コンテンツの質を考慮した多言語Wikipedia記事の差異情報抽出手法の提案

甲南大学　藤原裕也
名古屋大学　鈴木優
甲南大学　小西幸男
甲南大学　灘本明代

背景1

Wikipedia

特徴

- 複数のユーザがコンテンツ育成
- 250を超える言語版が存在

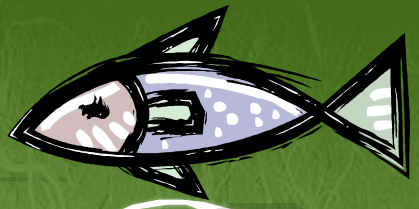
各言語版は異なる成長を遂げている

言語版によって書いてあることが違う

問題

自国の言語版だけでは情報が不足

例:イギリスの文化



検索Query:フィッシュ・アンド・チップス

日本語版

目次

1 概略
2 歴史
3 食べ方
4 関連項目

英語版

Contents

1 History
  1.1 England
  1.2 Scotland
  1.3 Ireland
2 Composition
  2.1 Cooking
  2.2 Thickness
  2.3 Batter
  2.4 Choice of fish
  2.5 Accompaniments
3 Vendors
4 Cultural impact
5 Environment
6 See also
7 Footnotes
8 External links

日本人ユーザ

背景2

- ユーザは母国語のサイトを調べる場合が多い
- ユーザが母国語以外を読むことが可能
  - 理解に時間がかかる
  - 例
    - 日本人ユーザ => 「Fish and chips」
    - 外国人ユーザ => 「平等院」

