

Wikipedia エントリ構造抽出ツール: Wik-IE

Wik-IE: A tool to extract structure of Wikipedia entries

森竜也^{1*} 増田英孝¹ 清田陽司² 中川裕志²
Tatsuya Mori¹ Hidetaka Masuda¹ Yoji Kiyota² Hiroshi Nakagawa²

¹ 東京電機大学情報メディア学科

¹ Department of Information System and Multimedia Design, Tokyo Denki University

² 東京大学情報基盤センター

² Information Technology Center, University of Tokyo

Abstract: Web 上のフリー百科事典 Wikipedia では全データが公開されてる。Wik-IE はそのデータファイルから各種情報を抽出してテキストファイルに出力するツールである。

1 はじめに

近年 Wikipedia を辞書・シソーラス作成の情報源やコーパスとして用いることが注目されている。Wikipedia では各言語版の全データが配布されていて [1], 誰でも自由に利用できるようになっている。しかしそれらのデータは XML や SQL ダンプの形式で提供されていて, どの記事がどのカテゴリに属しているか, ある記事に対してどのようなリダイレクトが設定されているか, といった情報を取り出したい場合には, その情報を使いたい研究者が必要なデータをデータベースに格納して操作する必要がある。また記事からある情報を独自に抽出したい場合には, そのためのツールを初めから自分で作成しなければならぬ。このように Wikipedia のデータの利用にはある程度の敷居の高さがある。またそれぞれの研究者が自分で使うためのツールやプログラムを自分で作成するのは無駄が多い。そこで Wikipedia の情報を抽出するツール Wiki-IE を作成した。Wik-IE はオプションを指定して実行するだけで配布データから各種情報を抽出できる Java プログラムである。情報の抽出に使う部分のクラスファイルを定義することで機能を追加することもできる。

2 Wik-IE

2.1 Wik-IE の機能

Wik-IE は標準で次の種類のデータを抽出する機能を持っている。

*E-mail: mori@csl.im.dendai.ac.jp

node エントリのタイトルと ID とページ種別

edge エントリのカテゴリとリダイレクト関係

link Wiki 内リンク

lang 言語間リンク

isbn 記事に記述されている ISBN コード

size 記事の文字列の長さ

2.2 Wik-IE の実行

Wik-IE は使用する XML ファイルと機能を指定するだけで実行できる。Wik-IE の実装にはオープンソースの分散処理プラットフォーム Apache Hadoop[2] を利用している。Hadoop がインストールされている環境下で日本語版 Wikipedia の XML ファイル jawiki-latest-pages-meta-current.xml を使って node データを抽出しディレクトリ result に結果を出力する場合には次のようなコマンドで実行する。

```
hadoop jar Wik-IE.jar node jawiki-latest-pages-meta-current.xml result
```

Hadoop ライブラリが組み込まれているスタンドアロン版 Wik-IE の場合は次のようになる。

```
java -jar Wik-IE.jar node jawiki-latest-pages-meta-current.xml result
```

2.3 生成データの形式

Wik-IE が最終的に出力するファイルは、抽出したデータを表現するタブ区切りのテキストファイルである。出力ファイルの形式は 1 行あたり次のとおりである。

node

```
id title kind
```

id エントリに付与されている ID。

name エントリのタイトル。

kind エントリの種類。leaf(ふつうの記事)・node(Wikipedia の Category)・redirect の 3 種類のいずれか。

edge

```
id(current) id(target) relation
```

id(current) edge の元のエントリの ID。

id(target) edge の先のエントリの ID。

relation エントリ間の関係。hypernym(カテゴリ関係)・target(リダイレクト関係)の 2 種類のいずれか。

link

```
id(current) id(target) section anchorText
```

id(current) リンクが記述されているエントリの ID。

id(target) リンク先のエントリの ID。

section エントリ中のセクションへのリンクの場合、セクションの名前。セクションへのリンクでない場合は空文字列。また同一エントリ中のセクションへのリンクの場合、id(current) と id(target) は同じになる。

anchorText アンカーテキストが設定されている場合は、その文字列。アンカーテキストが設定されていない場合は空文字列。

lang

```
id title1 title2 title3...
```

id エントリの ID。

title 言語間リンク先。言語用のプレフィックスがタイトルの前に付与される。1 レコードに連続して記述していく。

isbn

```
id isbn
```

id エントリの ID。

isbn ISBN コード。ISBN コードには 10 桁の旧規格と 13 桁の新規格があるが、10 桁のコードは自動的に 13 桁に変換される。

size

```
id size
```

id エントリの ID。

size エントリの文字列の長さ。Wikipedia のタグ付けを排除した文字列の長さとなる。

2.4 生成データの実例

図 1 は日本語版 Wikipedia に存在するエントリ構造の一部を例として抜き出したものである。「日本」と「アメリカ合衆国」という記事が存在し、それぞれ「G8 加盟国」というカテゴリに属している。さらに「日本」はカテゴリ「島国」にも属している。「日本」には「ニッポン」、「アメリカ合衆国」には「USA」、「米国」というリダイレクトが設定されている。

このようなエントリ構造を表現する node ファイルと edge ファイルは次のようになる。

node

```
1554722 日本 leaf
1026301 アメリカ合衆国 leaf
621088 category:G8 加盟国 node
196090 category:島国 node
1271943 ニッポン redirect
45407 USA redirect
2389 米国 redirect
```

edge

```
1554722 621088 hypernym
1554722 196090 hypernym
1026301 621088 hypernym
1271943 1554722 target
45407 1026301 target
2389 1026301 target
```

