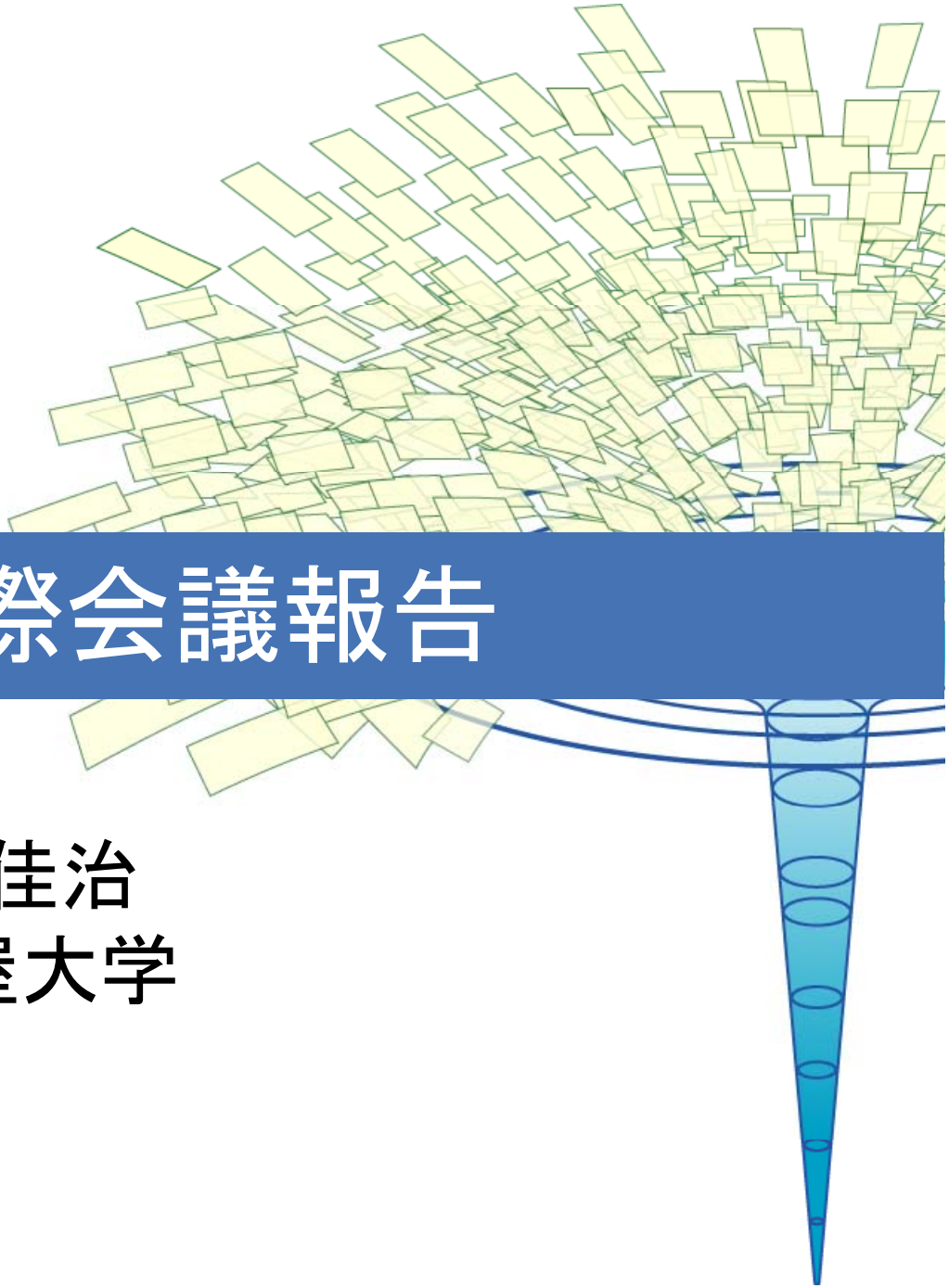


ICDE 2009国際会議報告

石川佳治
名古屋大学

目次

- ICDE 2009について
- 全体の傾向
- クラウドコンピューティング
- セキュリティ・プライバシー
- スカイライン問合せ
- トップk問合せ・ランク集約
- キーワードに基づく問合せ・マイニング
- Most Cited Papers in ICDE 2009

ICDE 2009について

- ICDEとは
  - International Conference on Data Engineering
  - IEEE主催
  - ACM SIGMOD, VLDBに並ぶデータベース分野の主要会議
  - 2005年には東京で開催
- ICDE 2009
  - 2009年3月29日～4月2日，上海にて
  - http://i.cs.hku.hk/icde2009/
- 先端的データベースとWeb技術動向講演会（ACM SIGMOD日本支部大会）
  - 6月13日，東工大にて
  - ICDE2009の報告あり