Etc…

問題

母国語以外の記事を全て読むことは困難

背景3

- Web上の情報は必ずしも質の良いもので構成されてない
- 嘘の情報も存在
  - Ex：口コミサイトで嘘のクチコミによる店の評価やランキングの操作

    →誤った情報を提示することは社会問題に発展

質のある情報を提示することは有益

目的

母国語版

提示

他の言語版

多言語Wikipediaの差分情報を抽出，提示するシステム提案

全体の流れ

比較対象記事の決定

- リンク構造
- 関連度

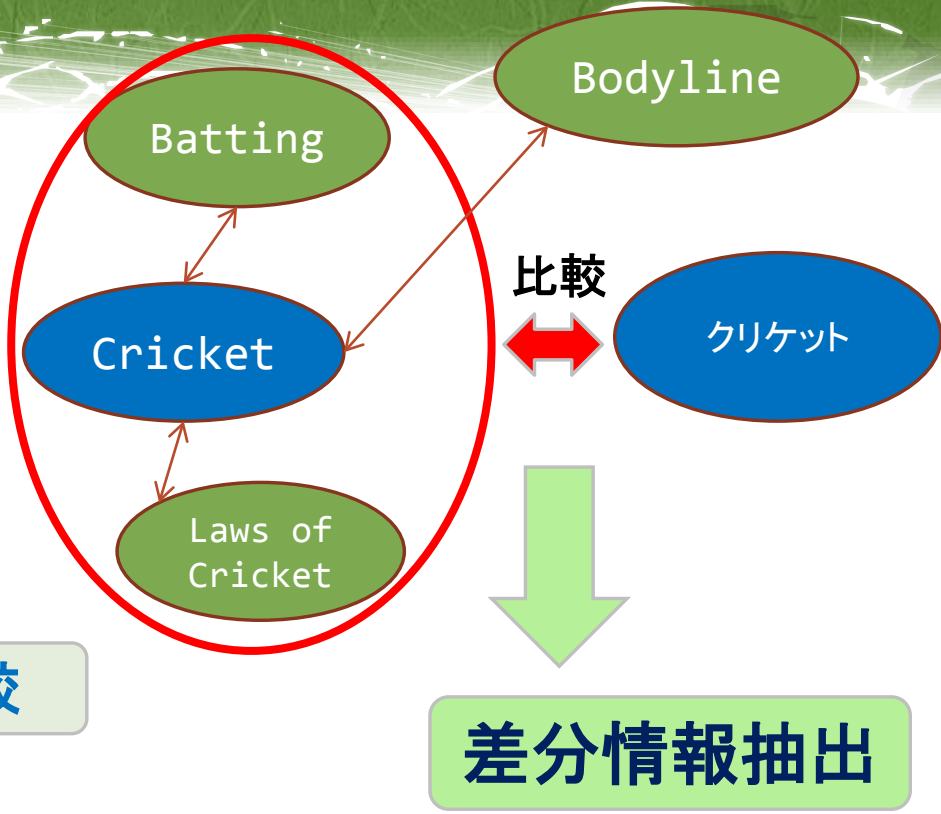
差分情報抽出

- コンテンツの比較

情報の質の計算

- 記事の残留度

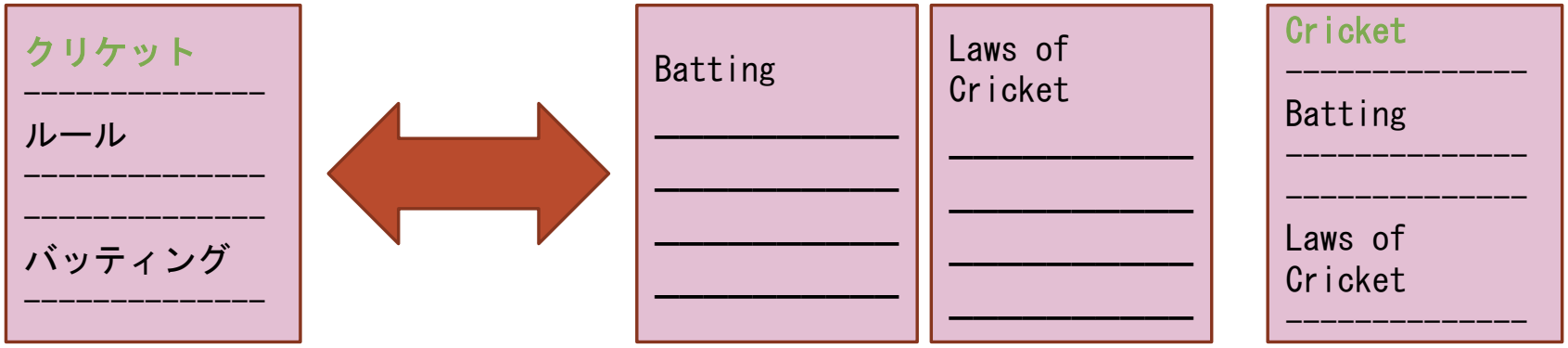
差分情報の提示



比較対象Wikipediaの記事の決定

- 言語や文化の違いから情報の粒度が異なる
  →対応する記事が複数にまたがる場合がある

  - Ex:「クリケット」
    - 日本語版:バッティングやルールの説明が含まれている
    - 英語版:Batting、 Laws of Cricketの記事が各々存在する



