ウェブインテリジェンスを支えるXML(再)入門

XML情報検索(再)入門

同志社大学 文化情報学部

波多野 賢治

本講演の内容

- XML情報検索の起こり
  - XML文書の型
  - データベースと情報検索
- XML情報検索に求められていること
- 第一世代XML情報検索
- 第二世代XML情報検索
- XML情報検索の今後の動向
  - データベース業界
  - 情報検索 業界
    - INEXプロジェクト

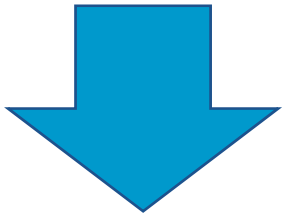
XML情報検索の起こり

- データの爆発的な増加
  - さまざまなデータに対する検索の要求が高まる
    - さまざまなアプリケーション
    - さまざまな使用環境

- データの種類，アプリケーションに応じて検索に関する研究が行われた
  - データベース業界
  - 情報検索業界
  - 人工知能業界
  - 自然言語処理業界 etc.

XML情報検索の起こり

- XMLデータの爆発的な増加
  - XMLデータに対する検索の要求が高まる
    - さまざまなアプリケーション
    - さまざまな使用環境
  - データフォーマットはXML一種類



単なるメタ言語であったXMLが
さまざまな型を持ったデータに変貌

XML情報検索の起こり

- XML文書の型

データ指向型XML文書

- 値，レコード
- 構造＝値の属性

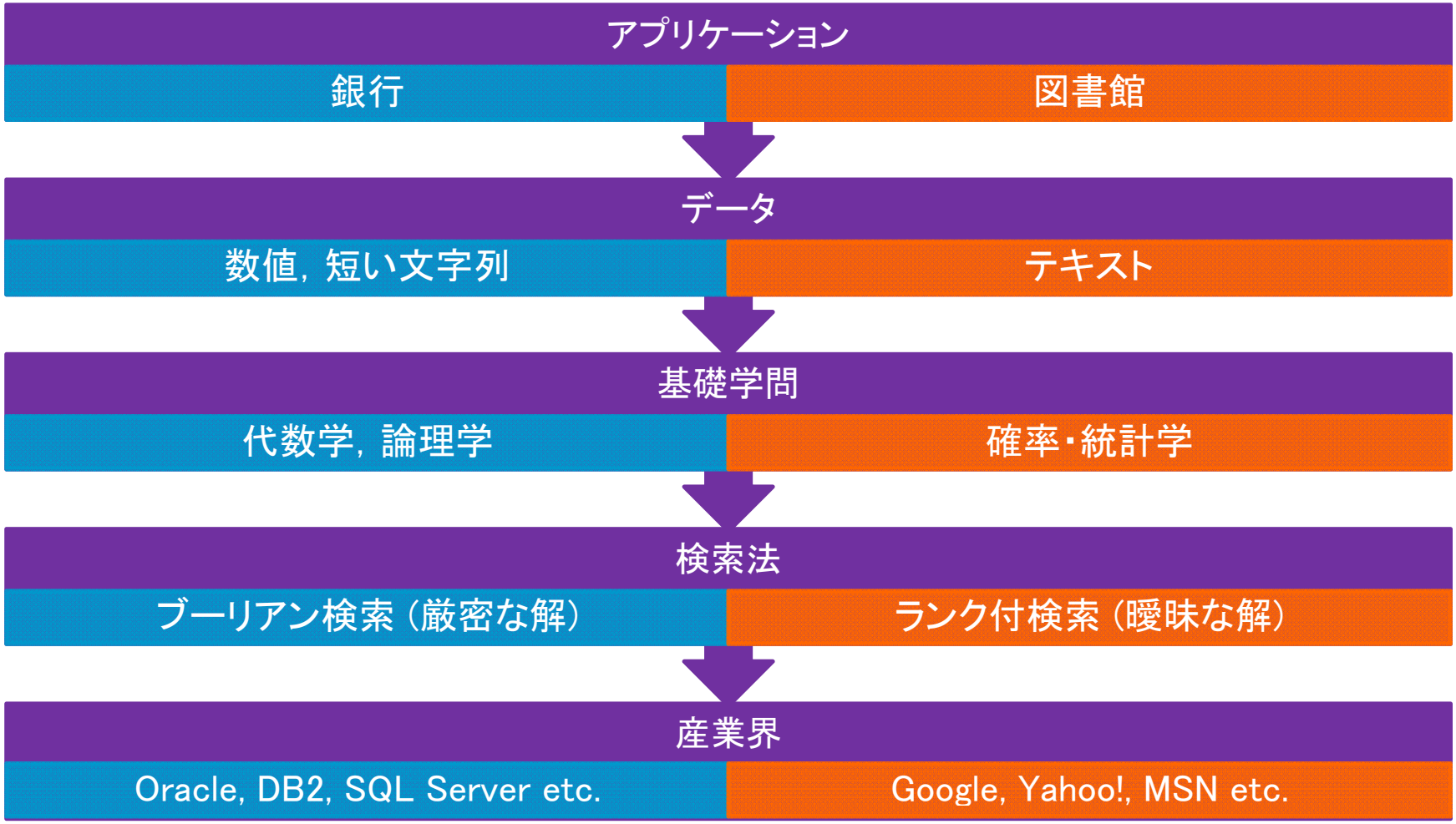
データベース

文書指向型XML文書

- 文章（部分）
- 構造＝文書の粒度

情報検索

DB & IR


アプリケーション
銀行
図書館
データ
数値，短い文字列
テキスト
基礎学問
代数学，論理学
確率・統計学
検索法
ブーリアン検索（厳密な解）
ランク付検索（曖昧な解）
産業界
Oracle, DB2, SQL Server etc.
Google, Yahoo!, MSN etc.

