お茶の水女子大学
Ochanomizu University

ミニサーベイ

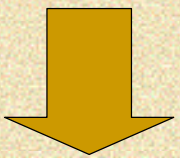
XML文書を対象とした情報検索

お茶の水女子大学
総合情報処理センター
絹谷弘子

データベースと情報検索の統合

- データベースからのデータ抽出と多量の文書からの情報検索をシームレスに行いたい
- データベース研究者と情報検索研究者のコミュニティは分かれていた

XMLの出現



- XMLを介したデータベースと情報検索の統合に向けた動き

発表の流れ

- データベース分野におけるXML検索技術
  - XQuery and XPath Full-Text
  - TeXQuery, FleXPath, GalaTex
- 情報検索分野におけるXML検索技術
  - XIRQL
  - INEXプロジェクトの役割
- データベースと情報検索の統合に向けた動き
  - TDM04
  - SIGIR04 Workshop

XML登場以前

- データベースの分野
  - 1970 関係データベース
  - 1990 入れ子関係モデルとオブジェクト指向データベース
  - 1995 半構造データベース
- 文書の分野
  - 1974 SGML (Structured Generalized Markup Language)
  - 1990 HTML (Hypertext Markup Language)
  - 1992 URL (Universal Resource Locator)

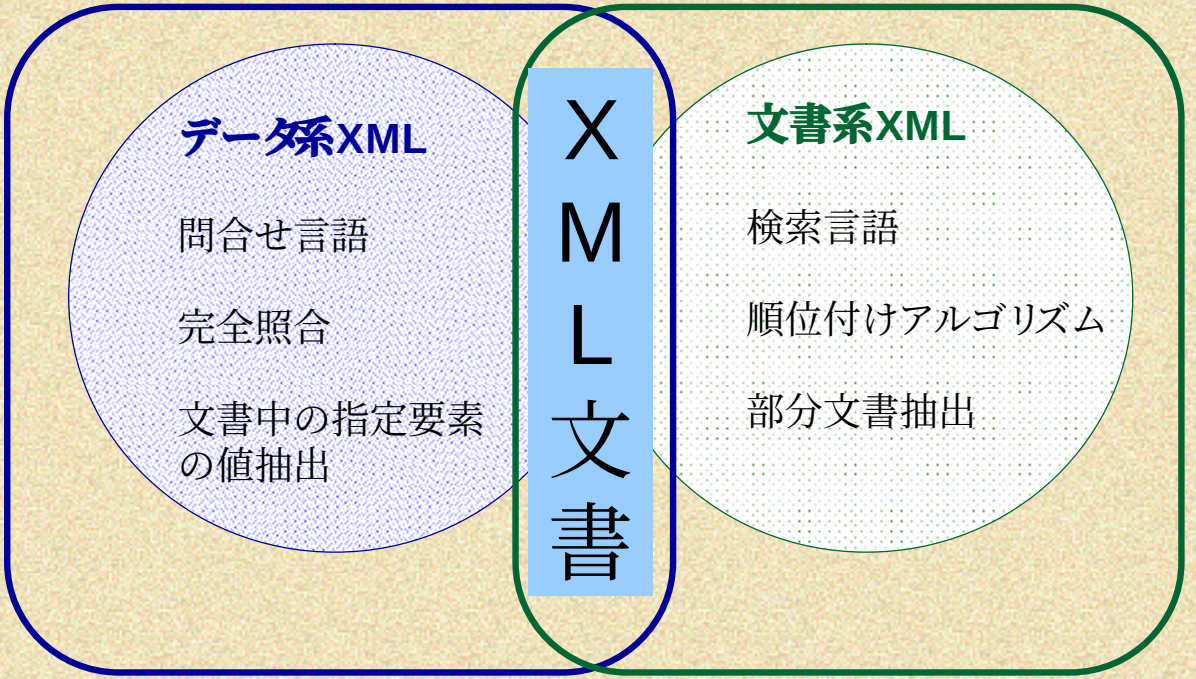
XML登場

- データベースの分野
  - 1970 関係データベース
  - 1990 オブジェクト指向データベース
  - 1995 半構造データベース
- 文書の分野
  - 1974 SGML (Standard Generalized Markup Language)
  - 1990 HTML
  - 1992 URL