複数ページと比較して差分情報を抽出する必要がある

比較対象Wikipediaの記事の決定

リンク構造解析

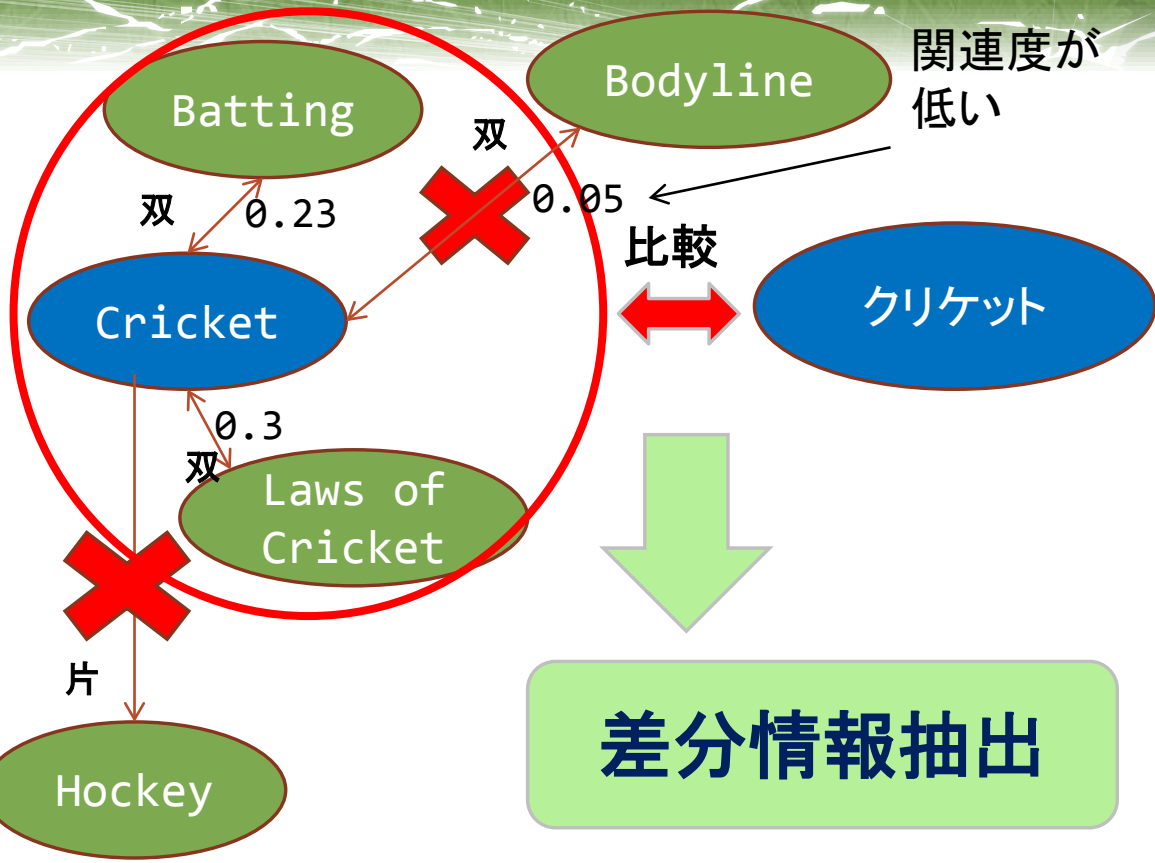
関連づけられた
記事をつないでるため

リンクグラフの生成

強連結：関連性が強い

関連度計算

比較対象ページの取得



比較対象Wikipediaの記事の決定

- 過去の手法
  - リンクグラフのノード間をCos類似度で計算
    - 適合率：35%，再現率：49%，F値：41

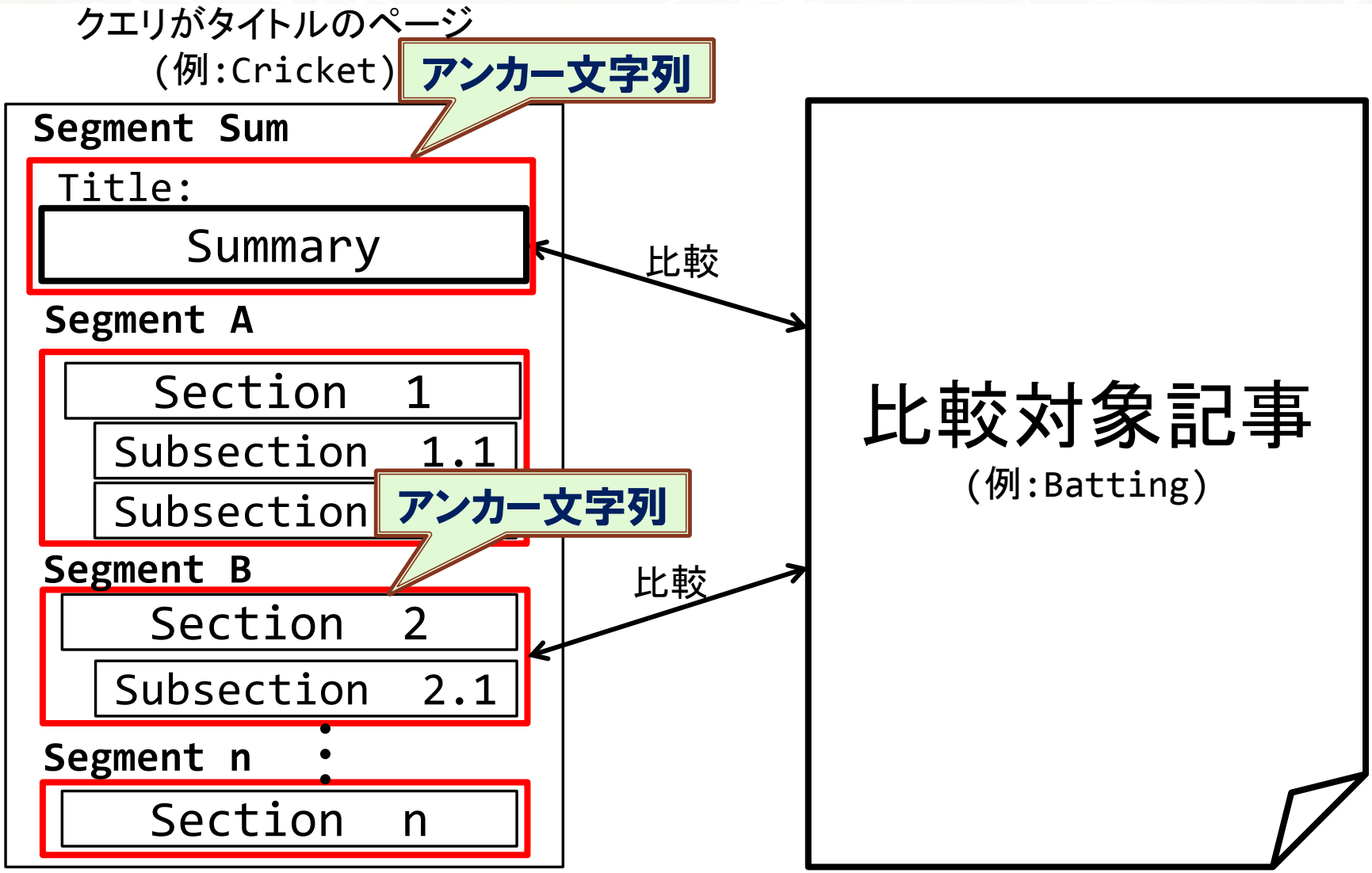
⇒精度が低かった

双方向リンク注目した
記事と記事との関連する度合い

関連度

- **アンカー文字列の出現位置**
  サマリにリンクを張っている記事は関連性が高い
- **アンカー文字列の出現回数**
  記事に何度も出現するアンカー文字列関連性が高い
- **部分的な情報との類似性**

関連度


クエリがタイトルのページ
(例:Cricket)
アンカー文字列
Segment Sum
Title:
Summary
Segment A
Section 1
Subsection 1.1
Subsection
アンカー文字列
Segment B
Section 2
Subsection 2.1
Segment n
Section n
比較
比較
比較対象記事
(例:Batting)

