【パート II：情報分野研究者のためのオンリーワン共有イノベーションプラットフォーム】

# 1. SlothLib/EaRDB：マイサーチエンジン開発環境支援

大島 裕明[*1] 中村 聡史[*1] 赤星 祐平[*1]
中田 秀基[*2] 喜連川 優[*3] 田中 克己[*1]

＊1 京都大学
＊2 産業技術総合研究所
＊3 東京大学

## マイサーチエンジンを作る

広く一般に Web が利用されるようになり，Web サーチエンジンが情報インフラとして非常に重要な存在となってきている．Web における情報爆発の現象が進むのに伴い，Web サーチエンジンも，Google，Yahoo!，Live Search などのように Web 全般を検索対象とするものから，画像サーチ，動画サーチ，ブログサーチなどの多様な検索サービスが数多く現れており，Web サーチの多様化が進んでいる状況がうかがえる．

Web サーチの研究においても，これから新たに必要とされるサーチシステムを模索し，さまざまな研究が行われている．それらの研究における試作システムの実装では，各種 Web サーチのサービスが提供する API を利用したり，自然言語処理のツールや既存のクラスタリング手法を利用したりすることなどが広く行われる．これらの API やツールを利用するための準備，既存のアルゴリズムの実装にかかるコストは無視できないほど大きい．また，システムの改良時に別ツールや別アルゴリズムを試すためにはその変更に大きなコストをかけることになる．

今後，さまざまな状況に対応するサーチシステム，たとえば，個々のユーザが自分で構成する「マイサーチエンジン」とも呼ぶべき検索システムが求められるようになると考えられる．そこで，そのようなシステムを比較的容易に構築可能とするプラットフォームとして，我々はプログラミングライブラリ SlothLib[1)]，関係データベース環境における Web データマネジメント環境 EaRDB[2)] と実デバイスマネジメント環境 PeRDB[3)] を開発している．

## SlothLib：Web 研究お手軽ライブラリ

SlothLib[☆1] は，Web サーチ研究におけるソフトウェア作成時のコストを低減させることを目的とするプログラミングライブラリである．SlothLib は新しい機能をライブラリ化して提供するものというよりも，むしろ，すでに存在しているさまざまなツールや既存のアルゴリズムを網羅的に利用可能な環境を提供するものである．Web サーチ API の利用，Web 上の情報取得，文字コード判別，クラスタリング，自然言語処理ツールなどが利用できる．各種機能やアルゴリズムのうち，類似のものには共通インタフェースを持たせて実装しており，システムに変更を加える際には，容易に機能を入れ換えることができる．**図 -1** は SlothLib が持つ機能の概略を表している．

### ● Visual Studio での利用

SlothLib を利用できる環境の 1 つが，Microsoft Visual Studio 2005/2008 である．SlothLib は C# で書かれ，dll ファイル群として提供されており，利用するプロジェクトから必要な SlothLib の dll ファイルの

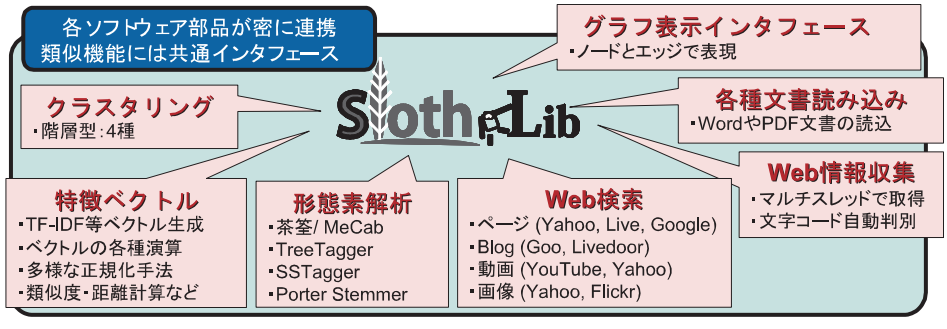


図 -1 SlothLib の機能の概要

☆1 SlothLib は http://www.dl.kuis.kyoto-u.ac.jp/slothlib/ よりダウンロード可能．

参照を行う．

**図 -2** のコードは，簡単な利用例である．まず，Yahoo! ウェブ検索で「京都」をクエリとして検索を行い，上位 10 件の検索結果を取得し，検索結果の概要部分のみを取得している．その文章を形態素解析器 MeCab を用いて解析し，品詞が「名詞」か「動詞」の語のみを抽出している．その結果を利用して，出現回数を重みとする特徴ベクトルを作成している．