データベース＋文書＝情報
1998 XML (Extensible Markup Language)
URI (Universal Resource Locator)

XML文書に対する検索
＝　データベース検索＋文書検索

XML文書


データ系XML
問合せ言語
完全照合
文書中の指定要素の値抽出
XML文書
文書系XML
検索言語
順位付けアルゴリズム
部分文書抽出
データベース検索
文書検索

XML文書

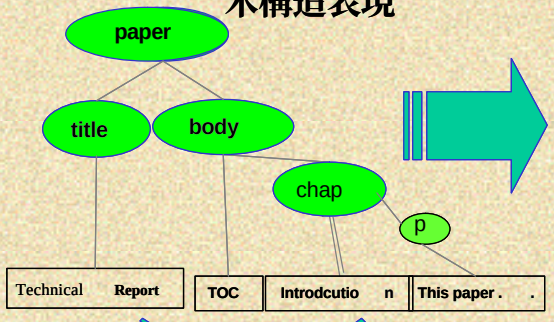
DTD

```
<!ELEMENT paper (title, body)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT body (chap*,p*)>
<!ELEMENT chap (p*,#PCDATA)>
<!ELEMENT p (#PCDATA)>
```

XML　文書

```
< paper>
<title>Technical Report</title>
<body>TOC
<chap>Introduction
<p>This paper  . </p>. .
</chap>
</.body>
</paper>
```

木構造表現



内容情報

Technical Report
TOC
Introduction
This paper ...
…

構造情報（経路式）

/paper[1]/title[1]
/paper[1]/body[1]
/paper[1]/body[1]/chap[1]
/paper[1]/body[1]/chap[1]/p[1]
1]
…

内容＋構造情報

technical	/paper[1]/title[1]
report	/paper[1]/title[1]
toc	/paper[1]/body[1]
introduction	/paper[1]/body[1]/chap[1]
this	/paper[1]/body[1]/chap/p[1]
paper	/paper[1]/body[1]/chap/p[1]
… ..	

発表の流れ

- **データベース分野におけるXML検索技術**
  - **XQuery and XPath Full-Text**
  - **TeXQuery, FleXPath, GalaTex**
- 情報検索分野におけるXML検索技術
  - XIRQL
  - INEXプロジェクトの役割
- データベースと情報検索の統合に向けた動き
  - TDM04
  - SIGIR04 Workshop

XMLデータベースに対する研究

研究課題

7th EDBT Summer School
Ioana Manolescu のスライドより

- XMLデータのデータベースへの格納方法
  関係データベース, ネイティブデータベース
- XMLデータの再構築方法
- XMLデータの木構造中のノードへの符号化(Encoding)方法
  データベース内部でのXMLデータへの効率的なアクセスのため
- XMLデータのディスク格納時の分割方法（水平分割, 垂直分割）
- XMLデータベースへの問合せ言語と問合せ処理

XML問合せ言語 XQuery (2001～)

```
<?xml version="1.0"?>
<bib>
  <book>
    <title>TCP/IP Illustrated</title>
    <author>Stevens</author>
    <publisher>Addison-Wesley</publisher>
  </book>
  <book>
    <title>Advanced Unix Programming</title>
    <author>Stevens</author>
    <publisher>Addison-Wesley</publisher>
  </book>
  <book>
    <title>Data on the Web</title>
    <author>Abiteboul</author>
    <author>Buneman</author>
    <author>Suciu</author>
  </book>
</bib>
```

XML文書

XQueryでの問合せ

```
<authlist>
 {
   for $a in distinct-values(input()//author)
   order by $a
   return
     <author>
       <name>
         { $a/text() }
       </name>
       <books>
         {
           for $b in input()//book[author = $a]
           order by $b/title
           return $b/title
         }
       </books>
     </author>
 }
</authlist>
```

XML問合せ言語 XQuery

```
<authlist>
  <author>
    <name>Abiteboul</name>
    <books>
      <title>Data on the Web</title>
    </books>
  </author>
  <author>
    <name>Buneman</name>
    <books>
      <title>Data on the Web</title>
    </books>
  </author>
  <author>
    <name>Stevens</name>
    <books>
      <title>TCP/IP Illustrated</title>
      <title>Advanced Unix Programming</title>
    </books>
  </author>
        :
        :
```