全体の流れ

比較対象記事の決定

- リンク構造
- 関連度

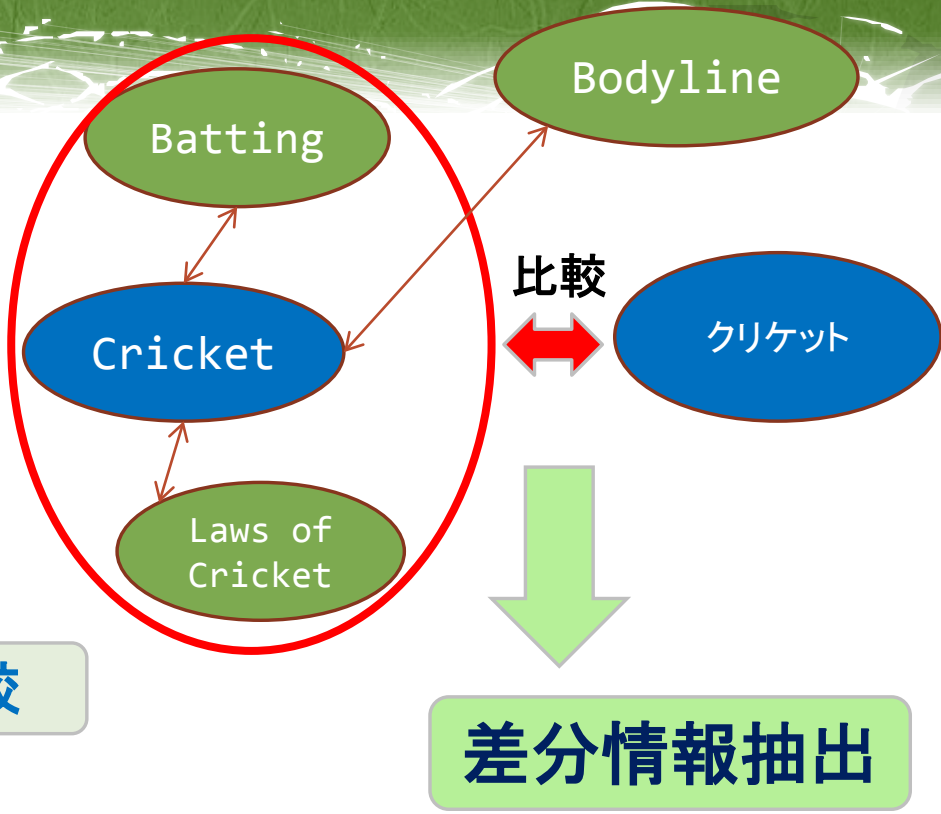
差分情報抽出

- コンテンツの比較

情報の質の計算

- 記事の残留度

差分情報の提示

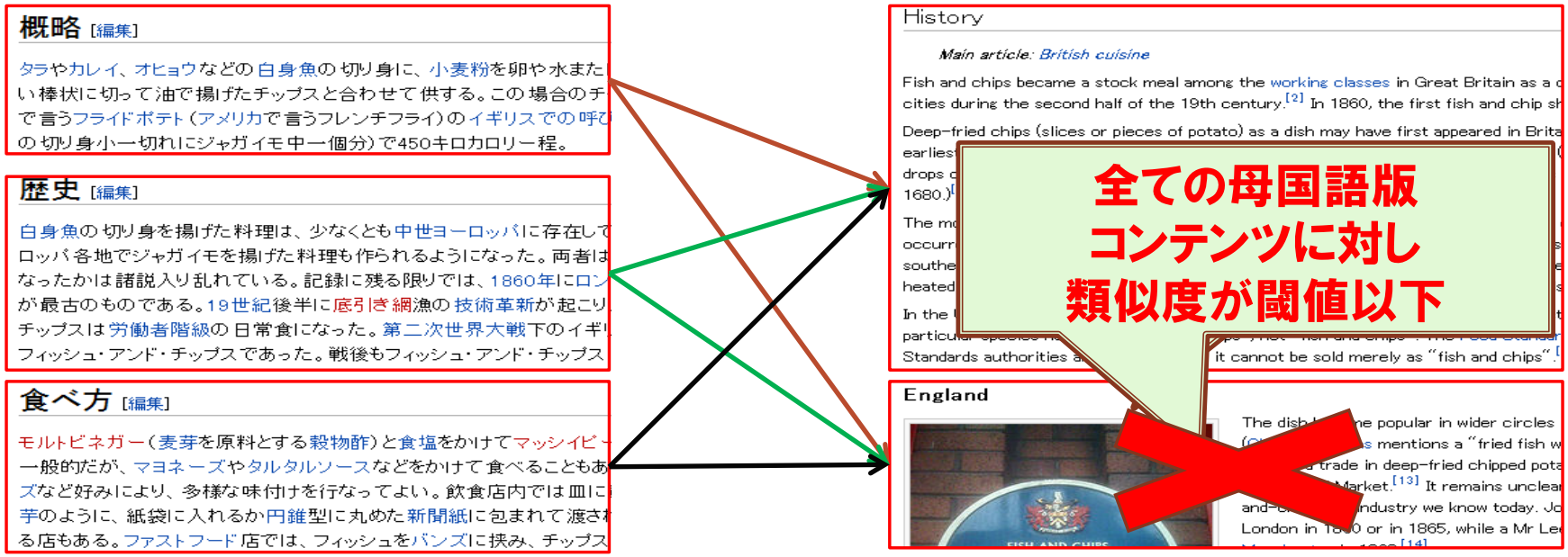


コンテンツの比較

・Wikipediaの記事は構造に基づいて段落に分かれている
　→意味的に分かれている可能性が高い

・各々の段落におけるコンテンツ同士の類似度を求める
・全てのコンテンツに対し**ある閾値以下**である段落を差分情報として抽出

例：フィッシュ・アンド・チップス



全体の流れ

比較対象記事の決定

- リンク構造
- 関連度