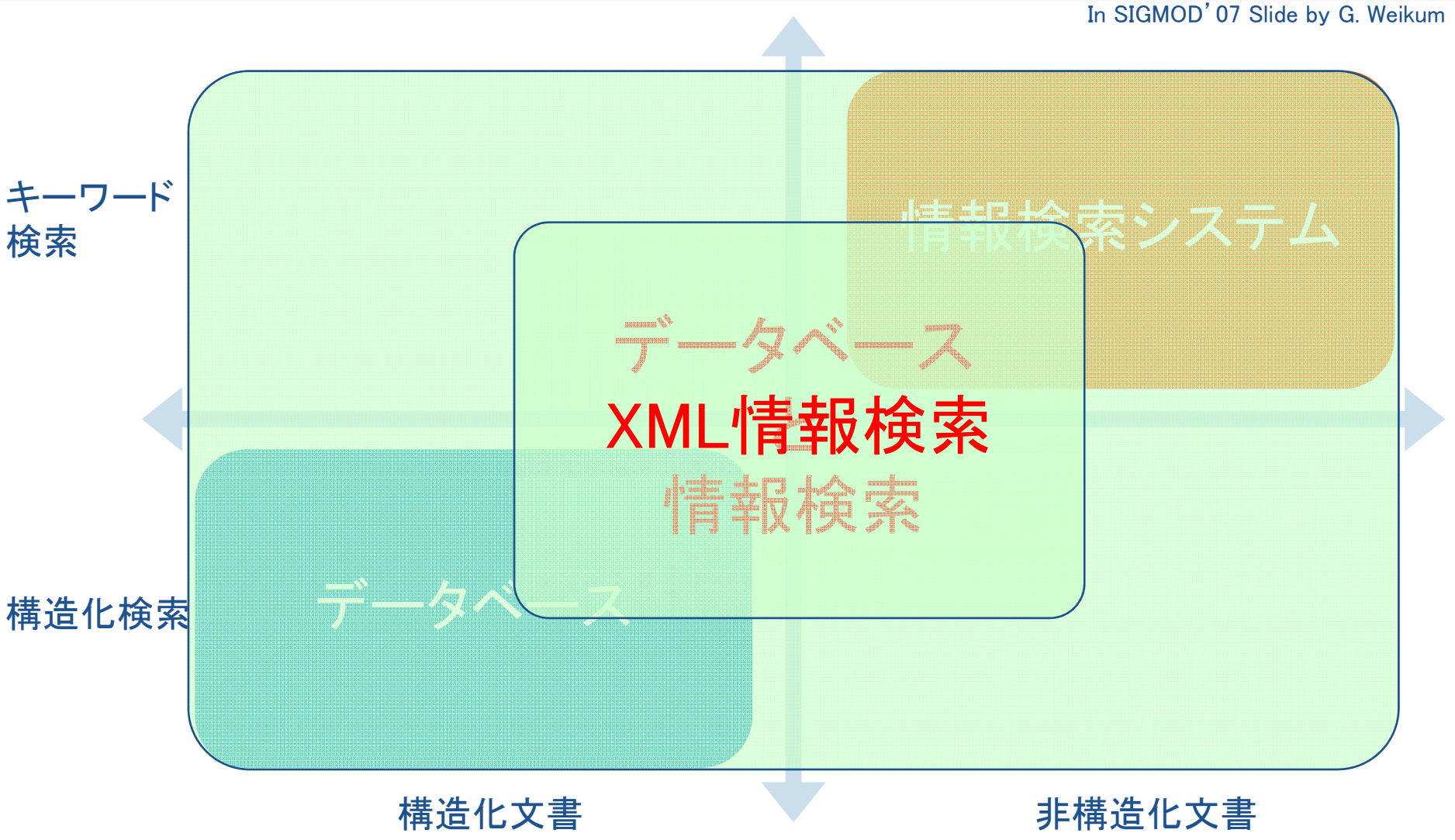
DB & IR

- データベース業界と情報検索業界
  - 歴史的な理由から別々に発展
    - 研究対象の違い
    - 方法論，枠組みの違い
  - それらの統合に関する議論
    - ACM SIGMOD’05 パネル
    - ACM SIGMOD’07 キーノート
    - ACM SIGMOD Record Vol.37, No.3, Sep. 2008

XML (Web 2.0) を素材として
二つの業界を統合できないか?!