結果として出力された文書

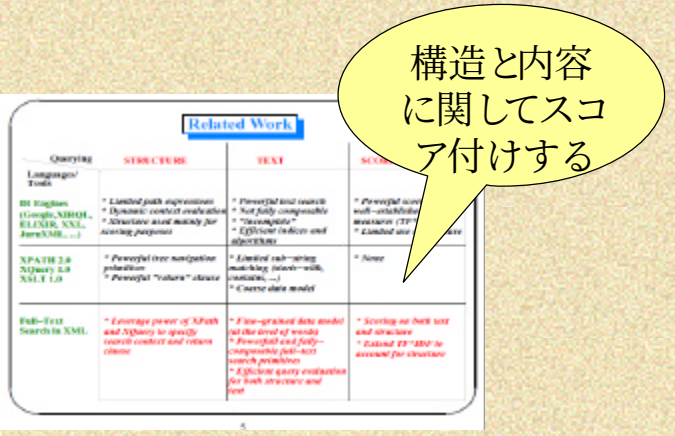
XQueryでの問合せ

```
<authlist>
{
  for $a in distinct-values(input()//author)
  order by $a
  return
    <author>
      <name>
        { $a/text() }
      </name>
      <books>
        {
          for $b in input()//book[author = $a]
          order by $b/title
          return $b/title
        }
      </books>
    </author>
}
</authlist>
```

XML問合せ言語への全文検索機能

EDBT Summer School http://edbtss04.dia.uniroma3.it/program.html

7th EDBT Summer School　Sihem Amer-Yahiaスライド
http://edbtss04.dia.uniroma3.it/AmerYahia.pdf



XQuery Full-Text

- XQuery and XPath Full-Text Requirements(2003,5)
- XQuery 1.0 and XPath 2.0 Full-Text(2004,7)
- XQuery 1.0 and XPath 2.0 Full-Text Use Cases(2004,7)
- 全文検索としての主要な機能

  単語検索, フレーズ検索, 不要語処理, 接頭語指定, 接尾語指定, 単語を単位とする近接検索, 順序を指定した近接検索, AND, OR, NOT, 語の正規化, 語尾処理, 語の区切り処理, 順位付け, 関連度（スコア）

データベース研究者の研究
XQueryへの全文検索機能の追加

- **Sihem Amer-Yahia (AT&T)**
  - **Phrase matching in XML**, S. Amer-Yahia, M. F. Fernandez, D. Srivastava, Yu Xu, VLDB03
  - **TeXQuery**: A Full-Text Search Extension to XQuery, S. Amer-Yahia, C. Botev, J. Shanmugasundaram, WWW2004
  - **TeXQuery-Based XML Full-Text Search Engine**: C. Botev, S. Amer-Yahia, J. Shanmugasundaram, SIGMOD04
  - **GalaTex**: http://www.galaxquery.org/galatex

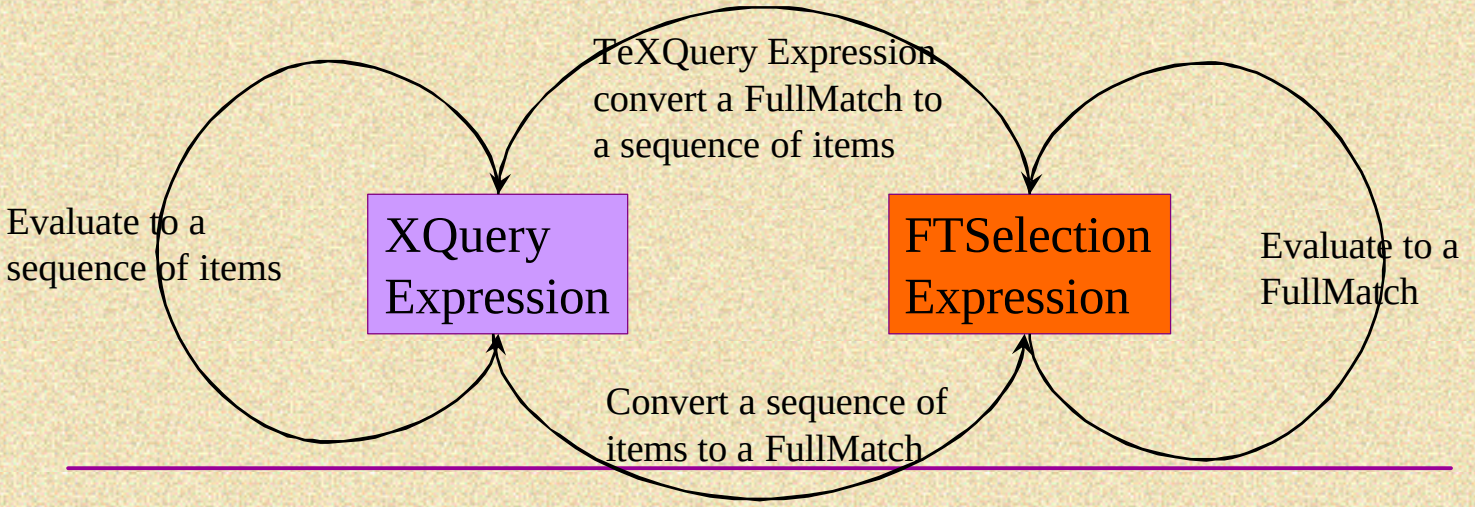
データベース研究者の研究
問合せ,検索結果の緩和(Relaxation)