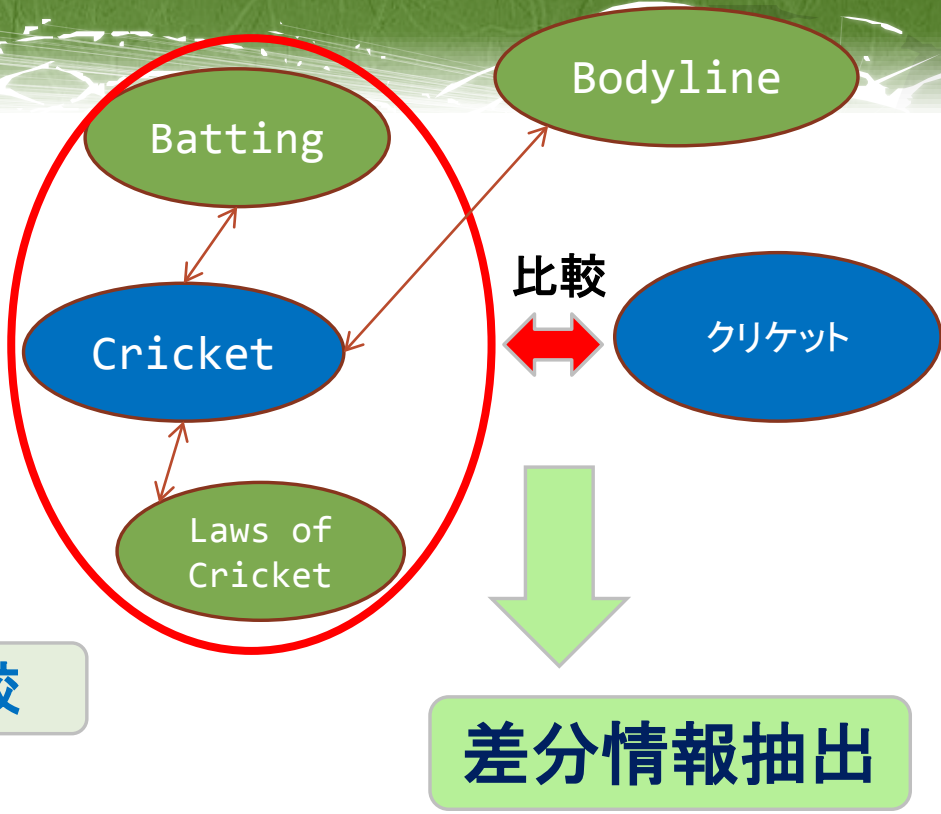
差分情報抽出

- コンテンツの比較

情報の質の計算

- 記事の残留度

差分情報の提示



情報の質の計算

- 得られた差分情報は必ずしも有益な情報であるとは限らない
  - Wikipediaの記事には質の低い情報が混在

記事の残留度に基づく質の算出方法

我々の提案する記事の残留度に基づく質の算出方法を用いて，得られた差分情報の質を計算を行う

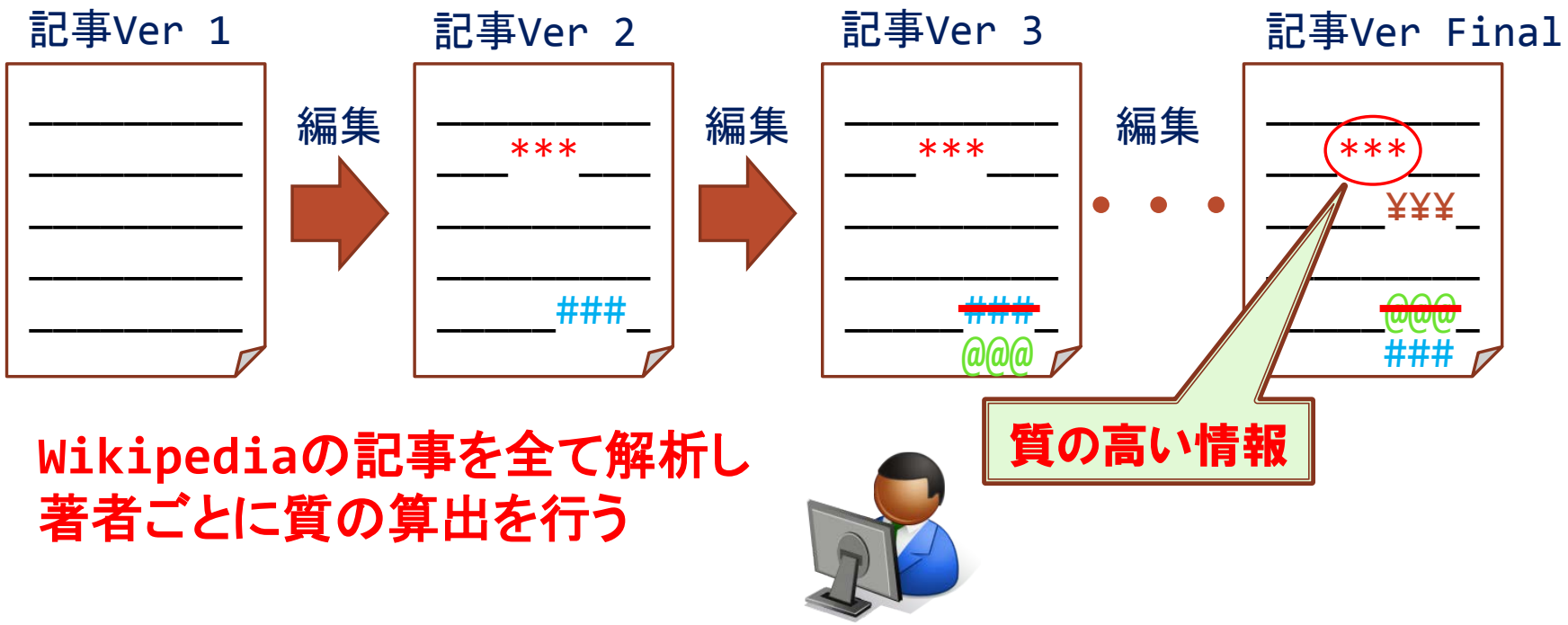
⇒質の高い情報のみユーザに提示

参考文献：鈴木優，吉川正俊，“Wikipedia におけるキーパーソン抽出による信頼度算出精度及び速度の改善”
情報処理学会論文誌 データベース Vol.3 No.3 20-32 (Sep. 2010)