基調講演

- *Search Computing*
  Stefano Ceri (Politecnico di Milano, Italy)
  - サーチコンピューティングについて
- *Why can't I find my data the way I find my dinner?*
  David Carlson (International Polar Year Program Office, UK)
  - 国際極年プロジェクトについて
- *Data Management in the Cloud*
  Raghu Ramakrishnan (Yahoo!)
  - Yahoo!におけるクラウド研究開発

ワークショップ

- DBRank: Ranking in Databases
  - ランク付け問合せ
- WISS: Workshop on Information & Software Services
  - 情報・ソフトウェアサービス
- SMDB: Self-Managing Database Systems
  - データベースの自律的管理
- MOUND: Management and Mining of Uncertain Data
  - 不確実なデータの管理・マイニング
- M3SN: Modeling, Managing, and Mining of Evolving Social Networks
  - ソーシャルネットワーク

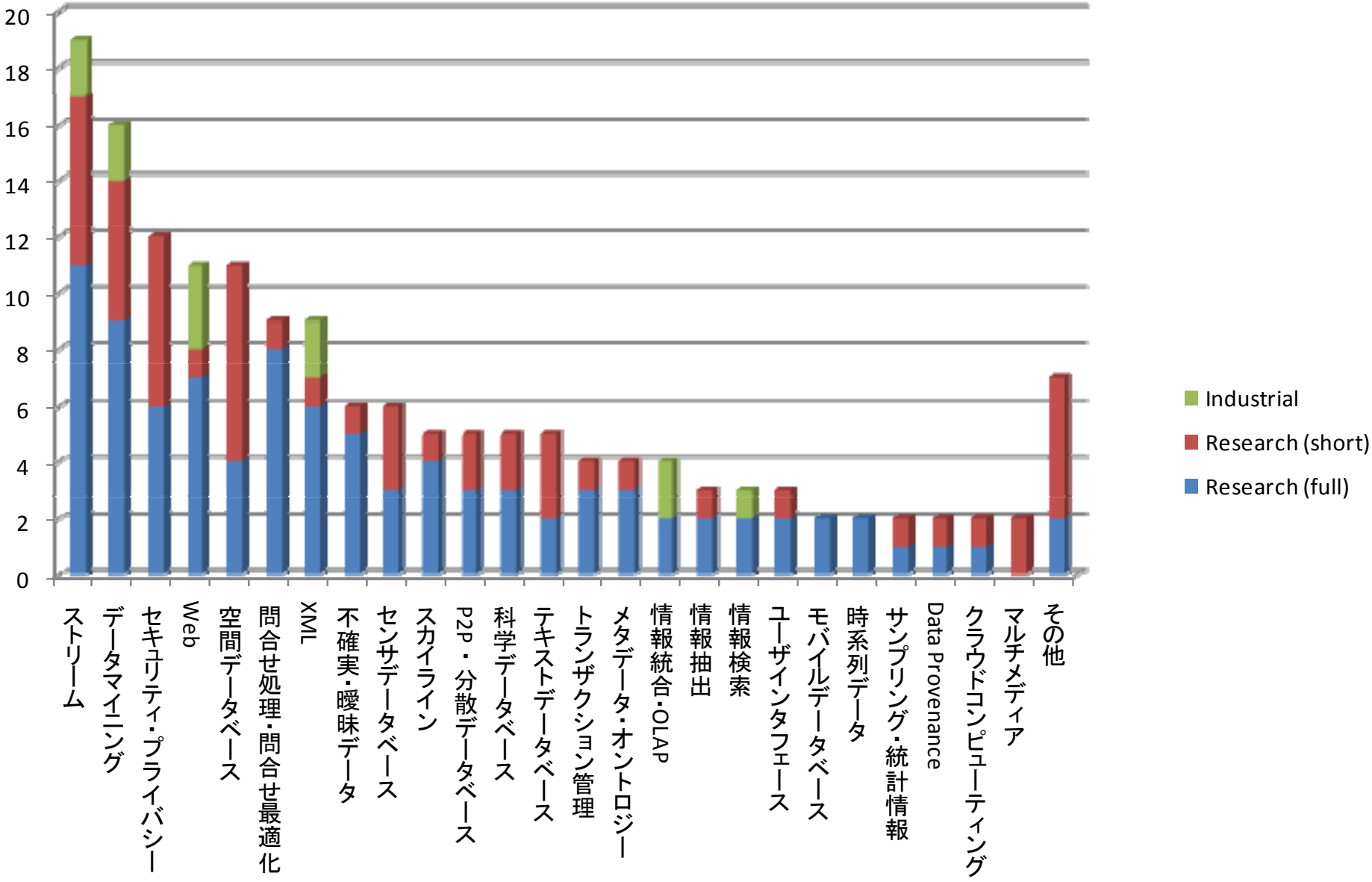
Areas in Research Track

- 14のエリアを設定
  - データの近似，不確実データ，確率的DB
  - 情報統合，メタデータ管理，相互可用性
  - データマイニング
  - プライバシーとセキュリティ
  - ストリーム，センサネットワーク
  - データウェアハウス，OLAP，データグリッド
  - ユーザインタフェース，情報可視化
  - 個人化，社会情報管理，注釈・キュレーション
  - 問合せ処理，問合せ最適化，チューニング
  - 科学DB，バイオインフォマティクス
  - トランザクション，ワークフロー，DBアーキテクチャ
  - ユビキタス，モバイル，分散，P2Pデータベース
  - Webデータ管理
  - XMLデータ管理