- **Approximate Tree Embedding for Querying XML Data,** T. Schlieder, F. Naumann, SIGIR00 Workshop on XML and Information Retrieval (ApproXQL)
- **Aproximate matching of XML Queries** : Sihem Amer-Yahia, Nick Koudas, Divesh Srivastava, ICDE03
- **FleXPath**: Flexible Structure and Full-Text Querying for XML, S. Amer-Yahia, Laks V. S. Lakshmanan, Shashank Pandit, SIGMOD04
- **Approximate XML Query Answers**, Neoklis Polyzotis, Minos N. Garofalakis, Yannis E. Ioannidis, SIGMOD04 (TREESKETCH)
- **CoXML (Cooperative XML)** : UCLA http://www.cobase.cs.ucla.edu/projects/CoXML/coxml.html, RLXQuery(ReLaxed XQuery)

TeXQuery(1) : XQuery全文検索

- TeXQueryデータモデル
  - XQueryデータモデルを拡張
  - XQueryのコア部分には手を入れない

    XQueryの外部関数として定義, 呼び出す



TeXQuery (2)

- 部分文書のランキング
  - ユーザの問合せに与えられる重み
  - 単語の出現頻度による重み (tf-idf など)
  - 問合せの文書構造と検索結果の文書構造の比較 (structural similarity)
- GalaTex : TeXQueryの実装

  XQueryの実装Galax(http://www.galaxquery.org/)を拡張してTeXQueryを実装

FleXPath

- FleXPath：XQueryの問合せの緩和
  - キーワードに関連のある部分文書のXPath式の簡略化 (relaxation)
    - 条件を緩和していく
  - 部分文書のためのランキングスキームの提案
    - structural scoreとkeyword scoreからanswer scoreを求める
      - IRの分野で提案されているもの(XIRQL, JuruXML etc.)とは異なる
      - 構造スコアとキーワードスコアの和, 平均, 中間値
  - top-Kランキングのための効率的なアルゴリズムの提案

発表の流れ

- データベース分野におけるXML検索技術
  - XQuery and XPath Full-Text
  - TeXQuery, FleXPath, GalaTex
- **情報検索分野におけるXML検索技術**
  - **XIRQL**
  - **INEXプロジェクトの役割**
- データベースと情報検索の統合に向けた動き
  - TDM04
  - SIGIR04 Workshop

情報検索分野におけるXML検索

- Web サーチエンジンは非常に有用
  - ページ間のリンク構造を考慮したランキング付き検索結果表示
  - 索引を利用した高速な検索
  - 利用者に優しい問合せインタフェースによるキーワード指定
- XML文書検索エンジンが持つべき機能
  - 上記機能は必須
  - それに加えて
    - 属性や要素などの文書構造を考慮したランキング
    - 文書構造, 文書内容の問合せ方法
    - 文書単位の検索に加え部分文書単位の検索
    - 利用者が利用しやすい検索結果の提示法

