DA-01

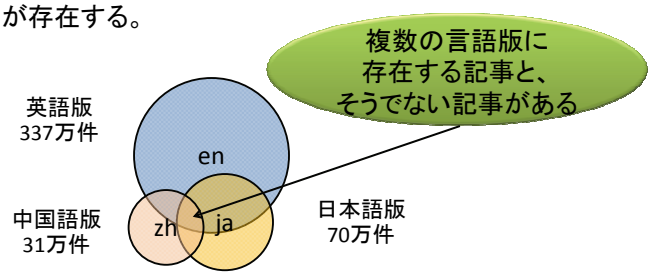
森竜也　増田英孝（東京電機大学）　中川裕志　清田陽司（東京大学）

言語間差異を活用したWeb情報資源へのアクセスシステムに関する研究

はじめに

•一般のユーザが情報を容易に発信できるシステムの登場
　→**ソーシャルメディア、ユーザ参加型コンテンツ**

•例えば、百科事典サイトWikipedia
　＊250以上の言語版
　＊全言語版の総記事数1000万件以上
　＊ただし、異なる言語で同じ事柄について書かれている記事が存在する。

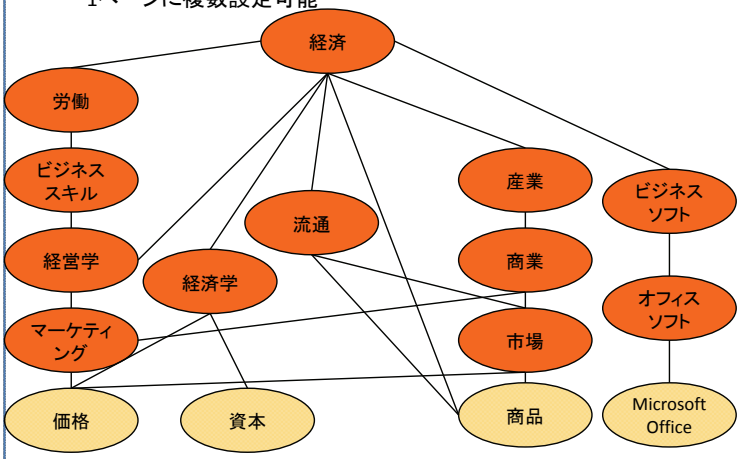


•複数の言語版でのみ書かれている概念・言葉
•多くの言語版で書かれている概念・言葉
　＊言葉の重要度や、国や文化を超えた普遍的な価値などが異なる

•**目的**:　ある話題に関する記事が、各言語版でどれだけ書かれているか分析、比較をする
•**目標**:　外国ではすでに注目を集めている話題をいちはやく察知する

Wikipediaの構造

•記事とカテゴリからなるグラフ構造
　＊記事：通常のページ
　＊カテゴリ：ページを分類するためのページ
　　・自由な文字列によるタグ付け
　　・1ページに複数設定可能



他の言語版に存在する同じ事柄について書かれている記事は**言語間リンク**という特殊なリンクで接続されている

ページ探索

•始点となるカテゴリを指定し、**関連度**の低いカテゴリを排除しつつ、未出現のページがなくなるまで、再帰的にグラフ構造を探索し、記事を数え上げる。
　＊**関連度**の指標を以下の式で得られる文字列の類似度とする
　＊閾値を実験的に0.25に設定

$$\text{関連度} = \frac{|s(b(T_1), b(T_2))|}{\sqrt{|b(T_1)| \cdot |b(T_2)|}}$$

T1, T2: タイトル
b: 文字列のbi-gram集合を得る
| |: 与えられた集合の要素数を得る
s: 与えられた2つの集合のうち、共通の要素を得る

　＊関連度高：経済、経済学
　＊関連度低：経済、ビジネスソフト

•**日本語版だけでも数万ノードからなるグラフ構造を構築するために並列システムを利用**
•同様の探索を言語間リンク先のカテゴリに対しても行う

探索結果の例

	北京オリンピック	マラソン	トライアスロン
ドイツ語	267	221	**301**
フランス語	**2396**	203	65
日本語	182	**312**	34
中国語	458	25	0

勝てる競技が人気があり、記事が書かれる
フランスではオリンピック記事が盛んに書かれている

おわりに

•Wikipediaの記事が書かれる傾向には、言語版によって確かに違いがあることが分かった
•タイトル文字列の類似度のみを関連度の指標にするには精度が低いことも判明した
•リンクテキストやテンプレート、一覧ページなどの他の特徴データからページ間の関連度を算出することを目指す

•Wikipediaの記事からこれらの情報を取り出す仕組みが必要
•Wikipediaが採用しているWikiシステムMediaWikiはWebブラウザ上にHTMLを出力する機能に特化
•パーサおよび出力部分のプログラムを改造・作成し、汎用的なXML文書として表示するAPIを作成中

API出力例

```
<?xml version="1.0" encoding="UTF-8"?>
<wiki>
<title>Hoge</title>
<body>
<p>本文<a href="/mediawiki/index.php/テキスト" title="テキスト">テキスト</a></p>
<template name="テンプレート名">
  <argument name="引数名1">引数値1</argument>
  <argument name="引数名2">引数値2</argument>
</template>
</body>
<categories>
  <category name="カテゴリ1" />
  <category name="カテゴリ2" />
</categories>
<interlangs>
  <interlang lang="en" name="Hoge" />
  <interlang lang="de" name="Hoge" />
</interlangs>
</wiki>
```