2.5 Wik-IE の特徴

Hadoop を使った実装

Wik-IE の実装には Apache Hadoop[2] を利用している。Hadoop は Apache で開発されている分散処理および分散ファイルシステムの Java による実装で、オープンソースで公開されている。英語版 Wikipedia の配布データは全言語版の中で最も大きく、2009 年 1 月時点

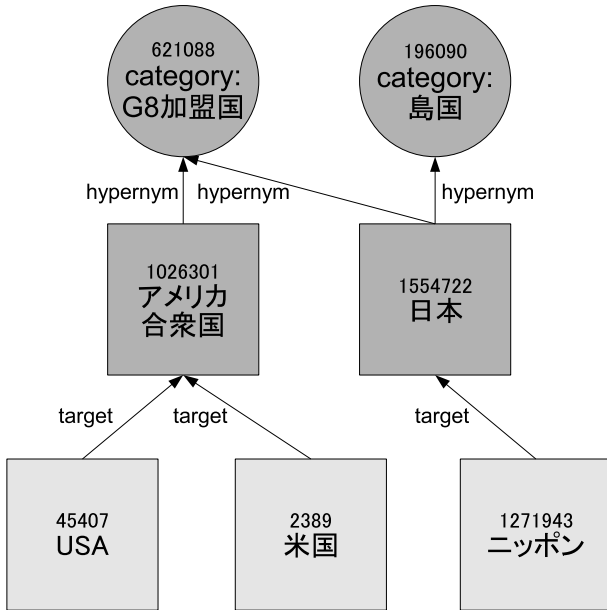


図 1: Wikipedia エントリ構造の例

の最新ファイルでおよそ 40GB あるが、Hadoop を使った分散処理によって高速に処理できる。Hadoop がインストールされていない環境でも Hadoop のライブラリを組み込んだスタンドアロン版 Wik-IE は 1 台の PC 上で動作させることができる。

機能の追加

Hadoop は MapReduce アルゴリズムでデータを処理する。MapReduce アルゴリズムとは、Map と Reduce という 2 段階に分けられた工程でデータを処理するアルゴリズムである。Mapper と Reducer はそれぞれキーと値のペアが入力と出力になっている。Mapper へは入力ファイル (Wik-IE なら Wikipedia の XML ファイル) を切り分けて入力ペアが与えられる。入力ファイルをどのように切り分けて入力ペアを作るかはクラスファイルによって定義されるが、Wik-IE では値を Wikipedia エントリ 1 個分の XML 要素、キーをその要素の XML ファイル内の位置としている。このエントリ 1 個分の要素は page という名前が付けられている。Mapper は 1 個ずつ与えられる page 要素を解析して新しい出力ペアを作り Reducer へ与える。Reducer はキー毎にまとめられた入力ペアを処理し出力となるペアを作る。Reducer の出力ペアが最終的にファイルとして出力されるときは形式もクラスファイルで定義されるが、Wik-IE では 1 つのペアを 1 行とし、キーと値の文字列をタブで区切ったテキストファイルとしている。

Wik-IE はこのように Wikipedia の XML ファイルを解析しているが、この工程で使う Mapper と Reducer

を独自に定義すれば後から機能を追加することもできる (図 2)。Wik-IE は標準でいくつかの機能を持っているが、これは、それらの機能を実現するクラスファイルを Wik-IE にあらかじめ組み込んであるということである。

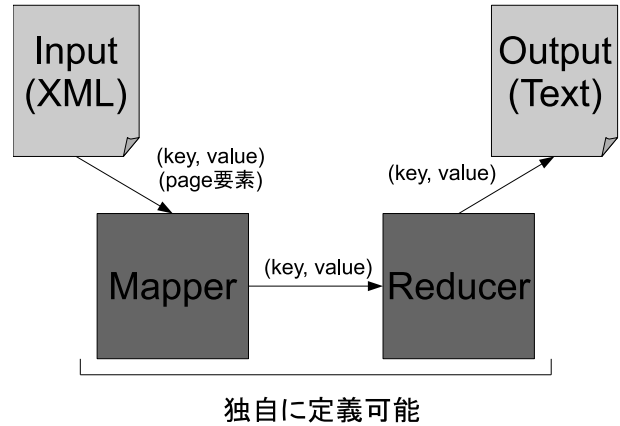


図 2: Wik-IE のイメージ

実行が容易

通常 Wikipedia で配布されている XML ファイルは専用ツールを使って SQL ダンプに変換した後にデータベースに格納して利用する [1]。Wik-IE はデータベースを使用せず、直接 XML ファイルを解析して情報を抽出する。そのため実行が容易で XML ファイルと機能を指定するだけで目当ての情報が抽出できる。

全言語版に利用可能

Wikipedia は世界中の言語で展開されていて、2009 年 1 月の時点で 264 の言語版の Wikipedia が存在する。Wik-IE の処理は特定の言語版のデータに依存しておらず、全言語版のデータに利用可能である。

2.6 配布

Wik-IE は Web 上で公開されている。
<http://wik-ie.sourceforge.jp/>

3 今後の課題

Wik-IE で使う Wikipedia の XML ファイルはサイズが大きく、ダウンロードや圧縮ファイルの展開にも時間がかかる。また、Wik-IE のユーザの目的はあくまで Wik-IE によって抽出できるデータを利用すること

が目的であり，自分の PC 上で Wik-IE を動作させなくても必要なデータが得られるなら便利である．そこで Wik-IE の生成データを配布するか，生成データにアクセスする仕組みを検討中である．

参考文献

- [1] Wikipedia:データベースダウンロード
<http://ja.wikipedia.org/wiki/Wikipedia:データベースダウンロード>
- [2] Apache Hadoop
<http://hadoop.apache.org/>