情報検索分野における XML文書検索の研究

- 2000年　SIGIR Workshop
  - XML and Information Retrieval 第1回
- 2002年　論文誌JASIST に特集
- 2002年　SIGIR Workshop
  - XML and Information Retrieval 第2回
- 2004年　SIGIR Workshop
  - XML and Information Retrieval 第3回
- INEXプロジェクト
  - INEX2002 2002年4月～12月
  - INEX2003 2003年4月～12月
  - INEX2004 2004年4月～12月

情報検索分野における XML文書検索の研究

- **XIRQL**：N. Fuhr, K. Grossjohann, SIGIR00 Workshop on XML and Information Retrieval, SIGIR01
- **Searching Text-rich XML Documents with Relevance Ranking**: Y. Hayashi, J.Tomita, G.Kikui, SIGIR00 Workshop on XML and Information Retrieval

XIRQL

- XML問合せ言語XQLを拡張し，指定したXPath式を満たすXML部分文書に対してキーワード検索を実行できる問合せ言語
  - 検索モデルは確率モデル
    - 検索結果の順位づけは部分文書の内容によって定まる
  - キーワードに対するAND，OR検索だけではなく，重みづけ検索なども可能
  - 索引づけはシステム設計者が選別した部分文書だけに対して実行
- precision-recallを用いて評価
  - 人によって選別された部分文書を索引づけしているため検索精度は高い
  - 索引づけされていない部分文書は検索できない
    - 決めうちによる弊害

INEXプロジェクト

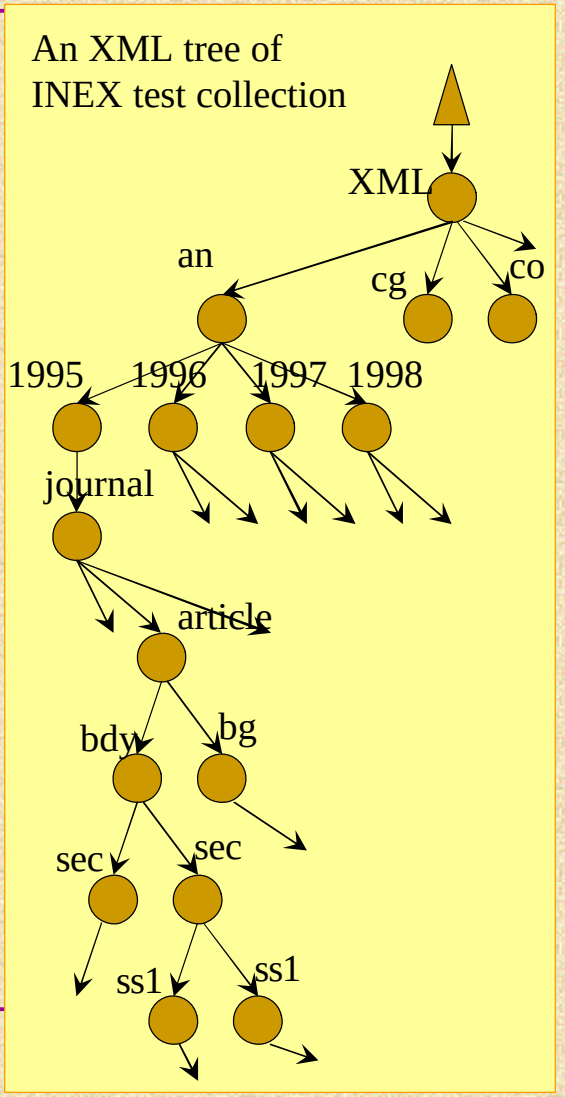
Initiative for the Evaluation of XML Retrieval (INEX)

- XML検索システムの効果を評価する手段が必要　→　テストコレクションを作成する
  - 文書集合，トピックス，正解集合

- 2002年 4月プロジェクト発足
  - オーガナイザ　DELOS Network of Excellence for Digital Libraries
  - プロジェクトリーダ　Norbert Fuhr　（Dortmund大学）
    Mounia Lalmas　（Queen Mary大学）