XML情報検索の起こり

In SIGMOD'07 Slide by G. Weikum

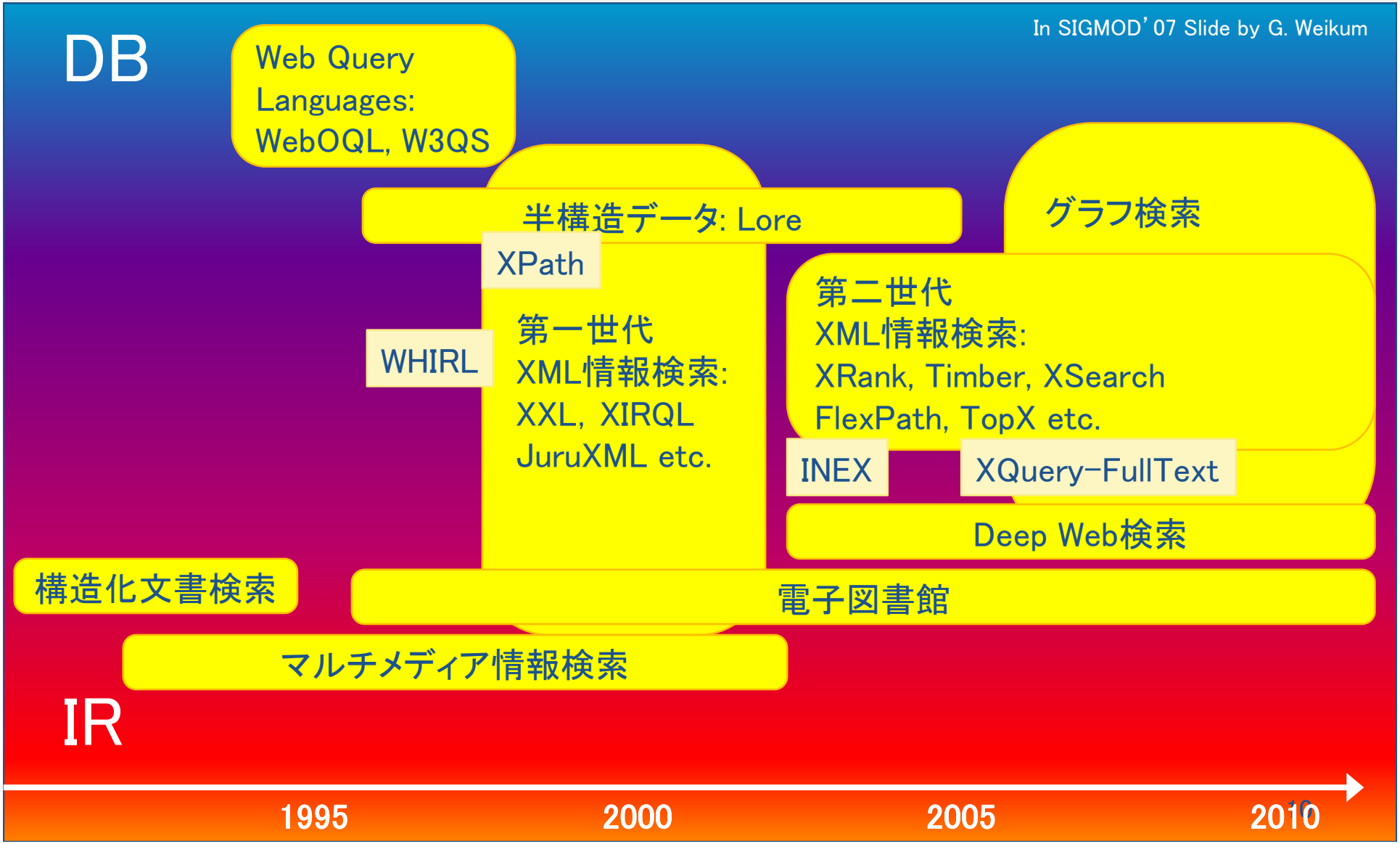


XML情報検索で求められていること

- 柔軟なランキングアルゴリズム
  - 膨大／ゼロな検索結果への対応
- 知識処理
  - オントロジによる問合せの書き換え (relaxation etc.)
- キーワードと文書構造を利用した複雑な問合せ
  - XPath/XQuery Full-Text Query with ranking [W3C, 2008]
- 処理の高速性
  - 高負荷時の処理や更新処理

[W3C, 2008] World Wide Web Consortium, “Xquery and Xpath Full Text 1.0, W3C Candidate Recommendation, 2008.

XML情報検索の関連研究



第一世代の研究

- WHIRL: IR over Relations [Cohen, *SIGMOD'98*]
  - テキストデータの類似度計算を関係代数に付加

Movies

タイトル	プロット	上映年度
Matrix	Computer hacker Neo … fight training …	1999
Hiro	…fights Broken Sward…	2002
Shrek 2	…lovely hero fights with cat killer …	2004

Reviews

タイトル	コメント	評価
Matrix	cool fights .. new techniques ...	4
Matrix Reloaded	more fights … … fairly boring	1