目次

- ICDE 2009について
- 全体の傾向
- クラウドコンピューティング
- セキュリティ・プライバシー
- スカイライン問合せ
- トップk問合せ・ランク集約
- キーワードに基づく問合せ・マイニング
- Most Cited Papers in ICDE 2009

トピックごとの論文数


20
18
16
14
12
10
8
6
4
2
0
ストリーム
データマイニング
セキュリティ・プライバシー
Web
空間データベース
問合せ処理・問合せ最適化
XML
不確実・曖昧データ
センサデータベース
スカイライン
P2P・分散データベース
科学データベース
テキストデータベース
トランザクション管理
メタデータ・オントロジー
情報統合・OLAP
情報抽出
情報検索
ユーザインタフェース
モバイルデータベース
時系列データ
サンプリング・統計情報
Data Provenance
クラウドコンピューティング
マルチメディア
その他
Industrial
Research (short)
Research (full)

全体の傾向

- 継続的にホットな話題
  - ストリーム処理，XML：各論へ（狭く深く）
  - データマイニング
    - グラフマイニング（リンク解析），データクリーニング，曖昧・匿名データからのマイニングなど
  - Web：Webコミュニティ，インターネット広告
- 最近の話題
  - セキュリティ・プライバシー
  - 不確実・曖昧データ
  - スカイライン問合せ
  - センサデータベース
  - Data Provenance / Lineage：データの来歴を管理：信頼性の確保
  - クラウドコンピューティング / DaaS (Database as a Service)

目次

- ICDE 2009について
- 全体の傾向
- クラウドコンピューティング
- セキュリティ・プライバシー
- スカイライン問合せ
- トップk問合せ・ランク集約
- キーワードに基づく問合せ・マイニング
- Most Cited Papers in ICDE 2009

クラウド環境での大規模データ管理

- 分散コンピューティングを基盤とするアプローチ
  - 大規模並列化によるスケーラビリティ
- キーベースのアクセス
  - Google: GFS, BigTable, MapReduce
  - Yahoo!: Hadoop & hBase
- SQLのサポート
  - SCOPE (Microsoft), Pig Latin (Yahoo!), Dryad (Microsoft), Clustra (Wisconsin U)
- トランザクションのサポート
  - PNUTS (Yahoo!), Database on S3 (ETH Zurich), Cassandra (Facebook)

提供するモデル・機能
一貫性の支援などに違い

K.S. Candan, W.-S. Li, T. Phan, M. Zhou,
Frontiers in Information and Software as Services
ICDE 2009 WISS Workshop

DaaS (Database as a Service)

- データベースをサービスとして提供
  - 例：MySQLをクラウドを介して利用
  - 通常のRDBMSとして見える：新たな学習は不要
- スケーラビリティをどう実現するか？
  - Multi-tenant Data Management
- 実現のアプローチ
  - Database Space：DBMSを共有，データは別々
  - タグ付け（Tagging）：タプルの所有者を表す属性を追加し，複数のテーブルを一つにまとめる
  - スキーマ統合：複数のスキーマを一つにマッピング [1]

[1] Aulbach, Grust, Jacobs, Kemper, Rittinger,
Multi-tenant Databases for Software as a Service:
Schema-Mapping Techniques, SIGMOD 2008.

ICDE 2009論文紹介：M-Store

- Hui, Jiang, Li, Zhou, *Supporting Database Applications as a Service*
- M-Store について：タグ付け方式を採用
- 分析
  - 一つのDBMSインスタンスで多数のテーブルを支援する方式はスケーラブルでない
    - 各テーブルごとの管理情報（主記憶上で管理）やバッファ領域が累積
    - 50,000テーブル程度が限界?
  - 複数のテーブルを一つの物理テーブルで実現する方式が有望