情報の質の算出

- ある著者が以前の著者の記述をどれだけ残したのかによって情報の質を算出

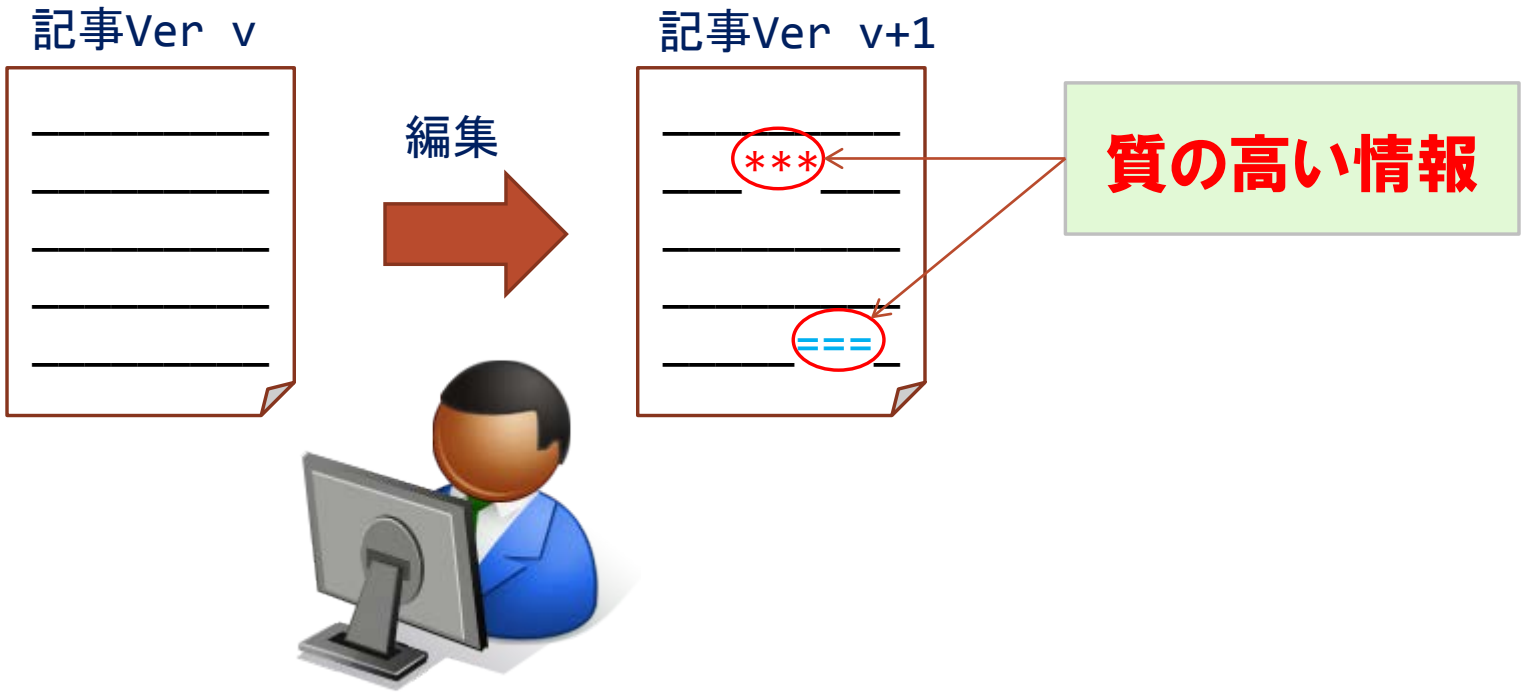
  →残っている時間が長い情報ほど著者達が残留すべき情報であると判断したため



Wikipediaの記事を全て解析し著者ごとに質の算出を行う

情報の質の算出

- 質の高い情報を記述してきたユーザの記述する情報は質が高い



実験1:比較対象記事の決定

- 比較対象記事の決定
- 差分情報抽出
- 情報の質の計算

- 比較対象記事決定の精度を測った
  - 実験内容
    - 関連度とCos類似度
    - 再現率，適合率，F値を比較
    - 比較言語版
      - 日本語版
      - 英語版
    - 条件
      - α=1～10を1刻み
      - 閾値0～1を0.05刻み

クエリ
バノック
ウォリック城
ブラックドッグ(亡霊)
フィッシュ･アンド･チップス
グッドウッド･フェスティバル･オブ･スピード
ローンボウルズ
ブルー･プラーク
バーレスク
スコットランドの国旗
ゲーリックハンドボール
キッパー(魚料理)
スコットランド国際美術館
リプトン