Matrix Eigenvalues	.. matrix spectrum ..	5
Hiro	... fight for peace ...	5

```
SELECT * FROM Movies M, Reviews R
WHERE M.プロット ～ "fight" AND M.上映年度 > 1990 AND R.評価 > 3
      AND M.タイトル ～ R.タイトル AND M.プロット ～ R.コメント
```

[Cohen, SIGMOD'98] W. W. Cohen, "Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity", pp.201-212, 1998.

第一世代の研究

- WHIRL: IR over Relations
  - “～” が類似計算のための演算
    - M.タイトル ～ R.タイトル: タイトル間のコサイン尺度
    - コサイン尺度の基準はTF-IDF索引語重み付け
  - 問合せ時におけるデータベースと情報検索のコラボレーション
  - 類似計算のためのモデルに難あり

第一世代の研究

[Theobald et al., WebDB'00] A. Theobald, G. Weikum, "Adding Relevance to XML", WebDB'00, pp.105-124, 2000.

- XXL [Theobald et al., *WebDB'00*]
  - テキストデータと XPath の類似度計算法の提案

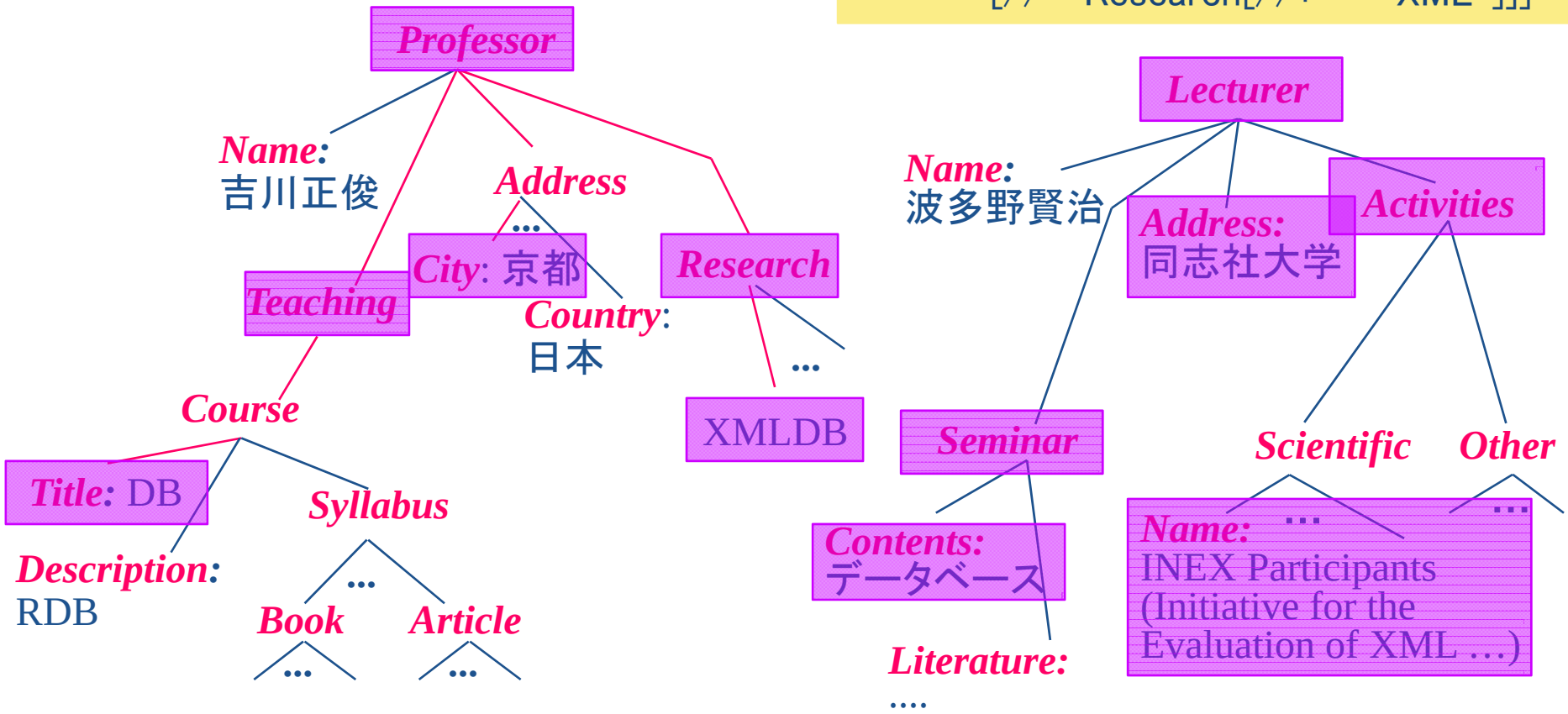
Query:
京都でデータベースの講義を持っていて
XMLの研究をしている教授は?

