NLP若手の会第3回シンポジウム
ポスターセッション発表

Wikipedia解析ツール
Wik-IE

森 竜也、増田 英孝 (東京電機大学)
清田 陽司、中川 裕志 (東京大学)

作成の背景

- 近年Wikipediaを辞書・シソーラス作成の情報源として用いることが注目されている。
- 「車輪の再発明」状態の発生
- Wikipediaにはカテゴリやリダイレクト関係があり、普通のWebサイトより構造的な情報が豊富である。
- Wikipediaでは全ページのデータファイルを配布している。
- 手軽にWikipediaのデータを利用できるようなツールがあれば便利である。

概要

- 目的
  - Wikipediaで配布されているデータファイルから各種データを抽出し、容易に利用できるようにするツールを作成する
- データ例
  - 記事のタイトルや種類
  - 記事とカテゴリ間の関係
  - ページ間リンク
  - 言語間リンク　など

Wik-iEでできること

- 「地上デジタルテレビジョン放送の同義語は？」
- 「日本の都道府県は韓国語でなんと書くか？」

Wikipediaの現状

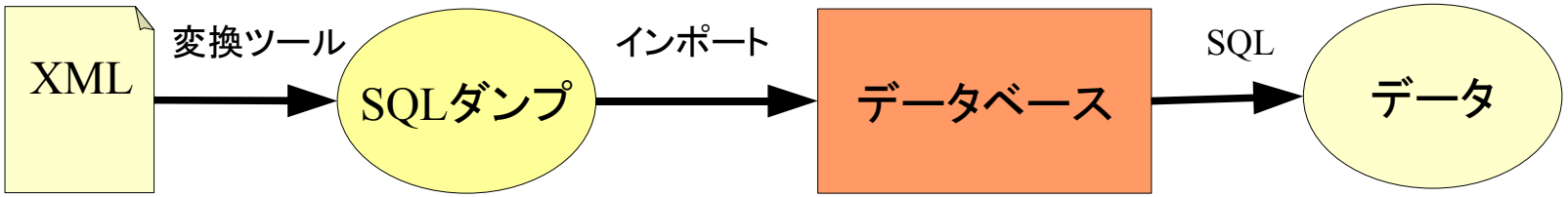
- 日本語版Wikipedia
  - 記事数約50万件
  - カテゴリー数約5万個
  - リダイレクト数約30万個
- 英語版Wikipedia
  - 記事数約250万件
  - カテゴリー数約40万個
  - リダイレクト数約580万個

特長

- 特長
  - データベースを介さずに直接解析
  - Hadoopプラットフォーム上での分散処理
  - スタンドアロンでも動作可能（その場合必要なのはJavaの実行環境のみ）
  - 全言語版Wikipediaに使用可能
- Hadoop
  - Javaで書かれたオープンソースソフトウェア
  - 分散コンピューティング＆ファイルシステム
  - MapReduceアルゴリズム

