Vol.2011-DBS-152 No.3
Vol.2011-IFAT-103 No.3
2011/8/2

Wikipediaの言語間比較による差異情報抽出手法の提案

藤原裕也[†1]　灘本明代[†2]

フリー百科事典 Wikipedia は日本語版，英語版，などの各言語版によって記述されている情報量が異なる．故に自国の言語版だけでは情報量が不足する記事も存在する．特に文化の記事は顕著である．そこで本研究では多言語 Wikipedia を用いて，日本と海外における日本の伝統文化の紹介の差異情報を抽出し提示するシステムを提案する．具体的には，ユーザの入力した日本の伝統文化の日本語 Wikipedia と英語 Wikipedia を比較しその差分情報を提示する．この時，情報の粒度の違いから，複数の日本語 Wikipedia の記事と英語 Wikipedia の記事を比較対象とする．本論文では，この比較対象の範囲を Wikipedia のリンク構造を用いて決定する手法の提案及び多言語間の差分情報抽出手法の提案を行う．

Extracting of Difference Information between Multilingual Wikipedia

Yuya Fujiwara[†1] and Akiyo Nadamoto[†2]

Theare are multilingual articles on the Wikipedia. The information between multilingual articles is different. Especially, the case of the articles which is written about culture is very different between languages. In this paper, we propose the system which extracts different cultural information between Japan and other countries on the Wikipedia. As a first step of the research, we compare the Japanese cultural article between Japanese article and English article, and extract different information from them.

†1 甲南大学大学院 自然科学研究科
Konan University
†2 甲南大学 知能情報学部
Konan University

1. はじめに

Web 上で自由に編集が可能なフリー百科事典である Wikipedia[⋆1]は気軽に様々な事が調べられることから，多くの人々が利用している．Wikipedia が通常の百科事典と大きく異なる点はエンドユーザが記事を作成しコンテンツを育成，現在では 250 を超える言語版が存在しているなどが挙げられる．各言語版は他の言語と独立して管理されており，各々の言語のユーザが記事の作成，追記，修正が可能であるためある事柄についての記事が特定の言語版のみに存在していたり，ある事柄に関する記事が他の言語版と比較して多く作成されている言語版が存在するなど，各々異なる成長を遂げている．例えば，日本の伝統的な履物である「下駄」を日本語版 Wikipedia と英語版 Wikipedia とで見比べると，書いてある目次の項目やその数，記事の内容等が異なることがわかる．日本語版の目次項目の一つである「下駄のつく言葉」は日本語版だけに存在し，逆に英語版の目次項目の一つである「Geisha」は英語版だけに存在する (**図 1** 左参照)．また，逆にイギリスの食べ物である「フィッシュアンドチップス」を英語版と日本語版で比べてみると日本語版のほうが目次の量が少ない，それに対し英語版は自国の言語圏内である故に目次の量が多く内容も詳しく書いてあることがわかる (図 1 右参照)．例えばこの「フィッシュ・アンド・チップス」の日英両方の目次に出現する「歴史」という項目の内容は，英語版にはイングランドでの歴史やスコットランドでの歴史についてかいてあるが，日本語版にはそれがなく情報が異なっており網羅の度合いが違う．さらに，国や民族，文化によって人々の興味や関心が異なるので，Wikipedia に書かれている記事の種類がジャンルによって異なっている．例えば，日本語版 Wikipedia の全閲覧数のうち 8 割が大衆文化についての記事に集中しているのに対し，英語版は 4 割である．逆に英語版は日本語版に比べると政治や地理などの記事の人気が高い[1)]．故に自国の言語版だけでは情報が不足しているという問題がある．そこで我々は，ユーザが閲覧している Wikipedia の記事の中で不足している情報が他の語版にある場合，その情報を提示するとユーザにとって便利であると考え， Wikipedia の多言語性を用いた差分情報抽出を行うシステムを提案する．具体的には，ユーザの入力したクエリの日本語 Wikipedia と英語 Wikipedia を比較しその差分情報を提示する．この時，情報の粒度の違いから，複数の日本語 Wikipedia の記事と英語 Wikipedia の記事を比較対象とする．そして得られた差分情報の信頼度，重要度を計算し，他方の言語版に加え提示する．本論文では，はじめの一歩とし