Query:
//Professor[[//*="京都"]
[//Course[//*="DB"]]
[//Research[//*="XML"]]]

第一世代の研究

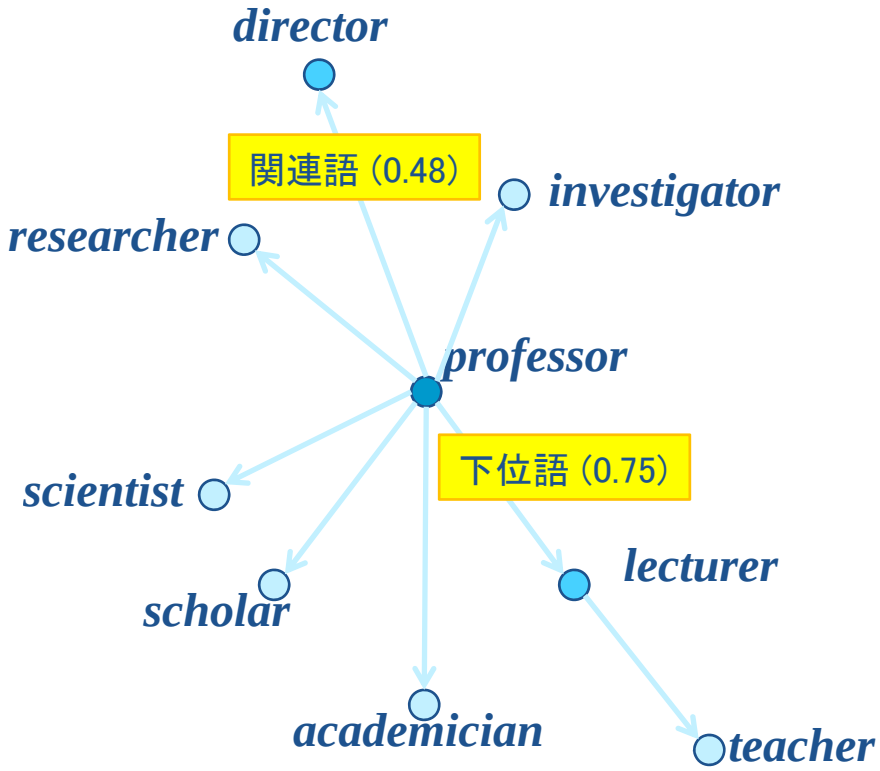
- XXL
  - WHIRLのように“～”を利用

Query:
//～Professor[[//*=“～京都”]
[//～Course[//*=“～DB”]]
[//～Research[//*=“～XML”]]]



第一世代の研究

- XXL
  - TF-IDF索引語重み付け
  - オントロジを利用してタグ名の類似度検索
  - XMLを利用した最初のデータベースと情報検索のコラボレーション



