟁 유발 접근법과 비교해 볼 때 70%만큼 메시지 총통신비를 줄일 수 있는 것으로 드러났다.

나. 엔지니어링 웹기반 기업 애플리케이션

Srinivasa Narayanan(미국, Tavant 테크놀로지)
Subbu N. Subramanian(미국, Tavant 테크놀로지)
Manish Arya(미국, Tavant 테크놀로지)

웹기반 애플리케이션과 웹이 가능케 하는 기존의 애플리케이션 구축의 맥락에서, 기업은 SOA와 요구응답(On-demand) 기술을 빠르게 채택하고 있는 바 이러한 것들은 새

로운 엔지니어링 문제를 일으키고 있다. 우리는 Fortune지 선정 100명의 기업 소비자에게 기업 애플리케이션을 구축한 우리의 경험을 기초로 해서 이러한 문제점을 토론하였다. 통합은 이러한 애플리케이션에서 중요한 역할을 하고, 부분적으로 통합할 필요성, 시스템 사이의 연결성 문제점을 다루는 것과, legacy 시스템과 통합하는 것 등의 도전과제를 제시한다. 메타데이터와 파트너 시스템에 대한 데이터 등을 캐싱(caching)하는 것과, 통합에서 수행능력 문제점을 언급하려는 사용자 기대를 기초로 해서 의미를 조정하는 테크닉을 이용하였다. 많은 경우, 고객은 요구에 따라 해결책이 이용가능해질 것을 요구한다. - 다시 말해서, 다수의 고객 사이에서 전체 하드웨어와 소프트웨어 스택(stack)을 공유하는 능력, 애플리케이션 다운타임(down-time)없이 소프트웨어 업그레이드를 실시하는 능력, 그리고 고객에게 다양한 맞춤 생산 특징을 제공하는 능력과 더불어 복수 차용(multitenant) 모드로 호스트화 된다. 우리는 이러한 애플리케이션을 구축한 시스템 엔지니어의 관점에서, 이러한 맥락에서 우리가 직면한 경험과 도전과제를 공유한다.

9. 개별 강의자료

가. 사생활을 보호하는 자료 공표: 일반화로부터 해부까지

Yufei Tao(중국, 홍콩중국어대학)

회사 및 단체는 연구 목적으로 기관에 고객 정보를 발표할 필요가 있는 경우가 있다. 예를 들어서 병원은 정기적으로 환자의 진료 기록을 발표하여 의학자들이 질병과 여러 다양한 요인과의 상관관계를 연구할 수 있도록 한다. 자료 공표에 있어서 사생활 보호는 중요한 문제이다. 첫째, 어떤 환자의 경우든 상대방이 정확한 병력을 알아낼 수 없도록 자료 공표는 충분히 불분명해야한다. 한편 공표 자료는 효율적 분석이 가능할 만큼 충분히 정확해야한다. 본 강의는 환자의 사생활을 침해하지 않으면서 자료 조사의 정확성을 최대화하는 것 사이에서 균형을 이루기 위한 방법을 검토한다.

나. 다국어로 된 DB 시스템

Jayant Haritsa(인도, 인도과학연구소)

오늘날의 세계화된 세상에서는 여러 개의 자연언어로 자료를 효율적으로 저장하고 검색하는 일은 매우 중요하다. 이와 같은 목표 달성의 중요한 필수조건은 기존의 표준 데이터 저장소가-관계형 DB 시스템- 효과적으로 그리고 빈틈없이 다국어로 된 데이터

를 지원해야한다는 점이다. 본 강의는 먼저 오늘날의 DB 시스템이(상업적 영역뿐만 아니라 공적인 영역 둘 다) 다국어로 된 자료의 저장, 관리 및 자료 처리 측면에서 얼마나 훌륭한지를 자세히 평가한다. 조사 결과, 라틴어 외의 글자에(Devanagari, Kanji, 키릴어 등) 기초한 언어에 대해서는 중대한 수행상의 비효율성이 있음을 살펴보기로 하자. 더불어 이 같은 문제점을 완화하기위한 방법을 약술할 것이다.

기능성 측면에서 SQL의 주요 한계는 서로 다른 자연언어간의 자료 검색, 다시 말해서 교차언어적(cross-lingual) 검색을 지원하지 못한다는 점이다. 이와 같은 허점을 해결하기 위해, 우리는 다국어 세상에서 이름의 음소에 기반을 둔 일치를 지원해주고, 개념의 온톨로지(ontology)에 기반을 둔 일치를 지원해주는 두개의 새로운 SQL 운용자(operator)를 제안할 것이다.

관계형 시스템과 이 새로운 운용자를 통합하기 위한 대수뿐만 아니라 관련 비용 모델, 선택성 추정량 및 접속 방법도 정의될 것이다. PostgreSQL에 대한 이 같은 운용자의 시범적 운행에 관한 우리의 경험을 집중적으로 살펴볼 것이다.

간단히 말해서, 본 강의는 "자연언어에 대해 중립적인" DB 엔진의 궁극적 목표 구현을 위한 실용적 접근법을 제시한다.

다. 안전한 데이터 아웃소싱

Radu Sion(미국, Stony Brook대학)

오늘날의 자료 관리 서비스의 네트워크화 되어진 속성과 점진적으로 증가하고 있는 유비쿼터스 속성은 악의성이 있거나 잘못된 행위를 탐지하고 방지하는 것을 보장하고 있다. 이것은 특히 아웃소싱 자료와 관련이 있으며, 여기서 고객은 자료 관리를 전문적인 서비스 제공자에게 위탁한다. 고객은 마지못해서 민감한 자료에 대해 기밀 유지 보장 없이 제 3자의 관리 하에 있게 한다. 그 외에도, 일단 아웃소싱이 이뤄지면, 사생활 보호 및 데이터 접속 정확도(데이터의 완벽성 및 검색의 완벽함)가 중요해진다.

