

Wikipediaと図書館資料の分類体系との対応づけ ～Web上の二次情報と図書館の一次情報の統合的活用に向けた取り組み～

東京大学情報基盤センター
図書館電子化研究部門 助手
清田 陽司

研究の意義

Webの世界と図書館の世界をつなげる

- 図書館の存在意義の再発見
- 図書館における情報検索の知恵をWeb検索に生かす

図書館の存在意義

- 過去の情報の蓄積
- 調査・研究・教育のためのインフラ
- 情報リテラシー教育

現状のWeb検索の問題点と図書館の利用

- 適切な検索キーワードを選ぶことが難しい
→ 情報要求を具体化させるためのサービス(レファレンスサービス)・情報資源(レファレンスブック)
- 検索結果(=ページの羅列)から必要な情報を見つけにくい
→ 情報の整理・組織化
- 信頼性に欠ける情報が少なくない
→ 一次資料(書籍・学術論文など)での検証

図書館の利用によって補完可能

レファレンスサービス

「図書館のコンシェル ジェ・サービス」

- 質問への回答
- 文献の提供
- 図書館利用法の援助・指導

利用者との対話によって、
漠然とした情報ニーズ
を具体化する

質問の例

- 利用案内
 - 判例時報はどこに置いてありますか
 - OPACの使い方を教えてください
- 所蔵調査
 - 「歌舞伎登場人物辞典」は図書館にありますか
- 文献調査
 - 江戸時代の農民一揆に関する本を探したい
 - 昭和20年代の東大の雰囲気を知ることができる資料が欲しい
- 事項調査
 - 大学いもの「大学」は東京大学のことを指しているのですか
 - 日本における英語教育制度の始まりについて知りたい

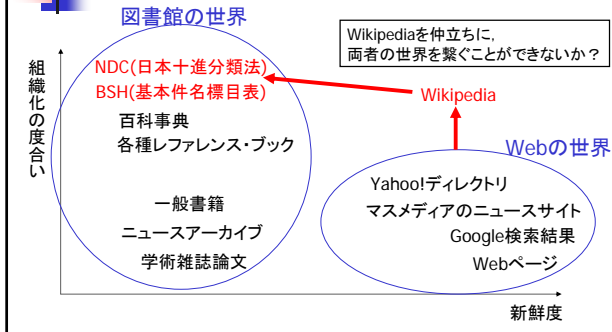
現状の図書館の問題点

- 情報の即時性に欠ける
 - 出版されてから貸出までのタイムラグ
- オンラインサービスが未発達
 - OPAC: 目録のみ、内容までは検索できない
 - レファレンス・サービス: 利用者との対話が難しい
- 利用者から閉じたギルド的世界
 - わかりにくい用語: 目録, 件名, 請求記号...
 - 提供されるサービスが利用者にわかりやすい形で整備されていない
 - 個々のサービスが連携していない

解決策

- Web上の情報資源と図書館の分類体系の対応づけ
 - Web上の情報資源の信頼性を図書館所蔵の一次資料によって検証可能とする
 - 図書館に存在しない最新の情報をWebによって補う
- 図書館資料探索に関するメタ知識の整備
- 図書館利用に関するポータル的なシステムの開発

情報資源の性質



NDC (日本十進分類法)

- 日本で最も普及している図書館資料の分類法
 - 日本図書館協会によって作成・維持
 - 1つの資料に1つのNDCコード
 - 他の分類法
 - DDC (デュイ十進分類法)
 - UDC (国際十進分類法)
 - NDLC (国立国会図書館分類表)
 - LCC (米国議会図書館分類表)
- | | |
|---------|-------------|
| 000 | 総記 |
| 002 | 知識・学問・学術 |
| 007 | 情報科学 |
| 007.1 | 情報理論 |
| 007.13 | 人工知能・パターン認識 |
| 100 | 哲学 |
| 200 | 歴史 |
| 300 | 社会科学 |
| 400 | 自然科学 |
| 500 | 技術・工学・工業 |
| 540 | 電気工学 |
| 548 | 情報工学 |
| 548.958 | 情報検索・機械検索 |
| 600 | 産業 |
| 700 | 芸術 |
| 800 | 言語 |
| 900 | 文学 |