⋆1 日本語版 Wikipedia http://ja.wikipedia.org/wiki/

Vol.2011-DBS-152 No.3
Vol.2011-IFAT-103 No.3
2011/8/2

図 1 多言語 Wikipedia の例
Fig. 1 Example of multilingual Wikipedia

て，日本の伝統文化に注目し多言語 Wikipedia における日本と海外での日本の伝統文化の紹介の差異情報を抽出し提示するシステムを提案する．なお本論文では，この比較対象の範囲を Wikipedia のリンク構造を用いて決定する手法の提案及び多言語間の差分情報抽出手法の提案を行う．

以下に処理の流れを示す．

(1) ユーザは調べたい日本の伝統文化をクエリとして入力する．
(2) 入力された伝統文化のタイトルの日本語と英語の Wikipedia の記事を各々取得する．
(3) 取得した英語と日本語の記事を目次構造に基づき分割するこの最小単位をセグメント呼ぶ．
(4) 英語版の記事をセグメントごとに形態素解析を行い，名詞を取得．
(5) (4) で取得した名詞を翻訳する．
(6) 比較対象となる日本語の記事群を (2) で取得した日本語の記事のリンク構造を解析し取得する．
(7) 日本語の記事群をセグメントごとに形態素解析を行い，名詞を取得．
(8) (4) で取得した英語の記事の情報と (8) で取得した日本語の記事群の情報を目次構造におけるコンテンツの比較を行い，差分情報を取得する．
(9) 取得した差分情報をユーザに提示する．

図 2 に提案システムのフローを示す．

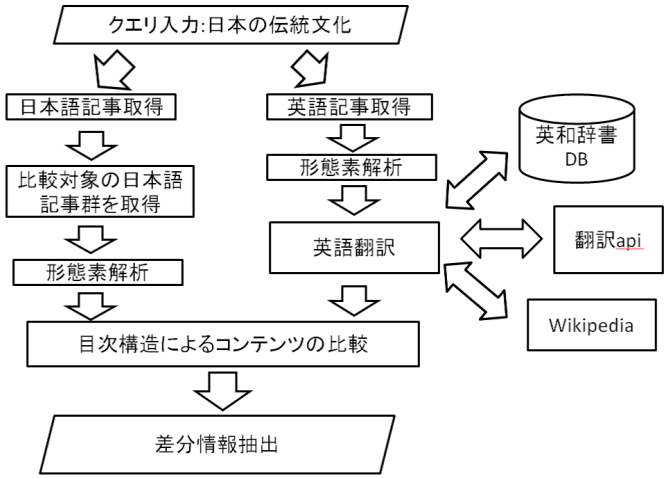

図 2 システムのフロー
Fig. 2 System Flow

以下，2 章では関連研究を，3 章では比較対象 Wikipedia の記事の抽出手法を，4 章では多言語記事の差分抽出について述べる．5 章ではプロトタイプシステムについて述べ，6 章では評価実験について述べ，7 章ではまとめと今後の課題について述べる．

2. 関 連 研 究

森ら[2] はある話題に属する記事の数の差異を抽出し興味や関心の違い抽出する手法を提案している．具体的には Wikipedia のあるカテゴリに属する記事の件数を用いて言語間の差異の抽出を行っている．下位カテゴリがなくなるまでそれと上位のカテゴリとの類似度を計算する．得られたカテゴリの数とそのカテゴリにある記事の件数をドイツ語，フランス語，日本語，中国語の Wikipedia と比較し，それぞれがどれだけそのカテゴリに興味があるのかを判定している．しかしながら，本研究では Wikipedia の記事の目次構造に基づきコンテンツを比較し，差分情報抽出する手法を提案している点が異なる．

斎藤ら[3] は日本語，英語，中国語のニュース記事から国際的な感覚の違いを抽出ためその記事の収集方法を提案している．具体的には日本語，英語，中国語のニュース記事から各々