IWebSearch はインタフェースであり，ここで利用している Yahoo! のクラスと同様に，Google や Live Search のクラスも同じインタフェースを実装している．これにより，あとから Yahoo! を別の Web サーチエンジンに置き換えたい場合にも，変更するのは 1 行だけで済む．

形態素解析を行う IMorphologicalAnalyzer インタフェースを実装するクラスには，MeCab のほかに，茶筌や，英語の形態素解析器である SSTagger，TreeTagger などがある．日本語用のシステムを英語に対応させる場合，これらがすべて共通のインタフェースを持っているため，容易に入れ換えることが可能になっている．

特徴ベクトルも各種が同じインタフェースを実装している．ここでは最も簡単に語の出現回数（TF）を重みとして作成したが，TF の Log を取る，正規化を行う，TF-IDF のような重みにする，といったことが容易にできるようになっている．

### ● Java での利用

SlothLib を利用できるもう 1 つの環境が Java 開発環境である．Java は非常に多くのプラットフォームでサポートされているため，SlothLib を広くさまざまな環境で実行することが可能である．現在開発の対象としている実行環境は，Windows，Linux，Mac OS X である．Java のプロパティファイルを利用することで環境の違いを吸収している．また，SlothLib から利用可能な外部プログラムのうち，Windows 専用のものに関しても，可能な限り代替プログラムを用意して同等の機能を提供している．Java 版 SlothLib は 1 つの jar ファイルとして提供され，いくつかの外部ライブラリとともに，CLASSPATH に追加することで，ユーザが作成するアプリケーションから利用できる．Java 版はメンテナンス性を高めるため，C# 版を極力逐語的に変換する方針で実装されている．そのため，基本的なクラス構成は C# 版と同一であり，たとえば図 -2 のサンプルコードを機械的に変更するだけで，ほぼそのまま実行することができる．C# 版との主な相違点は 3 点ある．まず，Java にはプロパティ機能がないため，必要に応じて getter/setter で実装されている．さらに，Java でのコンベンションに従い，メソッド名を小文字で始めている．また，Java では実行時例外の処理が C# よりも厳格であるため，例外処理コードの追加が必要である．

```
// Yahoo!でクエリ「京都」の結果を10件取得
IWebSearch yahoo = new YahooJpWebSearch("slothlib");
IWebSearchResult result = yahoo.DoSearch("京都", 10);
// 検索結果の概要部分を取得
StringBuilder sb = new StringBuilder();
foreach (IWebElement element in result.ResultElements)
  sb.Append(element.Description);
// 形態素解析器MeCabを利用する
IMorphologicalAnalyzer jm = new MeCab();
// 品詞が「名詞」か「動詞」の語を残すフィルタ
IMorphemeFilter pos = new PosFilter("名詞|動詞");
// 形態素解析の結果中から語の原型を取り出すフィルタ
IMorphemeToStringFilter org = new
                    RemainOriginalFilter();
// 形態素解析を行い，フィルタを適用する
IMorphologicalAnalyzerResult mr =
                    jm.DoAnalyze(sb.ToString());
IMorpheme[] morphs = pos.DoFilter(mr.Morphemes);
string[] terms = org.DoFilter(morphs);
// 語の出現回数による特徴ベクトル生成
IVector<string> v = new FrequencyVector<string>(terms);
```

図 -2　Visual Studio 環境におけるコード例

Java 版においては，Eclipse と Visual Editor もしくは Jigloo などの外部 GUI 作成ツールを併用することで，Visual Studio と同様に GUI を持つアプリケーション開発を容易に行うことができる．**図 -3** では，Java で Web サーチを行うプログラムを作成している．図右の Eclipse 上では，GUI を作成し，ボタンが押されたときにユーザがコンボボックスで選択中の検索エンジンを利用して検索を行い，その結果を出力するプログラムを記述している．図左がプログラムの動作例である．

## EaRDB：Web データマネジメント環境

研究の初期段階では，Web 上のデータを実際に解析して，方針の模索を行うことがしばしば行われる．しかし，そのためだけにプログラムを書くのはコストが高い．EaRDB は，関係データベース（RDB）環境において Web サーチの結果を取得して，自然言語処理を含む解析を行ったりすることが可能な環境である．SlothLib を利用するなどしてプログラムを作成する前の段階で，Web サーチを利用したさまざまな実験をアドホックに行うようなことが容易に行える．**図 -4** の上部が EaRDB の概念図である．Web サーチの結果を取得する機能や自然言語処理機能は，すべて仮想テーブル☆2 として実