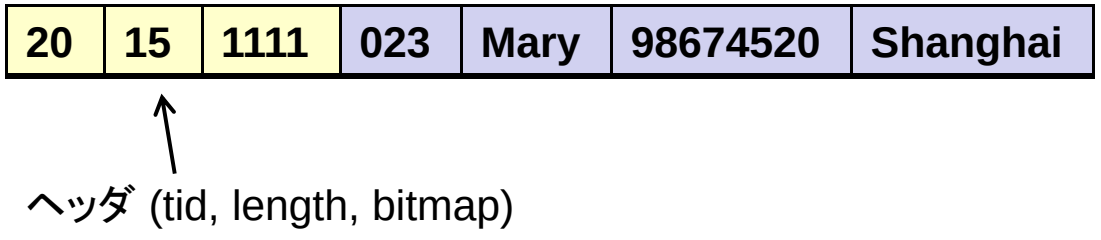
M-Storeのアイデア

- タグ付け方式を採用：類似したテーブルを統合

TID	Eno	Name	Age	Phone	Salary	Office
A	053	Jerry	35	NULL	NULL	NULL
A	089	Jacky	28	NULL	NULL	NULL
B	023	Mary	NULL	98674520	NULL	Shanghai
B	077	Ball	NULL	22753408	NULL	Singapore
C	131	Big	40	NULL	8000	London
C	088	Tom	36	NULL	6500	Tokyo

Sparse Wide Tableの格納構造はe-コマースのデータベースでも重要

  - 疎なテーブル
    - 単純な実装ではNULLの格納コスト大
- M-Storeのアプローチ
  - MySQLをベース
  - テナントのNULL属性を考慮したレコードのパッキング



目次

- ICDE 2009について
- 全体の傾向
- クラウドコンピューティング
- 不確実なデータ・確率的データベース
- セキュリティ・プライバシー
- スカイライン問合せ
- トップk問合せ・ランク集約
- キーワードに基づく問合せ・マイニング
- Most Cited Papers in ICDE 2009

不確実・曖昧なデータ（Uncertain/Imprecise Data）

- さまざまな要因で発生
  - センサデータ
  - 科学データベース：実験・観察の結果
  - Webなどからの情報収集の結果
  - 入力ミス，プライバシー，近似処理，…
- 例：Trio（Stanford大）[1]：不確実なデータを表現可能

(witness, car)	
(Amy, Honda) : 0.5 || (Amy, Toyota) : 0.3 || (Amy, Mazda) : 0.2	
(Betty, Acura) : 0.6	?

確信度（確率）

ORを指定

Maybe Tupleであることを表現

[1] Widom, Trio: *A System for Integrated Management of Data, Accuracy, and Linage*, CIDR 2005.

ICDE 2009の論文紹介

- Agrawal, Widom, *Confidence-Aware Join Algorithms*
  - 信頼度が伴うデータベースにおける結合問合せの処理手法
  - 信頼度によるトップ*k*問合せもしくは閾値ベースの問合せ
  - 信頼度の低いタプルは出力不要：新たな手法の必要性
- Cormode, Garofalakis, *Histograms and Wavelets on Probabilistic Data*
  - Best Paper Award
  - 確率が伴うデータ集合の要約法
  - 例：<**1**, 1/2>, <**2**, 1/3>, <**2**, 1/4>, <**3**, 1/2>

可能世界とその出現確率：これをヒストグラム or ウェーブレットで要約

W	∅	1	12	122	123	1223	13	2	22	23	223	3
Pr[W]	1/8	1/8	5/48	1/48	5/48	1/48	1/8	5/48	1/48	5/48	1/48	1/8

目次

- ICDE 2009について
- 全体の傾向
- クラウドコンピューティング
- 不確実なデータ・確率的データベース
- セキュリティ・プライバシー
- スカイライン問合せ
- トップk問合せ・ランク集約
- キーワードに基づく問合せ・マイニング
- Most Cited Papers in ICDE 2009

セキュリティ・プライバシーの研究

- 主な研究課題
  - プライバシー保護技術
  - データベース監査（Auditing）：ログをもとに不正な動作を監視・記録
  - アクセス権管理
- 例：プライバシーを保護したデータ出版（Data Publishing）

年齢	性別	ZIP	病名
23	男	11000	肺炎
27	男	13000	消化不良
35	男	59000	消化不良
59	男	12000	肺炎
61	女	54000	かぜ
65	女	25000	胃炎
65	女	25000	かぜ
70	女	30000	気管支炎

汎化による匿名化

ℓ-多様性の例 $(\ell = 2)$：
Bobの個人情報を知っていても高々 $1/\ell$ の確率でしか病名が特定できない

年齢	性別	ZIP	病名
[21,60]	男	[10001,60000]	肺炎
[21,60]	男	[10001,60000]	消化不良
[21,60]	男	[10001,60000]	消化不良
[21,60]	男	[10001,60000]	肺炎
[61,70]	女	[10001,60000]	かぜ
[61,70]	女	[10001,60000]	胃炎
[61,70]	女	[10001,60000]	かぜ
[61,70]	女	[10001,60000]	気管支炎