Vol.2011-DBS-152 No.3
Vol.2011-IFAT-103 No.3
2011/8/2

キーワードを抽出し英語に翻訳し，索引化を行っている．その翻訳の際に Wikipedia の言語間リンクを用いて翻訳を行っている．本研究でも Wikipedia の言語間リンクを用いて翻訳を行っている．しかしながら，本研究では Wikipedia の記事同士を比較しその記事の差分情報抽出する手法を提案している点が異なる．

中崎ら[4] はあるトピックに対して日本語と英語のブログを比較し文化間の違いを抽出する手法を提案している．具体的にはあるトピックの日本語版，英語版 Wikipedia の記事を取得し，その本文から太字と他の記事へのリンクを関連語として抽出する．そしてあるトピックの日本語，英語のブログサイト群から関連語を用いてランキングし分析を行っている．しかしながら，本研究ではブログでの多言語比較ではなく Wikipedia の記事の目次構造の比較になっている点が異なる．

立床ら[5] は時間情報や空間情報と結び付けられたコンテンツ Wikipedia から抽出した知識ベースを用いた検索手法を提案している．具体的には Wikipedia の記事同士の関連度，そして記事に関する時間情報，空間情報の関連度の計算を行い地球科学データに応用している．この際に Wikipedia の記事同士の関連度の計算として中山ら[6] の提案した pfibf を用いている．pfibf はある記事から別の記事へのパスの多さ，そしてある記事から別の記事への最短距離，その２つ要素を考慮している．それに対し，本研究ではある記事から別の記事へのリンクの数の多さ，そして Wikipedia の目次構造におけるそのリンクの出現位置を考慮し関連度の計算を行う手法を提案している．

3. 比較対象 Wikipedia の記事の抽出手法

3.1 リンク構造解析

日本語と英語の Wikipedia の記事を比較する時，多言語 Wikipedia では言語や文化の違いから情報の粒度が異なり，対応する記事が複数にまたがる場合がある．特に日本の伝統文化の場合，英語の Wikipedia では 1 記事であるのに対し，日本語の Wikipedia では詳細に書かれており複数の記事になっている場合がある．例えば,「和歌」の場合，英語の Wikipedia では和歌の形式の１つである長歌や短歌の説明が和歌という記事１つに書いてあるのに対し，日本語の Wikipedia では和歌の記事だけでなく長歌，短歌の記事が各々存在し複数ページにまたがっている．そこで，我々は日本語の Wikipedia のリンク構造を解析する事により，比較対象の記事を抽出する．以下に抽出手順を示す．

(1) ユーザの入力したクエリと同じタイトルを持つ日本語の記事から記事をノードとしリンクをエッヂとするリンクグラフを作成する．このリンクグラフを基準リンクグラフと呼ぶ．そして基準リンクグラフ内のユーザの入力したクエリと同じタイトルを持つ記事を示すノードを基準ノードと呼ぶ．

(2) 基準ノードと双方向にリンクされているノードは，基準ノードの記事に深く関連すると考え，双方リンクされているかつ英語版に存在しないノードを残し，その他の基準ノードをリンクグラフから削除する．(**図 3**)

(3) 基準ノードと基準リンクグラフ内のその他のノードの関連度を求める．

(4) その関連度が閾値以下であるものを削除し，残ったノードの記事を比較対象の日本語記事群とする．

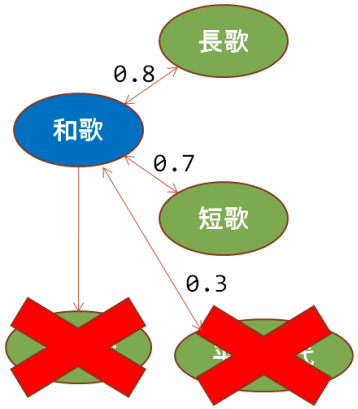

図 3 リンク構造の解析
Fig. 3 Analysis of the link structure

3.2 関 連 度

これまで我々の研究[7] では上記の (3) の手順をコサイン類似度を用いて記事同士の類似度を測り比較対象の日本語記事群を取得していた．しかし，それだけでは比較対象日本語記事