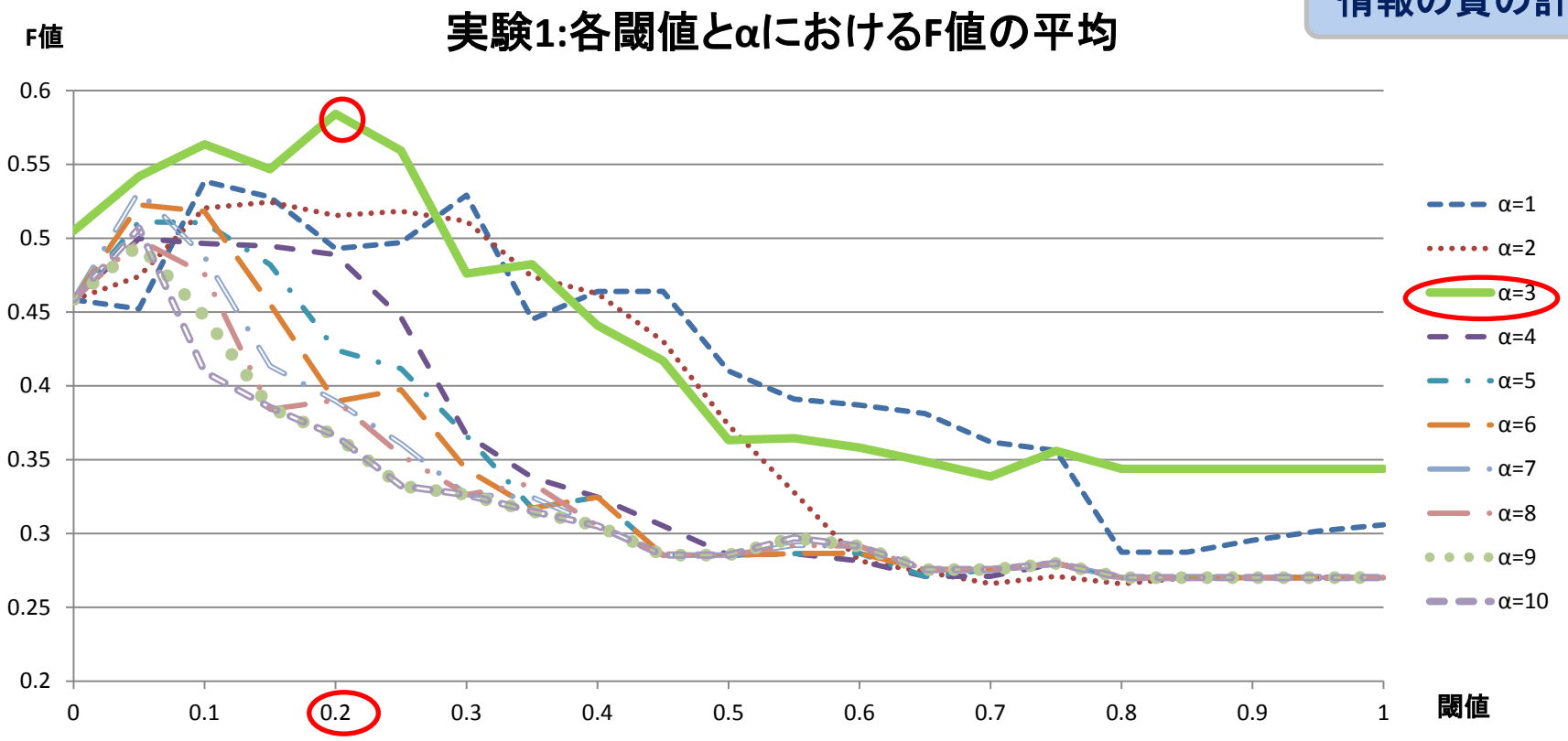
$$R_{kl} = \{\alpha \cdot (tf_{sum} \cdot S_{sum}) + \sum_{i=1}^{n} (tf_i \cdot S_i)\} / \max(R_{lm})$$

実験1:比較対象記事の決定

比較対象記事の決定

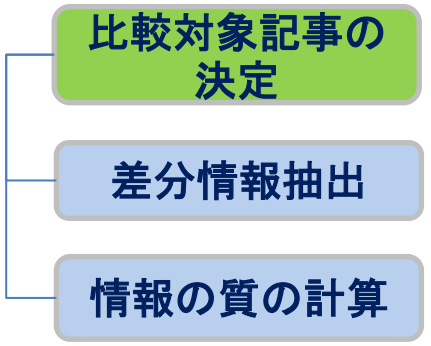
差分情報抽出

情報の質の計算



グラフよりα=3で閾値が0.2の時に最も高いF値を得ることができた
よってα=3、閾値を0.2と設定する

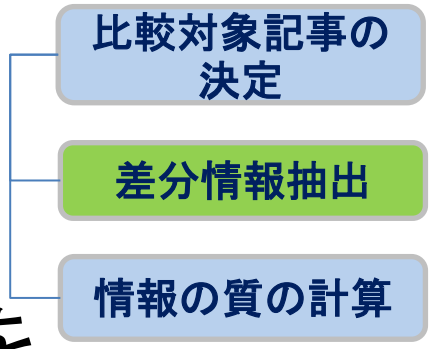
実験1:比較対象記事の決定



- 結果
  - 適合率:35%→57%
  - 再現率:49%→59%
  - F値　　:41 →58

- 考察
  - コサイン類似度では類似性に注目したため値が低かった
    - ex:Gaelic handballとGaelic handballの理事会の場合は理事会はGaelic handballの競技の説明をしているわけではない
  - 部分的な情報と類似性
    - Gaelic handballの記事の中にあるGaelic handballの理事会を説明している部分とGaelic handballの理事会の記事を比較

実験2:差分情報抽出



- 得られた比較対象記事と母国語記事を用いて差分情報抽出を行った最適な閾値を求めた
  - 内容
    - コンテンツの比較の際の閾値
    - 条件
      - 閾値0～1を0.05刻み
      - 差分情報の適合率、再現率、F値
    - 差分情報:母国語版にない情報
    - 比較言語
      - 日本語版
      - 英語版