BSH (基本件名標目表)

- 図書館資料の統制キーワードリスト
 - 日本図書館協会によって作成・維持
 - 1つの資料に複数の件名
 - 他の件名標目表
 - LCSH (米国議会図書館件名標目表)
 - NDLSH (国立国会図書館件名標目表)
- | | |
|------|--|
| 情報理論 | NDC007.1 |
| TT: | 情報科学 |
| BT: | 情報科学 |
| NT: | アルゴリズム、エントロピー、音声処理、可視化技術、形式言語、言語情報処理、コンピュータグラフィックス、人工知能、図形情報処理、... |
| UF: | サイバネティクス |
| TT: | 最上位標目 |
| BT: | 上位標目 |
| NT: | 下位標目 |
| UF: | 直接参照あり |
| RT: | 関連標目 |

Wikipedia

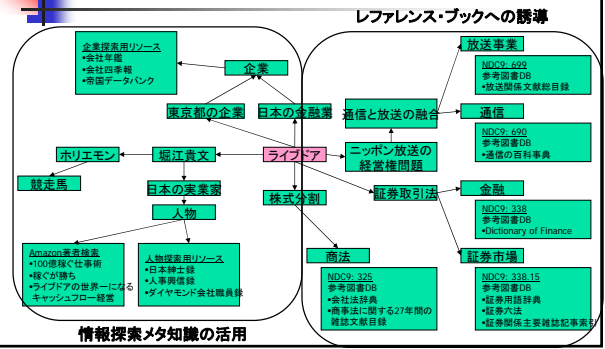
- インターネット上で共同編集されている多言語百科事典
 - Wikimedia財団によるプロジェクト
 - 各記事にはカテゴリが付与されている
 - 複数の上位概念を与えることができる(他の分類体系との大きな関連)
- 編集**
編集(かかく)とは、有形・無形の各種の**商品**(サービスを含む)の取引に際して提示される金額を言う。基本的には**需要と供給**のバランスによって決定される。一般には、**値段(ねだん)**とも呼ばれる。
- 関連項目**
・**アダム・スミス**
・**需要と供給**
...
- カテゴリ: **[経済学]** **[市場]**
-

複数の分類体系の統合的利用

- NDC, BSH, Wikipediaカテゴリを統合的に探索可能なインターフェースを試作した
- 文字列レベルで一致するカテゴリ名を対応づけ
- 上位、下位、関連項目などへのリンク
- 図書館資料(OPAC), Amazonなどとの連携



探索例: ライブドア



自動対応づけ手法の研究

- Wikipedia記事に出現するキーワードの利用
 - ベクトル空間法
- 体系構造の類似している部分の検出
- 機械学習手法の適用
 - 上位・下位・兄弟概念をfeatureとして与える

自動レファレンスサービスへの発展

- 情報探索メタ知識の整備
 - 図書館利用ガイドブック
 - 過去の質問応答事例の蓄積
 - 質問タイプ別の情報探索戦略
- 質問と知識データベースの照合
 - 一般概念への置き換え(e.g. 発電工学→電気工学)
 - 自然文同士の柔軟なマッチング
- 対話型インタフェース
 - cf. ダイアログナビ, 京大レファレンスサービスシステム

多様な情報探索戦略の組み合わせ

- キーワード検索
 - OPAC, Amazon, Google, Kiwi, ...
- 過去の質問事例との照合
 - 国立国会図書館 共同レファレンスデータベース事業
- テーマ別主要リソース提示(Top-down一般化)
 - NDCコードへの一般化
- 関連概念提示(Bottom-up一般化)
 - 上位・下位・兄弟概念
- 質問タイプへの一般化
 - 「いつ～」→歴史・日時情報
 - 「どこで～」→地理・地名情報