相関ルールからのプライバシー情報の導出（1）

- Zhu, Wang, Du, *Deriving Private Information from Association Rule Mining Results*
- 相関ルール集合（minsup = 0.3, minconf = 0.8)
  - 学歴 = 博士 ⇒ 給料 = 50K+ (conf = 0.83, sup = 0.42)
  - 学歴 = 博士 ∧ 性別 = 女性 ⇒ 給料 = 50K+ (conf = 1, sup = 0.33)
  - 性別 = 女性 ⇒ 給料 = 50K+ (conf = 0.89, sup = 0.67)
- 何が導出できるか
  - 博士号をもつ人の83%は50K以上の収入
  - 女性で博士号をもつと100%で50K以上の収入
  - 博士号をもった男性の中に低収入の人がいる
  - …

相関ルールからのプライバシー情報の導出(2)

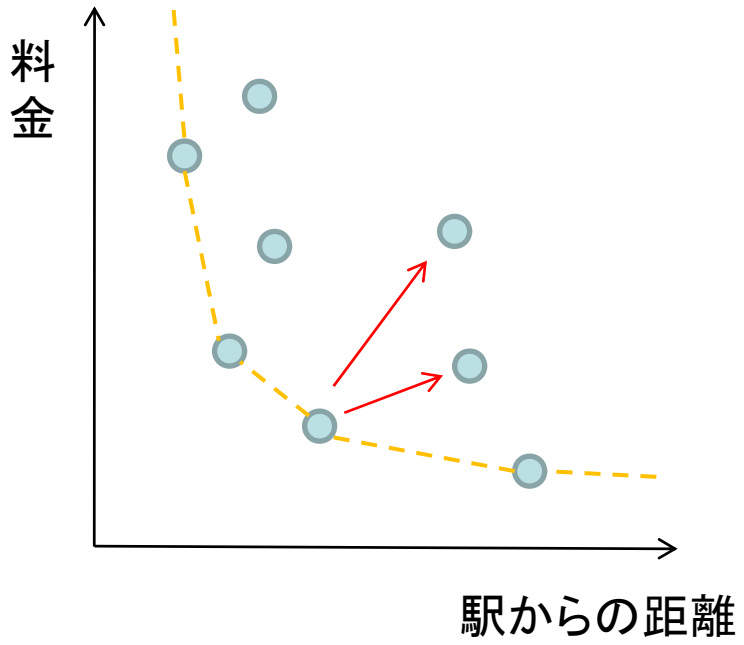
- 逆頻出アイテム集合マイニング(inverse frequent itemset mining)の一種
- 相関ルールの集合から，プライバシー属性に関する確率を推定
  - 例：Pr(給料 = 50K+ | 学歴 = 博士 ∧ 性別 = 男性)
- アプローチ
  - 相関ルールを組み合わせる
  - 存在しない相関ルールも考慮
  - 相関ルール集合だけでは元データを完全に復元できるわけではないので，さまざまな可能性が存在
  - 最大エントロピー法を用いて，相関ルールの制約のもとで妥当な確率を導出

目次

- ICDE 2009について
- 全体の傾向
- クラウドコンピューティング
- 不確実なデータ・確率的データベース
- セキュリティ・プライバシー
- スカイライン問合せ
- トップk問合せ・ランク集約
- キーワードに基づく問合せ・マイニング
- Most Cited Papers in ICDE 2009

スカイライン問合せ

- 例：宿泊ホテルの候補を探す



  - スカイライン：他のどのオブジェクトにも支配されないオブジェクトの集合

- スカイライン問合せ
  - Kossmann らが提案 [1]
  - 一般の設定に関しては，空間索引を利用するBBSアルゴリズム [2] が高速
- さまざまな状況・対象について拡張
  - 複数の判断基準がある際の意思決定などに活用
  - 半順序の支配関係の利用がキーポイント

[1] Börzsönyi, Kossmann, Stocker: *The Skyline Operator*, ICDE 2001.
[2] Papadias, Tao, Fu, Seeger: *An Optimal and Progressive Algorithm for Skyline Queries*, SIGMOD 2003.