☆2　ここでの仮想テーブルとは，SELECT 文で通常のテーブルのようにデータを取得することが可能だが，引数として与えられた値に応じて異なったデータを返すテーブルである．

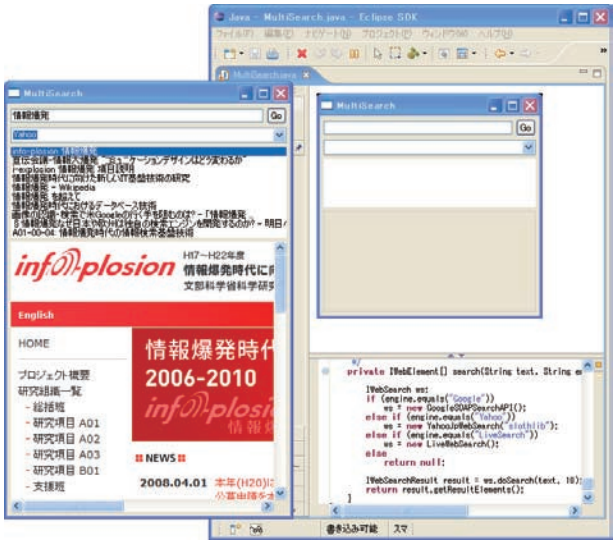

図 -3　Java 版 SlothLib を利用した GUI プログラム開発

装されており，RDB のクエリ言語である SQL を用いて扱う．たとえば，Web サーチの結果は，仮想テーブルに要求があったときに Web サービスの API を利用して取得され，テーブル形式で返されるようになっている．EaRDB は現在，Microsoft SQL Sever 2005 上で実装されている．ここでは実際に，SQL の優れた集約機能を利用して，Web サーチ結果からの知識取得を短い記述で実現する例を示す．

### ● Web から知識を得る

**図 -5** は Web 検索を利用して富士山の標高を求める SQL である．FROM 節で WebSearchJp という Web 検索を行う仮想テーブルと，WordExtractJp という指定したパターンにあてはまる語を抽出する仮想テーブルにアクセスしている．まず，「富士山 標高」をクエリとして Web サーチを行い，「メートル」や「m」などの直前に頻出する数値を求めている．この場合，富士山の標高として正しい答えである 3776 が，最も出現回数が多いという結果になった．Web サーチを上手く利用すれば，知識の獲得が可能であり，EaRDB ではそれを気軽に試すことができる．

**図 -6** は Web サーチの結果を，ソーシャルブックマークされている件数の多いものを上位にするように独自にランキングを行う SQL である．検索結果のランキングの変更を行っており，非常に簡単ではあるが自分好みの検索システムの実現ということができるだろう．

### ●実デバイスマネジメント環境

マイサーチエンジンの実現が情報空間にとどまる必要はない．現在，ユビキタスデバイス環境の整備が進み，実世界で扱えるデバイスは増え続けている．実世界の状況に応じて，センサやディスプレイを組み合わせた Web サーチを適宜構築することは遠い未来の話ではないと考える．PeRDB は EaRDB と同様に RDB 環境で実デバイスの操作を行うことができる環境である．図 -4 の下部が PeRDB の概念図である．PeRDB で管理されたセンサやアクチュエータに対する操作は，す