현재의 해결방안은 기본적으로 불법행위에 대해 불안하며 문제가 되기 쉽다. 왜냐하면 이러한 디멘션을 처리하지 않기 때문이다. 본 강의에서 우리는 기존의 해결방안과, (1) 정확도 (2) 기밀 유지 (3) 데이터 접속 사생활 보호 등의 강력한 안전 보장을 제공하는, 굳건하고 능률적이며 척도화 가능 데이터 아웃소싱 메커니즘에 대한 미래 디자인을 토론할 것이다.

그러한 보장 사이에는 강력한 관계가 존재한다. 예를 들어서 접속 패턴 사생활 보호

부족으로 인해서 통계적 공격을 허용하게 되며 자료의 기밀 유지와 타협이 이루어진다. 기밀 유지는 자료 암호화를 통해 달성될 수 있다. 그러나 실제로는 아웃소싱 자료 서비스는 기밀 유지와 타협이 이루어지지 않고도 과도한 고객 검색을 허용해야한다.(예를 들어서 임의 서술어와의 관계형 연결). 이것은 어려운 문제이다. 왜냐하면 복호화 키를(decryption key) 잠재적으로 신뢰할 수 없는 서버에게 직접 제공할 수 없기 때문이다. 게다가 멀리 떨어져있는 서버를 완전히 신뢰할 수 없는 경우, 프로토콜의 정확도는 필수적이 된다. 그러므로 이 세 디멘션을 언급하지 않는 해결 방안은 불완전하고 불안하다.

(i) 각각의 대상 자료 집합에 대해서 완전무결하고 완벽하게 검색이 실행되도록 하고, (ii) 암호화된 자료에 대해 기밀이 유지된 채 검색이 실행되는 것을 허용하고, (iii) 고객 검색 및 데이터 접속 패턴의 사생활 보호를 보장하는 아웃소싱 관계형 자료를 대상으로 하는 검색 메커니즘을 디자인하는 것이 중요하다. 우리는 신뢰받는 하드웨어의 존재에 대해 조정이 이루어지는 프로토콜을 토론한다. - 그러므로 중요한 기능이 안전하게 고객에게서 서버로 위임될 수 있다. 우리는 아웃소싱 시나리오에서 완전히 사생활이 보장되면서 이분지 서술어(predicate) JOIN를 처리하는 것을 실제 프로토콜을 가지고 예를 들어 볼 것이다.

라. 측정가능(scalable) 정보의 추출 및 통합

Sunita Sarawagi(인도, 봄베이, 인도 과학기술연구소)

텍스트에 대한 많은 애플리케이션은 방대한 구조화되어있지 않은 출처로부터 구조화되어 있는 데이터를 추출하고 통합하기 위한 효율적 방법이 필요하다. 본 강의에서는 정보 추출 및 중복 제거를 위한 최신 접근법을 검토하게 된다. 첫째, 조건적 랜덤 필드(conditional random field)와 같은 순차(sequential) 모델과 일반화 모형이 포함된 접근법을 기초로 한 기계학습을 특히 강조하면서, 사용되어지는 방법의 개요를 제시하고자 한다. 둘째, 방대한 구조화되지 않은 텍스트 집합에 대해 이러한 모델을 전개하기 위한 척도화 가능한 기법을 제시한다. 정보 추출 애플리케이션을 전문으로 하는 색인 기법뿐만 아니라 일반 목적 검색 엔진 사용을 포함해서, 정보 추출을 척도화하기 위한 주요 접근법을 검토한다. 또한 적절한 결합(join) 연산식과 퍼지(fuzzy) 색인 룩업(lookup)을 이용하여 추출되어진 정보를 통합하기 위한 척도화 가능 기법을 개략한다. 더불어 연구 기회와 미해결 도전과제를 조명하고자한다.

마. 데이터그리드관리시스템(DGMS) 소개

Arun Jagatheesan(미국, 캘리포니아대학, SDSC)

컴퓨터과학의 역사를 분석해보면, DBMS를 포함한 대부분의 기여는 기존의 과학기술의 한계를 극복하고자하는 필요의 결과였다. 오늘날의 그 같은 필요중 하나는 전통적 파일 시스템을 이용하여 분산된 각 국가의 거대한 구조화되지 않은 자료의 관리문제이다. Fortune지 선정 500위에 드는 회사의 일부는 현재 이 문제에 직면하고 있는데 그것은 상호 협조하는 아웃소싱 및 분산된 글로벌 팀 때문이다. 데이터그리드관리시스템(DGMS)은 여러 팀간의 방대한 양의 구조화되지 않은 자료를 합작으로 전 세계적으로 공유하는 것을 관리한다. DGMS의 핵심 개념은 전통적인 RDBMS와 매우 유사하다. DGMS는 여러 개의 하위조직으로 된 이질적 자료 저장 자원의 논리적 명칭 공간(namespace) (혹은 논리적으로 분산된 파일 시스템)으로 간주될 수 있다. DGMS는 관계형 DB에 의해 촉진되며 데이터를 구성 및 검색하기 위한 시스템과 사용자정의 도식(schema) 둘 다를 제공한다. 본 강의는 DGMS의 개념을 소개한 후 그러한 시스템이 매우 방대한 학술 자료 센터 및 주요 회사에서 왜 필요한 지를 실제 이용 사례를 들어 설명하고 있다. 분산된 자료 관리에 있어서 신참자와 전문가 모두 이 새롭게 부상하고 있는 과학기술과 연구 문제점 및 사업 기회에 대해 배울 기회를 갖게 될 것이다.