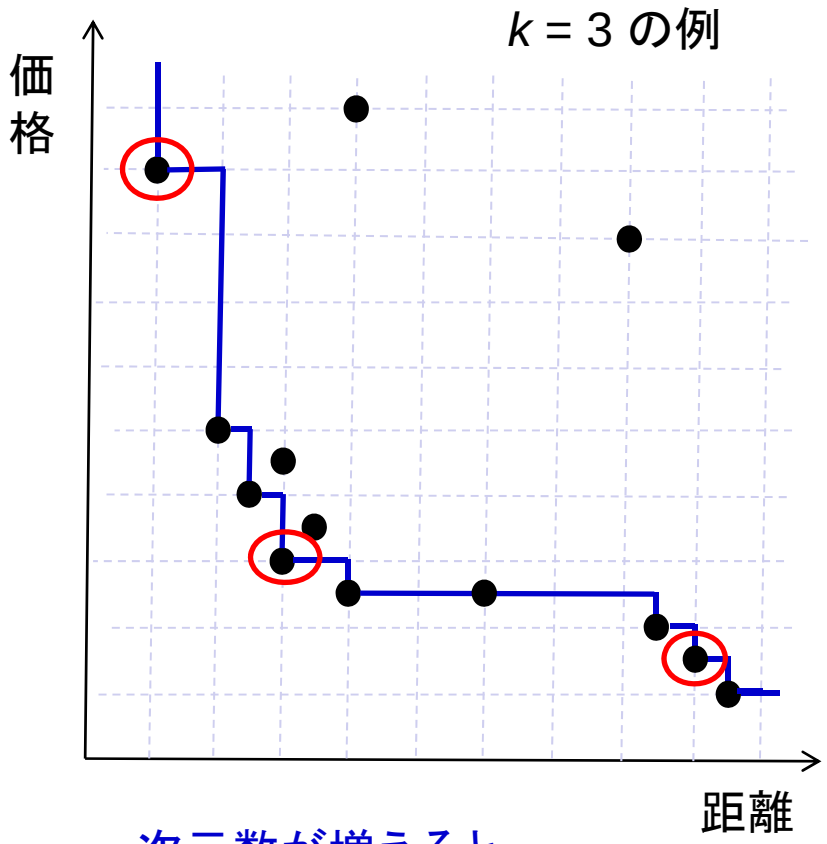
ICDE 2009論文：代表的スカイライン

- Tao, Ding, Lin, Pei, *Distance-based Representative Skyline*
- k 個の代表的なスカイライン オブジェクトを求める
  - スカイライン全体を求める 必要はない
- 基準：任意の代表スカイライン点から非代表スカイライン点への最大距離