データベースとの比較

データベース



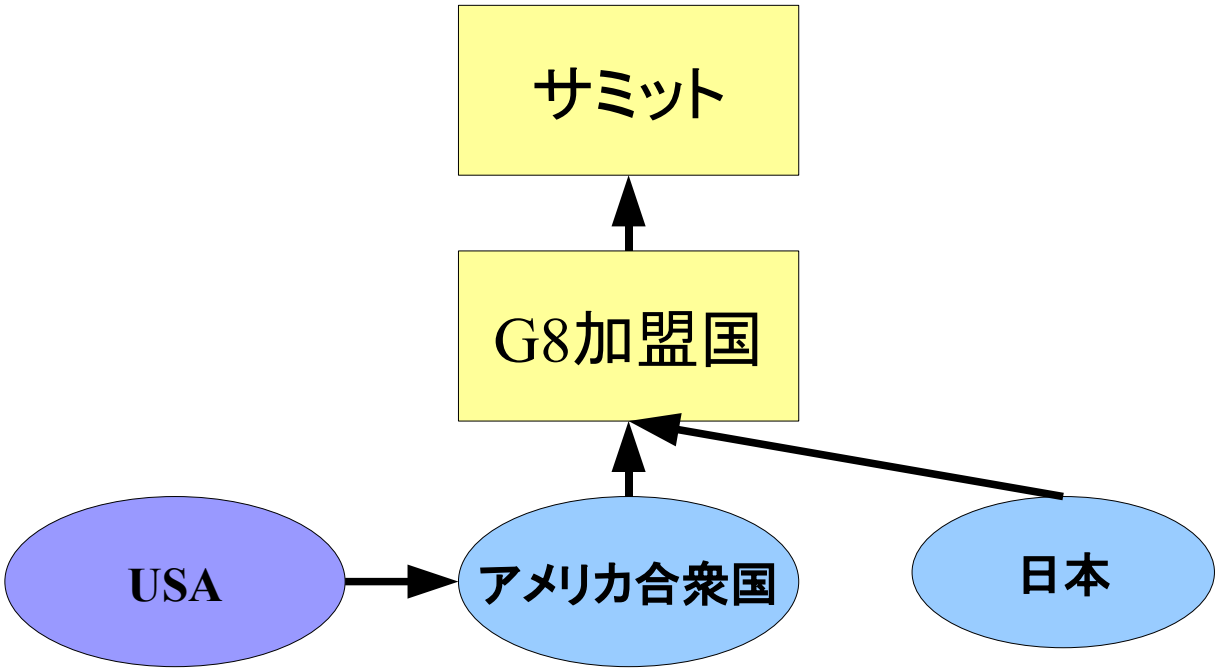
- 処理工程が多い
- 時間がかかる
- 分散不可

Wik-IE

- 処理工程が少ない
- スレーブを増やすほど高速化
- 分散可能

生成されるデータ

- XMLファイルを解析し、タブ区切り(TSV)形式のテキストファイルを生成する。
- TSVファイルは再利用が容易
- 例



node

- ページのタイトルと種類を表したデータ
- 例

Id	title	kind
28	日本	leaf
1026301	アメリカ合衆国	leaf
621088	category:G8加盟国	node
608926	category:サミット	node
45407	USA	redirect

edge

- 記事・カテゴリ・リダイレクト間の関係を表したデータ
- 例

Id(current)	id(target)	relation
28	621088	hypernym
1026301	621088	hypernym
621088	608926	hypernym
45407	1026301	target

interWiki

- 言語間リンクを表したデータ
- 例

Id	title1	title2	title3	...
28	en:Japan	fr:Japon	ru:Япония	
1026301	en:United States	fr:États-Unis	ru:Соединённые Штаты Америки	

variation

- ページ間リンクのアンカーテキストとリンク先エントリのタイトルから取得した表記ゆれ
- 例

title	variation
日本	日本国
日本	JAPAN
日本	ニッポン
日本	ジャパン

isbn

- 記事中に参考文献として記述されている出版物のISBNコード
- 例

Id	isbn
28	9784130561013

- Wikipediaには参考文献とそのISBNコードを記述する方針があるが、言語によって記述されている割合に差がある。日本語版は記述されている割合が低いほう。

各国語版の参考文献

対象	英語版	日本語版	ドイツ語版	フランス語版
記事数(redirect除く)	2,299,977	483,967	770,416	669,740
外部リンクの記述がある記事	42%	39%	52%	30%
参考文献の記述がある記事数	34%	6%	20%	10%
ISBNコードの記述がある記事数	6%	5%	14%	3%

ISBNコードの記述率はドイツ語版が突出

history

- 過去の編集頻度を週単位で集計したデータ
- あるエントリがどんな時に多く編集されているか分かる

title	date	frequency
日本	2008-6-30	9
日本	2008-7-7	13
日本	2008-7-14	5
日本	2008-7-21	4

利用例1

- 「地上デジタルテレビジョン放送の同義語は？」
- edgeとvariationを使う
- リダイレクトと表記ゆれから同義語を取得
  - 地上デジタル
  - 地上デジタルテレビ
  - 地上波デジタル放送
  - 地デジ
  - 地上波ディジタル放送…など

利用例2

- 「日本の都道府県は韓国語でなんと書くか？」
- node, edge, interWikiを使う
- Category:日本の都道府県の下にあるエントリを取得し、その言語間リンクを調べる。
  - 北海道　ko:홋카이도
  - 沖縄県　ko:오키나와 현
  - 東京都　ko:도쿄 도
  - 京都府　ko:교토 부
  - 大阪府　ko:오사카 부...など

実行時間

- 日本語版データファイル
- 使用マシン
  - OS: CentOS 5
  - CPU: Xeon 3060 2.40GHz
  - メモリ: 4GB
  - マスタ1台、スレーブ3台
- 前に挙げた全ファイルを生成: 約20分
- nodeとedgeファイルの生成: 約7分

おわりに

- 今までに挙げた機能は実装済み
- 以下のURLで公開中
- https://sourceforge.jp/projects/wik-ie/
- 自由にご利用ください
- みなさんの意見をください
  - こういったデータを取り出したい
  - こんな機能がほしい
  - Wikipedia以外のリソースを利用したい