Vol.2011-DBS-152 No.3
Vol.2011-IFAT-103 No.3
2011/8/2

を抽出しきれないという問題が生じた．例えば，上記の 4.1 でも挙げた「和歌」という記事とその記事と複数ページにまたがっている「長歌」の記事との類似度が低く抽出できずにいた．そこで本論文では Wikipedia の目次構造に基づいた関連度の計算を行うことで，基準ノードに関連する複数ページ群の抽出する．我々の提案する関連度とは，Wikipedia のある記事と双方向リンクされている別の記事がどのぐらい関わりが深いかを測るための尺度である．そこで，基準ノードである記事とそれに双方向リンクされている記事がどれぐらい関連しているかを Wikipedia の目次構造と記事間のリンクのアンカー文字の位置に注目する．以下，関連度を求める手順を示す．

(1) 基準リンクグラフの基準ノードの記事の目次構造に従って記事内の情報を分割する (**図 4**)．ここで分割された最小単位をセグメントと呼ぶ．

(2) セグメントをノードとし，目次構造に従って**図 5** のような木構造に変換する．この木構造をセグメントツリーと呼ぶ．

(3) 基準リンクグラフ内の各ノードが基準ノードのセグメントツリーのどのノードからリンクされているかを抽出する．例えば，図 3 の「短歌」のノードが図 5 のどのセグメントからリンクが張られているかを抽出する．

(4) 以下の式 (1) を用いて関連度を計算をする．

$$RW = af * cos(a,b) + \sum_{i=1}^{af} (\frac{1}{d_i})^{n_i} * (n_i - o_i + 1) \quad (1)$$

ここでの af は基準ノードの記事における基準ノードの記事と双方向リンクされているある記事のリンクの貼られている個数を指す．d_i は基準ノードの記事と双方向リンク貼られているノードの深さを指す．n_i は基準ノードの記事と双方向リンク貼られいてるノードの深さにおけるノードの数を示している．o_i は基準ノードの記事と双方向リンク貼られいてるノードの深さにおける左からの順番を示している．cos(a,b) はコサイン相関値であり， a は基準ノードの記事の名詞の出現頻度であり，b は基準リンクグラフ内のその他のノードの記事の名詞の出現頻度を示している．

我々は基準ノードとリンクが張られている回数の多い記事，図 5 より，リンクの貼られいてる位置が木構造の深さの浅い所にあり，尚且つ木構造の左側すなわち目次番号が早いほど基準ノードにとって関連性が高いと考え上記の (1) 式を提案した．関連度がある閾値以上の記事を，比較対象の記事とする．

図 4 Wikipedia の目次構造の分割
Fig. 4 segment contents structure of Wikipedia

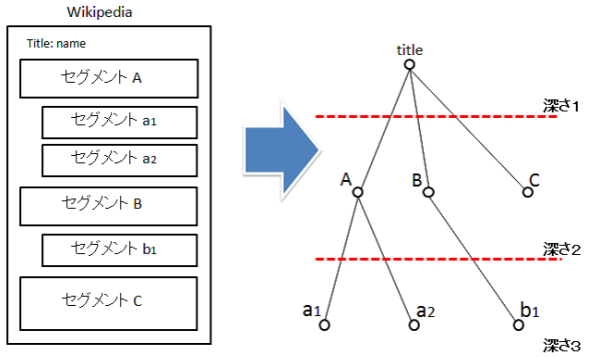

図 5 Wikipedia の記事を木構造化
Fig. 5 Tree posture Creator of the article of Wikipedia

Vol.2011-DBS-152 No.3
Vol.2011-IFAT-103 No.3
2011/8/2

4. 多言語記事の差分抽出

4.1 多言語 Wikipedia の比較

言語にかかわらず Wikipedia の記事は目次構造に基づいて段落に分かれている．つまりは，Wikipedia の段落は意味的に分かれている可能性が高いと考えられる．そこで，我々は多言語 Wikipedia を比較する際に Wikipedia の目次構造に注目し，目次構造に基づくコンテンツの比較を行う (**図 6** 参照)．類似している段落の中から，その差分情報を抽出することを行う．