$$Er(K, S) = \max_{p \in S-K} \{ \min_{p' \in K} \| p, p' \| \}$$

を最小化

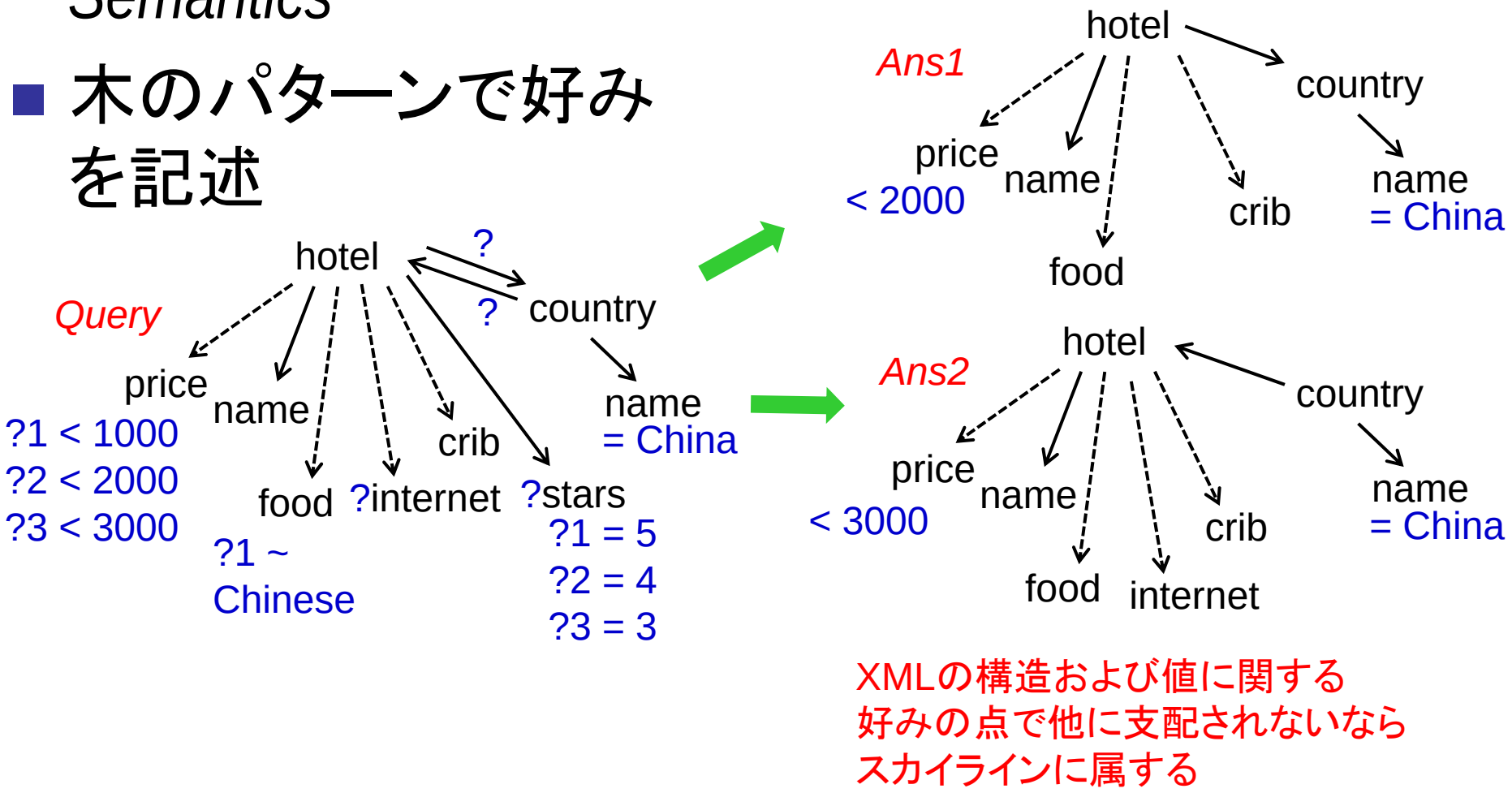
  - S：スカイライン全体集合
  - K：k 個の代表



次元数が増えると
スカイライン点が
急速に増えるので
代表点の考え方は有効

XMLにおけるスカイライン問合せ

- Cohen, Shiloach, *Flexible XML Querying using Skyline Semantics*
- 木のパターンで好みを記述



XMLの構造および値に関する
好みの点で他に支配されないなら
スカイラインに属する

目次

- ICDE 2009について
- 全体の傾向
- クラウドコンピューティング
- 不確実なデータ・確率的データベース
- セキュリティ・プライバシー
- スカイライン問合せ
- トップk問合せ・ランク集約
- キーワードに基づく問合せ・マイニング
- Most Cited Papers in ICDE 2009

トップ*k*問合せ・ランク集約

- トップ*k*問合せの動機
  - ユーザは必ずしもすべてのデータを必要としない
  - ランク付けして上位*k*件を提示
  - スカイライン問合せとも共通する動機
- 多くの研究が存在 [1]

[1] Ilyas, Baskales, Soliman, *A Survey of Top-k Query Processing Techniques in Relational Databases*, ACM Computing Surveys, 40(4), 2008.
[2] Fagin, Lotem, Naor, *Optimal Aggregation Algorithms for Middleware*, PODS 2001 / JCSS 2003.

- ランク集約（Rank Aggregation）
  - 複数システムからのランク付け問合せ結果を集約
  - FaginによるTAアルゴリズム [2]が有名
- ICDE2009の論文：Soliman, Ilyas, *Ranking with Uncertain Scores*
  - スコアに曖昧性があるデータ（例：レストランAの満足度は [5-8]）に基づくランク付け

目次

- ICDE 2009について
- 全体の傾向
- クラウドコンピューティング
- 不確実なデータ・確率的データベース
- セキュリティ・プライバシー
- スカイライン問合せ
- トップk問合せ・ランク集約
- キーワードに基づく問合せ・マイニング
- Most Cited Papers in ICDE 2009

データベース上のキーワード問合せ

- 例: 問合せ {maxtor, netvista}
- キーワードを含むタプルを外部キーで結合しランク付け

Hristidis, Gravano, Papakonstantinou, *Efficient IR-style Keyword Search over Relational Databases*, ICDE 2002.

Complaints

rid	prodId	custId	date	comments
c1	P121	c3232	6-30-2002	disk crashed after just one week of moderate use on an IBM Netvista X41
c2	P131	c3131	7-3-2002	lower-end IBM Netvista caught fire, starting apparently with disk
c3	p131	c3143	8-3-2002	IBM Netvista unstable with Maxtor HD

Products

rid	prodId	manufacturer	model
p1	p121	Maxtor	D540X
p2	p131	IBM	Netvista
p3	p141	Tripplite	Smart 700VA

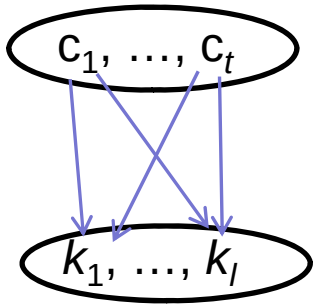
Customers

rid	custId	name	occupation
u1	c3232	John Smith	Software Engineer
u2	c3131	Jack Lucas	Architect
u3	c3143	John Meyer	Student

ICDE 2009論文：キーワードによるRDB中のコミュニティ検索

- Qin, Yu, Chang, Tao, *Querying Communities in Relational Databases*
- RDBをグラフとみる
  - ノード = タプル（テキスト属性含む）
  - 辺：外部キーによる参照．
    辺には重みが存在