第一世代の研究のまとめ

- 基本的にはデータベースにスコアリング（ランキング）の機能を追加
  - データベースインスタンス／スキーマの多様性に対応
  - 検索精度は低
    - ad-hoc 的なスコアリング
    - スコアリングアルゴリズムの定式化は高精度化のためにも重要
- オントロジの利用は意外に有用

第二世代の研究

- TopX [Theobald et al., *VLDB'05*]
  - 第一世代の成果を取り入れる
    - タグや文書構造の利用による精度向上
    - オントロジ利用によるタグ名/文書構造の多様性に対応することによる精度向上
    - Okapi BM25 による定式化されたランキング
    - 効率的な top-k クエリ処理
    - 自動クエリ書き換え機能によるクエリ拡張

[Theobald et al., VLDB'05] M. Theobald, R. Schenkel, G. Weikum, "An Efficient and Versatile Query Engine for TopX Search", VLDB'05, pp.625-639, 2005.

第二世代の研究

[Robertson et al., JASIS'76] S.E. Robertson, K. Sparck Jones, "Relevance Weighting of Search Terms", Journal of the American Society for Information Science, Vol.27, No.3, pp.129-146, 1976.

- TopX
  - Okapi BM25 [Robertson et al., *JASIS'76*]
    - 検索キーワードに重みを与えて確率的にスコアリングし検索結果に順位付けを与える方法
      - 文書dが検索質問に適合する確率と適合しない確率の比
      - 検索質問に適合する文書に索引語t_iが付与されている確率と検索質問に適合しない文書に索引語t_iが付与されている確率

$$S(d,q) = \frac{P[R \mid d]}{P[\overline{R} \mid d]} \approx \sum \log \frac{p_i}{\overline{p_i}} + \sum \log \frac{1-p_i}{1-\overline{p_i}}$$

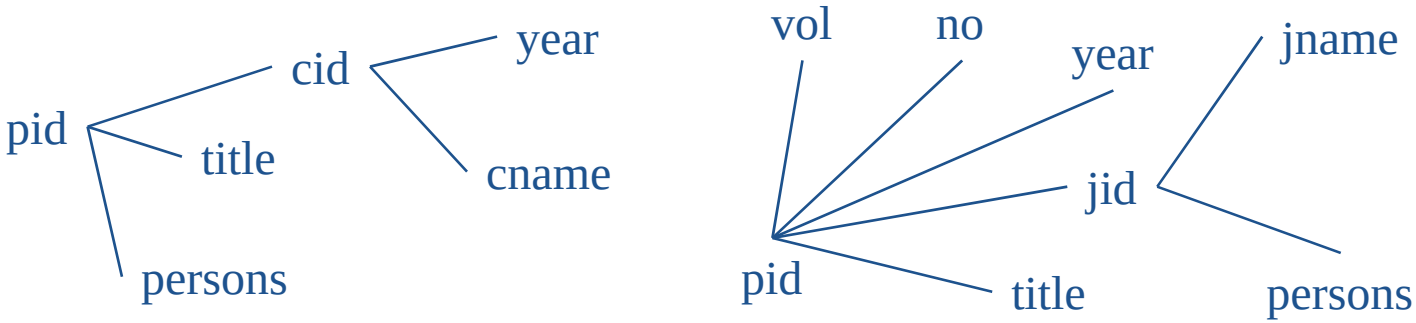
検索質問にも文書中にも出現する索引語の重み

検索質問には出現するが文書中には出現しない索引語の重み

第二世代の研究

- DBXplorer, BANKS, Discover, etc. *[ICDE'02 以降多数]*
  - 複数の表，木をまたいだ検索
  - 例:
    会議 (cid, cname, year)
    会議論文 (pid, title, cid)
    著者 (pid, persons)
    論文誌 (jid, jname)
    雑誌論文 (pid, title, jid, vol, no, year)
    編者 (jid, persons)

```
SELECT * FROM *
WHERE * CONTAINS "国島 天笠 波多野 坂本 的野 XML"
        AND YEAR > 2005
```



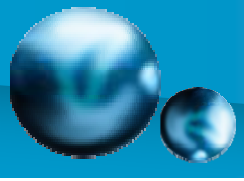
[Agrawal et al., ICDE'02] S. Agrawal, S. Chaudhuri, G. Das, "DBXplorer: A System for Keyword-Based Search over Relational Databases", ICDE'02, pp.5-16, 2002.
[Bhalotia et al., ICDE'02] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, S. Sudarshan, "Keyword Searching and Browsing In Databases using BANKS", ICDE'02, pp.431-440, 2002.