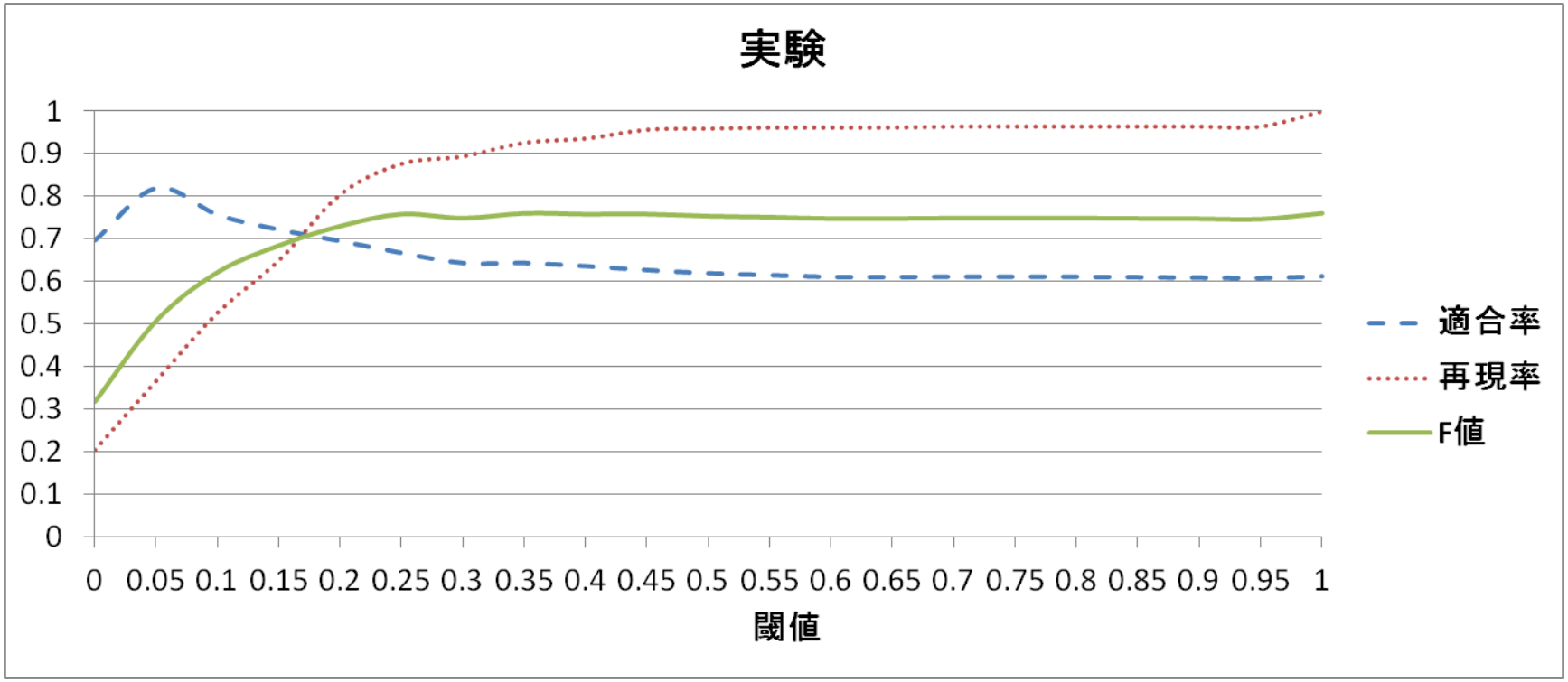
クエリ
バノック
ウォリック城
ブラックドッグ(亡霊)
フィッシュ･アンド･チップス
グッドウッド･フェスティバル･オブ･スピード
ローンボウルズ
ブルー･プラーク
バーレスク
スコットランドの国旗
ゲーリックハンドボール
キッパー(魚料理)
スコットランド国際美術館
リプトン

実験2:差分情報抽出

比較対象記事の決定

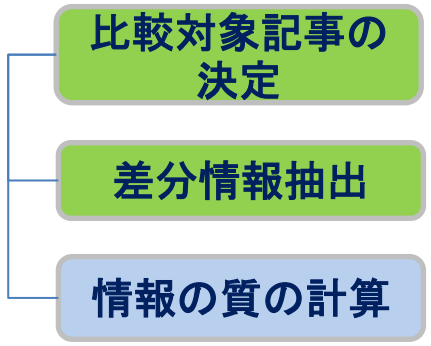
差分情報抽出

情報の質の計算



図より閾値が0.2の時に適合率，再現率が交わり共に高い値となった
これにより，本研究では閾値を0.2と設定し差分情報の抽出を行う

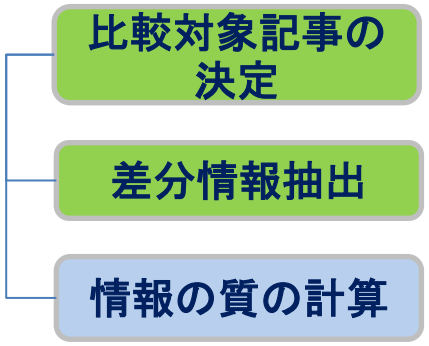
実験3:差分情報抽出



- 提案手法の有用性を示すために評価実験を行った
  - 内容
    - 差分情報の適合率、再現率、F値
    - 差分情報:母国語版にない情報
    - 閾値:0.2
    - 比較言語
      - 日本語版
      - 英語版

クエリ
バノック
ウォリック城
ブラックドッグ(亡霊)
フィッシュ･アンド･チップス
グッドウッド･フェスティバル･オブ･スピード
ローンボウルズ
ブルー･プラーク
バーレスク
スコットランドの国旗
ゲーリックハンドボール
キッパー(魚料理)
スコットランド国際美術館
リプトン

実験3:差分情報抽出



- 結果
  - 適合率:70%，再現率:81%，F値:73
- 考察
  - 結果の良いもの
    - リプトンであれば日本語版に存在しないイエローラベルの製作者や歴史についての情報が抽出された.
  - 結果の悪いもの
    - ブラックドッグなどは関連ページの一部の情報のみが関連する差分情報であるが，その関連ページの残り情報はクエリと関係のない差分情報が抽出される場合が存在した.

まとめと今後の課題

- まとめ
  - 複数の言語版に対しWikipedia上での差分情報を取得する手法の提案を行った
  - 提案手法
    - 比較対象記事の決定→関連度
    - 差分情報抽出→コンテンツの比較
    - 情報の質の計算→記事の残留度に基づく質の算出
- 今後の課題
  - 情報の質の実験
  - 差分情報の提示方法
  - 差分情報抽出方法