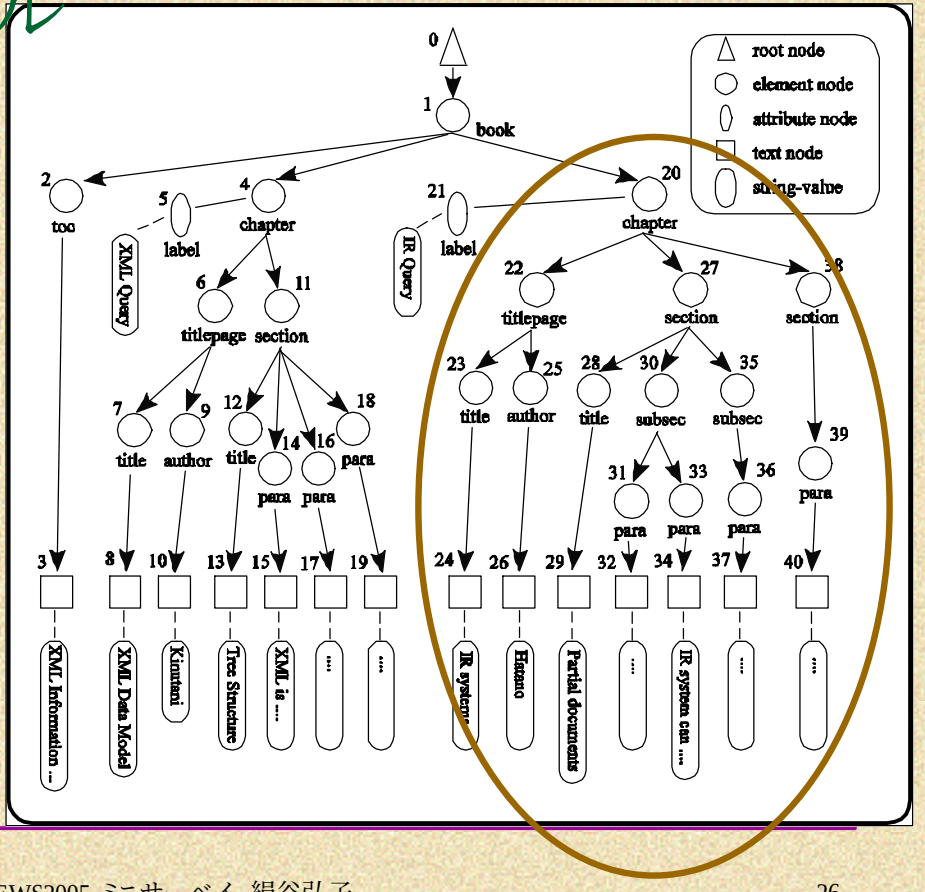
INEX テストコレクション文書集合

- IEEE Computer Societyの出版物
  - 12 magazines, 6 transactions
  - 494M Bytes
  - 12,107XML articles
  - 1,532 ノードarticle 木の深さ6.9



INEXテストコレクション XML文書モデル

- Non-overlapping lists
  - 部分文書単位 :XML文書を重複のない文字列に分割
- **Proximal nodes**
  - 部分文書単位 :XML文書の論理木構造のうち要素ノードを根とする部分木に対応する文書部分
  - 根ノードの要素を経路式で表現できる



INEX テストコレクション
トピックス

- 2002年, 2003年　アドホック検索に重点 :2種類の トピック
- Content-only (CO) トピック
  - 内容に関する条件だけを含み、文書構造を指定しない問合せ要求
- Content-and-structure (CAS) トピック
  - トピックステートメントで、XML構造についての参照を含み、context of interest や　context of certain search conceptsを限定する
  - VCAS　構造, 内容とも指定した条件に類似する文書部分を求める
  - SCAS　検索対象となる文書部分が指定した構造条件を厳密に満たす場合だけを求める

精度/再現率
- Recall and Precision -

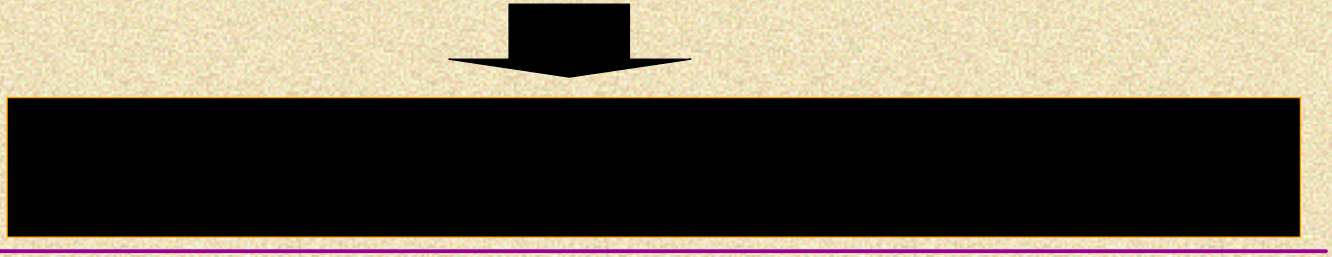
- 再現率 (Recall)
  - Fraction of relevant documents (*R*) which has been retrieved