第二世代の研究

- DBXplorer, BANKS, Discover, etc.
  - 複数の表，木をまたいだ検索
    - グラフデータからの検索（グラフ検索）
      - 検索結果はクエリキーワードを含んだノードを連結して構成されたグラフ
      - グラフのランキング
        - » ノードスコア: TF-IDFなど
        - » エッジスコア: ノード間の関係の強さ
  - 効率性を考慮（top-k クエリ処理）
    - 最も効率の良い木の構成法は?
      - 例）スタイナー木の構築

多くの応用分野
- XML
- RDF グラフ
- ER グラフ etc.

[Hristidis et al., VLDB'02] V. Hristidis, Y. Papakonstantinou, "DISCOVER: Keyword Search in Relational Databases", VLDB'02, pp.670-681, 2002.
[Wang et al., VLDB'06] S. Wang, Z. Peng, J. Zhang, L. Qin, S. Wang, J.X. Yu, B. Ding, "NUITS: A Novel User Interface for Efficient Keyword Search over Databases", VLDB'06, pp.1143-1146, 2006.



第二世代の研究のまとめ

- ノードスコア: テキストデータの確率的スコアリング
  - スコアリングの定式化
  - 高精度な検索結果
- エッジスコア: 文書構造のスコアリング
  - Tree edit distance をベース (FlexPath, Timber etc.)
- XML文書のスコアリング
  - テキストデータの確率的スコアリング
  - 文書構造スコアリング
  - オントロジの利用

[Amer-Yahia et al., SIGMOD'04] S. Amer-Yahia, L.V.S. Lakshmanan, S. Pandit, "FleXPath: Flexible Structure and Full-Text Querying for XML", SIGMOD'04, pp.83-94, 2004.
[Jagadish et al., VLDB J.] H.V. Jagadish, S. Al-Khalifa, A. Chapman, L.V.S. Lakshmanan, A. Nierman, S. Paparizos, J.M. Patel D. Srivastava, N. Wiwatwattana, Y. Wu, C. Yu, "TIMBER: A Native XML Database", The VLDB Journal, Vol.11, No.4, pp.274-291, 2002

第二世代の研究のまとめ

- さらに検索精度を上げるためには．．．
  - ユーザからのフィードバックによるクエリ生成が重要
    - ユーザの入力クエリや検索結果へのアクセス履歴を分析

例）“life scientist Kyoto University”

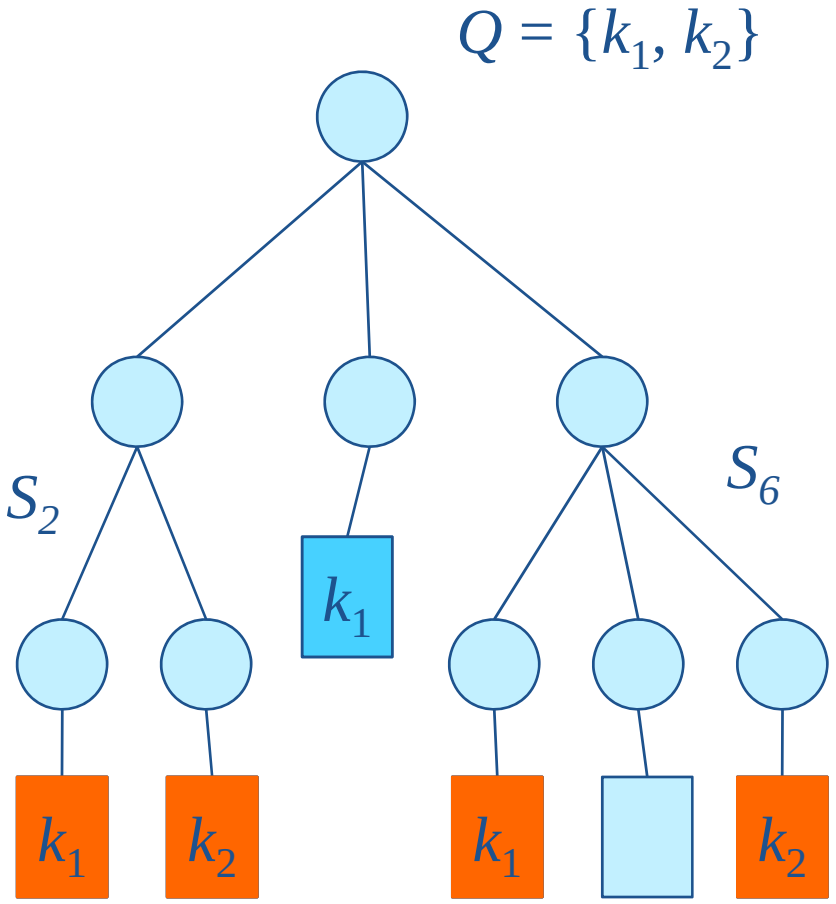
//article[[ftcontains(//person, “Kyoto University”)]
[ftcontains(//category, “scientist”)]]//biography