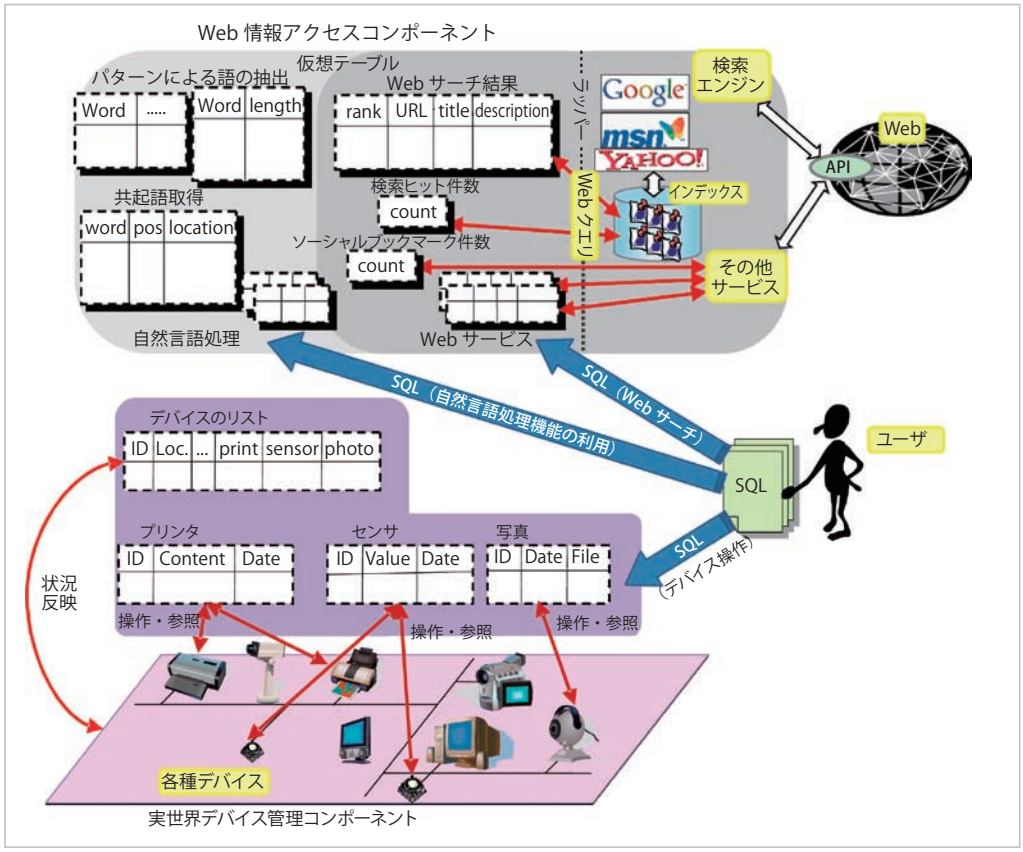


図 -4　関係データベース環境における Web とデバイスの操作

```
SELECT dbo.ToNumerical(word) height,
       count(dbo.ToNumerical(word)) c
FROM WebSearchJp('富士山 標高', 100) sr
 CROSS APPLY
  WordExtractJp(sr.description, '<term>(メートル|m|m)')
WHERE length = 5
GROUP BY dbo.ToNumerical(word)
ORDER BY c DESC;
```

| | height | c |
|---|---|---|
| 1 | 3776 | 28 |
| 2 | 2400 | 8 |
| 3 | 3000 | 5 |

図 -5　EaRDB で富士山の標高を求める

```
SELECT dbo.SocialBookmarkCount(url) c, *
FROM WebSearch('Java', 50)
ORDER BY c DESC;
```

| | c | rank | url | title |
|---|---|---|---|---|
| 1 | 2705 | 1 | http://java.sun.com/ | Java Technology |
| 2 | 2498 | 4 | http://java.sun.com/do | The Java_ Tutorials |
| 3 | 1541 | 14 | http://www.javaworld.c | Welcome to JavaWorld.com |
| 4 | 1438 | 20 | http://www.javalobby.o | Javalobby - Sun Java, JSP and J2EE te |
| 5 | 1203 | 42 | http://www.onjava.com/ | ONJava.com: The Independent Source |
| 6 | 1011 | 3 | http://www.java.net/ | java.net　The Source for Java Technolo |
| 7 | 484 | 27 | http://developers.sun.c... | Sun Microsystems - Sun Developer Net |
| 8 | 472 | 25 | http://forum.java.sun.c... | Sun Forums |