$$\mathrm{Re} = \frac{|R \cap A|}{|R|}$$

- 精度 (Precision)
  - Fraction of retrieved documents (*A*) which is relevant

$$\mathrm{Pr} = \frac{|R \cap A|}{|A|}$$

従来のIRでは検索結果文書間に重複がないことが前提



INEX 2002

http://qmir.dcs.qmul.ac.uk/inex/index.html

- 2002年4月～12月
- 49グループ登録，20(+6)グループ実働
- アドホック検索の基礎固め
- テストコレクション作成：XML文書（IEEE論文），トピックス，正解集合（XML部分文書）
  - 入れ子構造となっていて、どの階層でもretrievable unit
  - 文書内容と文書構造の条件を結合できるし，特定の要素に検索を限定することもできる
  - 文書構造を考慮したrelevance assessments を考慮しなければならない

    relevance assessmentの評価尺度：relevance, coverage

INEX 2003

http://inex.is.informatik.uni-uisburg.de:2003/

- 2003年4月～12月
- 46グループ登録, 28グループ実働
- トピックス指定にabout関数を導入

  //article[about(.//sec, “XML retrieval”, +XML, -”information retrieval”)]
- relevance assessmentの評価尺度を exhaustivity, specificity に変更

INEX 2004

http://inex.is.informatik.uni-duisburg.de:2004/

- 2004年4月～12月
- 55グループ登録2 7グループ実働
- アドホック検索だけでなくHeterogeneous, Interactive, Natural Query, Relevance Feedbackの各トラック別に活動
- 昨年に比べ, 検索行動分析やRelevance Assessments　の結果分析など、ワークショップでの発表の視点が多角化する
- 活発な活動を行うグループがいくつか出現し、オピニオンリーダとなっている。

INEXテストコレクションの認知

- 文書系XMLとして多数の研究グループが利用
- 多数の研究成果が報告されるようになってきた
- 活動の活発なグループ
  - Fuhr (Duisburg-Essen大学)
  - Queen Mary Information Retrieval (QMIR) research group: M. Lalmas, G. Kazaiら
  - Holger Meuss (European Southern Observatory)
  - IBM Haifa研究所：Y.Mass, D.Carmel,S.Maarek
  - アムステルダム大学：J.Kamps,B.Sigurbjörnsson

INEXプロジェクトの課題

- relevance assessmentでの2次元の評価値からprecision/recallを導く手法
- 検索結果の部分文書間にあるオーバーラップの扱い
- CAS (Content and Structure) トピックの正解とは何か？　構造の類似性, 内容の類似性の扱い
- 3年間のプロジェクトで参加者間のXML情報検索に対するコンセンサスはできた。しかし、情報検索研究者はprecision/recallに興味が集中しがち

情報検索とデータベースの統合に向けた動き

- 2004年，データベース研究者と情報検索研究者が集ってXML情報検索とXMLデータベースに関するワークショップが開催された

1. TDM:The first Twente Data Management Workshop on XML Databases and Information Retrieval (6月）
2. SIGIR04 Workshop (7月)
   - 第3回 XML and Information Retrieval
   - 第1回 Integration of IR and DB (WIRD)

TDM 04

- オランダTwente大学の研究者が主催
- XML情報検索の目標の問合せ(Fuhr)

  “an artist named Ulbrich living in Frankfurt, Germany about 100 years ago”
  - データベース的 :artist の属性 “Joseph Maria Olbrich, Darmstadt,1901”
  - ある程度の類似性を許すと