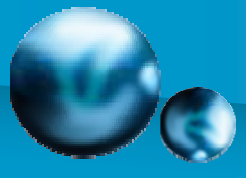
情報抽出とその利用にも関連

依然，効率的処理に関する多くの問題が残る

現在のXML情報検索

- ユーザによってクエリが与えられる
- クエリキーワードを含むノードを検索する
- 検索されたノードを含む最小部分グラフを検出する
- 最小部分グラフに対しスコアを計算する
- スコア付最小部分グラフを検索結果を返す





INEX

- 2002年春より続いているXML情報検索システム精度評価のためのデータセット，評価基準，ツールなどを提供するプロジェクト
  - 現在は150チーム，計300名ほどが参加
    - 日本からは4チームが参加
  - さまざまなタスクが用意
    - ad-hoc, book, efficiency, interactive, QA mining etc.
  - 提供されているデータセット，クエリ，評価データ
    - IEEE CS コーパス (2002-2005)
    - Wikipedia XML コーパス (2006-present)

INEXの動向（ad-hoc task）

- 検索結果となる最小部分グラフ
  - スニペットのようなもの
  - クエリに適切かどうかの判断は可能かもしれないが，最適な結果ではない
  - 文書から最適な部分を抽出する際の人間の行動を考慮する必要があるのではないか?
    - 人間による検索行動はそもそも曖昧なもの
    - 定式化がふさわしいかどうかは疑問
      - 同じ検索行動でも，欲する情報はユーザによって異なる

検索結果としてクエリキーワードを含んだ
最小部分グラフが返ってくることは本当に適切?

INEXの動向 (ad-hoc task)

- タスクの目的
  - 構造クエリが役に立つかの調査
    - キーワードによるクエリで十分でないか?
  - XML要素による検索に問題がないかの調査
    - パッセージ検索では不十分なのか?
- 方法
  - サブタスクを作成し，サブタスク間で比較実験を行う
    - Thorough task
    - Focused task
    - Relevant in context
    - Best in context

INEXの動向 (ad-hoc task)

- Focused task
  - 検索結果は部分グラフ／パッセージ
    - ただし検索結果同士のオーバーラップはなし
  - 検索結果上位の検索性能の評価　＝最小部分グラフ
- Relevant in context task
  - 検索結果は文書ごとにグループ化された部分グラフ／パッセージ
    - 一文書あたり一つの部分グラフ／パッセージが解
  - 検索結果全体の検索性能の評価　≠最小部分グラフ
- Best entry point task
  - 検索結果はユーザが読み始めるのに相応しい位置
    - 開始タグの位置，パッセージの開始点
  - 検索結果全体の検索性能の評価

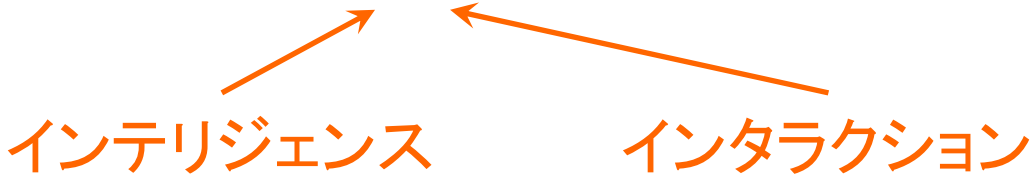
人間の検索行動に
合致するか否か

XML情報検索の今後の動向

- データベース業界
  - 情報検索業界との統合を促進したい
    - スコアリングの定式化
    - 処理の高速化
    - オントロジ構築，利用による高精度化
  - 今後も処理の高速化のための研究が続けられる
    - 高精度化には処理速度の問題は常に付きまとうため
    - 如何に高精度化のための各種処理を高速化するか?

XML情報検索の今後の動向

- 情報検索業界
  - 業界の統合には反対はしない?!
  - ただ業界内でやらねばならない研究はまだまだある
    - 人間の行動
    - 検索結果の粒度決定とそのスコアリング (ad-hoc)
      - 人間の視点にたったスコアリングとは何か?



WI2コミュニティにも大きく関連