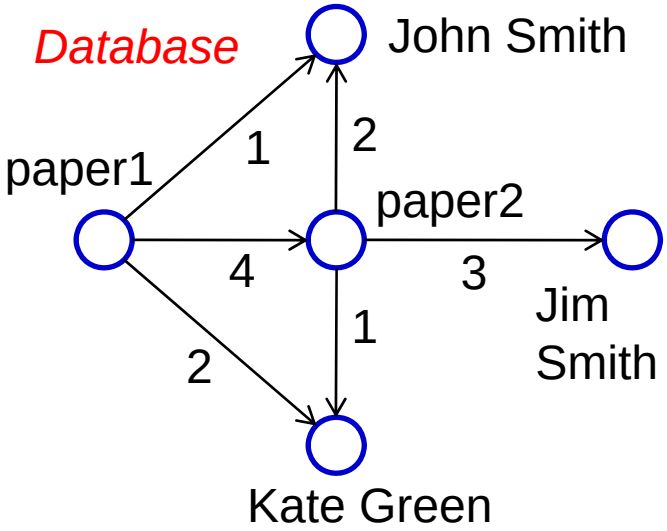
コミュニティ構造



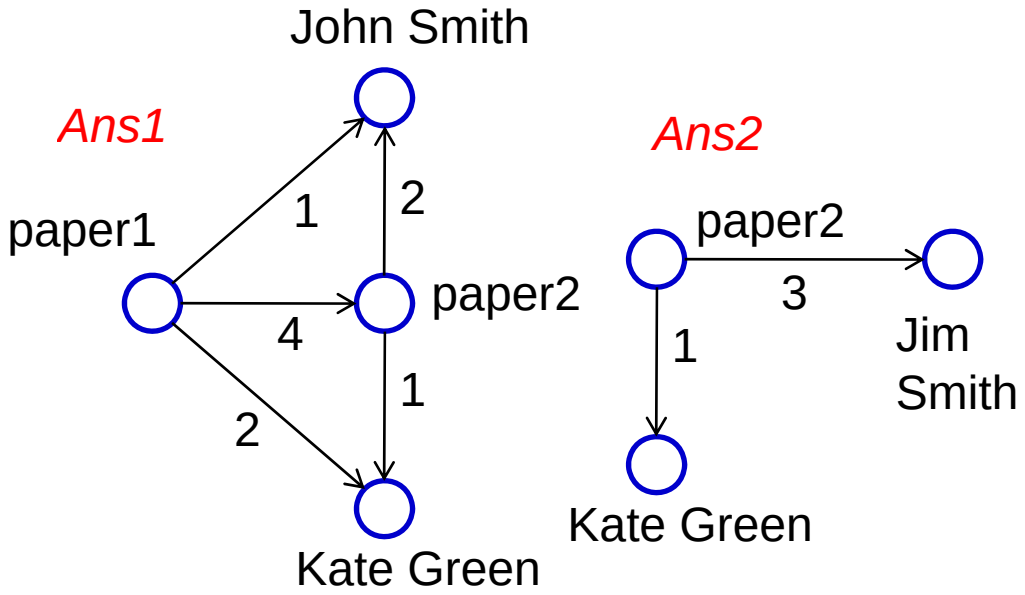
Center nodes

Path nodes

Keyword nodes



Query: {Kate, Smith}



目次

- ICDE 2009について
- 全体の傾向
- クラウドコンピューティング
- 不確実なデータ・確率的データベース
- セキュリティ・プライバシー
- スカイライン問合せ
- トップk問合せ・ランク集約
- キーワードに基づく問合せ・マイニング
- Most Cited Papers in ICDE 2009

Most Cited Papers in ICDE 2009 (1)

- Fagin, Lotem, Naor, *Optimal Aggregation Algorithms for Middleware*, PODS 2001 / JCSS 2003. [15] ランク集約
- He, Wang, Yang, Yu, *BLINKS: Ranked Keyword Search on Graphs*, SIGMOD 2007. [10] キーワードサーチ
- Börzsönyi, Kossmann, Stocker, *The Skyline Operator*, ICDE 2001. [9] スカイライン問合せ
- Bhalotia, Hulgeri, Nakhe, Chakrabarti, Sudarshan, *Keyword Searching and Browsing in Databases Using BANKS*, ICDE 2002. [8] キーワードサーチ
- Hristidis, Papakonstantinou, *DISCOVER: Keyword Search in Relational Databases*, VLDB 2002. [8] キーワードサーチ

Most Cited Papers in ICDE 2009 (2)

- Dalvi, Suciu, *Efficient Query Evaluation on Probabilistic Databases*, VLDB 2004 / VLDBJ 2007. [8] 不確実・曖昧DB
- Li, Ooi, Feng, Wang, Zhou, EASE: *3-in-1 Keyword Search Method for Unstructured, Semi-Structured Data*, SIGMOD 2008. [7] キーワードサーチ
- Liu, Yu, Meng, Chowdhury, *Effective Keyword Search in Relational Databases*, SIGMOD 2006. [7] キーワードサーチ
- Madden, Franklin, Hellerstein, Hong, *TAG: A Tiny Aggregation Service for Ad-hoc Sensor Networks*, OSDI 2002. [7] センサネットワーク
- 引用の総数ではスカイライン問合せ，不確実・曖昧データ，セキュリティ・プライバシー，ストリーム処理などが上位