図 -6　EaRDB でソーシャルブックマーク件数によって Web サーチ結果を再ランキングする

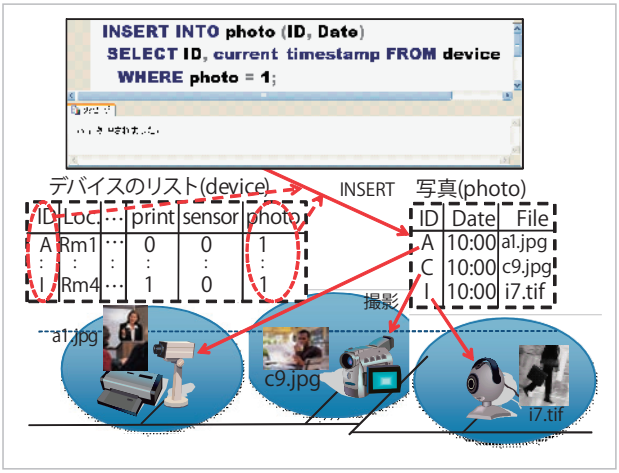


図 -7　PeRDB ですべてのカメラを動作させる

べて SQL によるテーブルへのタプル挿入などで記述される．EaRDB と PeRDB は共にすべて SQL で操作することができ，Web 情報空間と実デバイス情報空間を統合的に扱うアプリケーションを容易に実現することができる．RDB のトリガ機能を利用することなどにより，高度なアプリケーションでも十分に記述可能である．**図 -7** は簡単な利用例である．デバイス管理テーブル（device）から，写真機能を持つデバイスの情報を取得している．次に，写真テーブル（photo）にタプル挿入を行っているが，この操作により実際にデバイスが動作する．この場合，管理されたすべてのカメラデバイスが撮影を行う．写真テーブルには撮影された写真の保存情報が自動的に記録される．このように，多くのデバイスを一度に動作させたり，条件に合うデバイス（たとえば，最も大きなディスプレイ）を選択したりする操作は，SQL の優れた集約操作と高い親和性がある．

## 今後の展望

情報爆発の時代において，情報サーチはますます重要性を増していくと考えられる．そのような中，ユーザの違いや状況の違いに適応したマイサーチエンジンを構築するためのプラットフォームとして，プログラミングライブラリ SlothLib と，Web データマネジメント環境 EaRDB，実デバイスマネジメント環境 PeRDB を紹介した．今後，これらを含め，Web ユーザの誰もが自分のためのサーチシステムを容易に構築できる環境がますます求められるようになると考えられる．

参考文献
1）大島裕明，中村聡史，田中克己：SlothLib：Web サーチ研究のためのプログラミングライブラリ，日本データベース学会論文誌 DBSJ Letters, Vol.6, No.1, pp.113-116 (Mar. 2007).
2）大島裕明，小山　聡，田中克己：Web 集約質問処理のための検索エンジンの関係データベースインタフェース，情報処理学会論文誌：データベース , Vol.48, No.SIG20 (TOD36), pp.50-60 (Dec. 2007).
3）Akahoshi, Y,. Kidawara, Y. and Tanaka, K. : A Database-oriented Wrapper for Ubiquitous Data Acquisition/access Environments, Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication, Suwon, Korea, pp.25-32, ACM Digital Library (Jan. 2008).

（平成 20 年 5 月 1 日受付）

**大島　裕明（正会員）**：ohshima@dl.kuis.kyoto-u.ac.jp
京都大学大学院情報学研究科社会情報学専攻特定助教．2007 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主に Web 情報検索，Web マイニングの研究に従事．電子情報通信学会，日本データベース学会，ACM 各会員．

**中村　聡史（正会員）**：nakamura@dl.kuis.kyoto-u.ac.jp
京都大学大学院情報学研究科社会情報学専攻特定講師．2004 年大阪大学大学院情報学研究科博士後期課程修了．博士（工学）．主にヒューマンコンピュータインタラクション，Web 検索の研究に従事．日本データベース学会会員．

**赤星　祐平（学生会員）**：akahoshi@dl.kuis.kyoto-u.ac.jp
京都大学大学院情報学研究科社会情報学専攻博士後期課程在学中．2005 年京都大学大学院情報学研究科修士課程修了．主に，ユビキタスコンピューティング環境における情報利活用技術の研究に従事．日本データベース学会学生会員．

**中田　秀基（正会員）**：hide-nakada@aist.go.jp
産業技術総合研究所情報技術研究部門主任研究員．1995 年東京大学情報工学系研究科博士後期課程修了．博士（工学）．分散並列計算，グリッド計算の研究に従事．

**喜連川　優（正会員）**：kitsure@tkl.iis.u-tokyo.ac.jp
東京大学生産技術研究所教授．同所戦略情報融合国際研究センター長．1983 年東京大学大学院工学系研究科情報工学専攻博士課程修了．工学博士．データベース工学，並列処理，Web マイニングに関する研究に従事．本会フェロー．電子情報通信学会フェロー．ACM，IEEE Computer Society，日本データベース学会各会員．

**田中　克己（正会員）**：tanaka@dl.kuis.kyoto-u.ac.jp
京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．京大工博．主にデータベース，Web 情報検索，マルチメディアコンテンツ処理の研究に従事．本会フェロー．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，日本データベース学会等各会員．