    音の類似性： Ulbrich →　Olbrich

    画像の類似性： Frankfurt → Darmstadt　地図上で近い）

    データの類似性： 1901 → 1904

  類似性にも種類がある。
- XML情報検索へのデータベースアプローチ (G.Ram?rez)
  - 経路式上のjoin索引とコスト関数で問合せ処理時間を削減

SIGIR04ワークショップ

- David Hawking 招待講演）

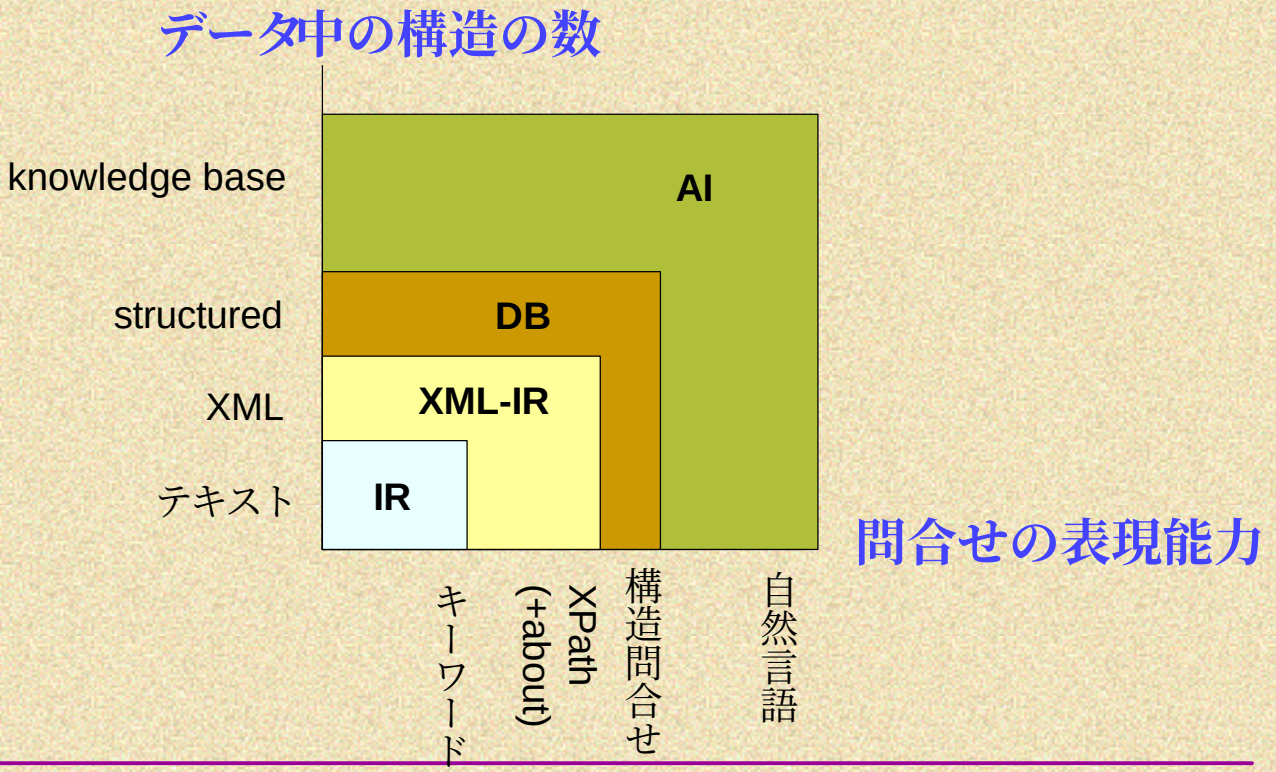
  Webサイト, データベース, email, 共有ファイルシステムなど多様な文書やデータ集合を統一的に検索可能なシステムが必要 (Enterprise Search タスク)
- XMLとIR 発表4件
- IRとDB 発表3件
- 今後の研究課題

  半構造検索, 問合せ言語, 構造の多様さ

  提案手法の評価方法, 効果的でスケーラブルな検索システム, オントロジが必要か否か, customer relationship managementへの関与

SIGIR04ワークショップ

- 議論　XML-IRの位置づけ



まとめ

XML文書を対象とした情報検索

- 情報検索技術とデータベース技術
- 情報検索研究者とデータベース研究者
- 多様な文書やデータをシームレスに利用
- 検索結果にランキングを入れる
  - 類似性とは？　部分文書が単位
- DB-XML-IR　という関係についての共通認識はできた？