ここでは，日英 Wikipedia 各々の記事の段落毎にテキストの形態素解析を行い名詞のみを抽出する．そして比較を行うため辞書を用いて英語の名詞を日本語に翻訳する．本研究では翻訳に GENE95 辞書[⋆1]を使用する．また，GENE95 辞書に載っていない単語は Google Ajax api[⋆2] と Microsoft Translator api[⋆3] の翻訳を使用する．しかし，下駄の種類である「おこば」や水墨画家である「雪舟」などの和製単語や人名は翻訳することができない．そこで，Wikipedia を用いて翻訳を行う．Wikipedia には，多言語へのリンクが存在し，それを利用することで該当する言語に翻訳することができる (**図 7**)．なお，翻訳時に単語の多義性が問題になるが，今回はこの多義性には考慮せず，今後の課題とする．次に日本語版 Wikipedia の記事と英語版 Wikipedia の日本語翻訳の記事の名詞の出現頻度を求める．そして，以下の式 (2) のコサイン相関値を用いて，各々の段落における類似度を求め，ある閾値以下段落を差分情報として抽出する．

$$cos(x, y) = \frac{\sum x_i * y_i}{\sqrt{\sum x_i^2 * \sum y_i^2}} \qquad (2)$$

なお，上記の式の x_i は日本語版のある一つの目次のコンテンツの名詞の出現頻度．y_i は英語版のある一つの目次のコンテンツの翻訳した名詞の出現頻度を表す．

5. プロトタイプシステム

以上の提案手法を用いて，開発言語に Ruby[⋆4]，日本語形態素解析器に Mecab[⋆5]，英語の

⋆1 GENE95 辞書 http://www.namazu.org/ tsuchiya/sdic/data/gene.html
⋆2 Google Ajax api http://code.google.com/apis/language/
⋆3 Microsoft Translator api http://www.microsofttranslator.com/dev/
⋆4 Ruby http://www.ruby-lang.org/ja/
⋆5 Mecab http://mecab.sourceforge.net/

図 6 目次構造とコンテンツ
Fig. 6 Table of contents structure and contents

図 7 Wikipedia の言語リンク
Fig. 7 Language link of Wikipedia

Vol.2011-DBS-152 No.3
Vol.2011-IFAT-103 No.3
2011/8/2

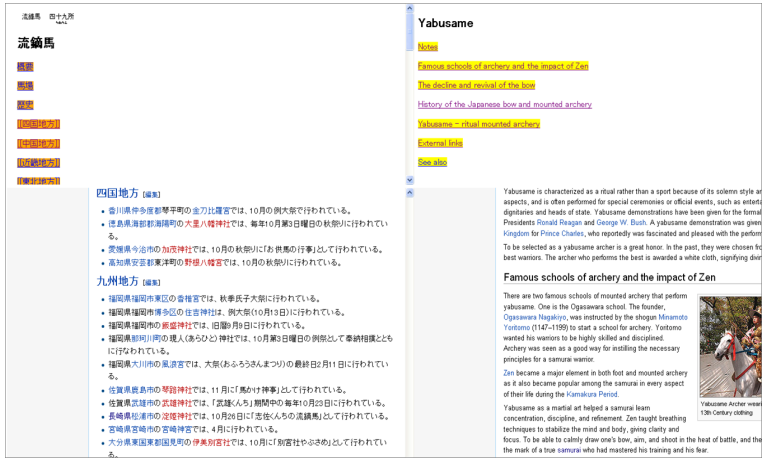

図 8 プロトタイプシステムの出力画面
Fig. 8 output of prototypesystem

Tagger に Tree Tagger[⋆1]，データベースに Mysql[⋆2]を使用しプロトタイプシステムを作成した．プロトタイプシステムではユーザはキーワードを入力すると，システムは出力画面**図 8** に示すように，日本語と英語の差分情報を提示する．図 8 に示す出力画面では左側に日本語記事の目次コンテンツを，右側に英語記事の目次コンテンツを示してある．日本語のページが複数にまたがる場合は左上に表示されてあるタブを押すことでその記事の目次に切り替えが可能である．なお，得られた差分情報は日本語版にしかない目次コンテンツをオレンジ色のマーカーで，英語版にしかない目次コンテンツを黄色のマーカーで示してある．目次をクリックすると Wikipedia 上のその記事の目次が下のフレームに表示され，コンテンツの内容を確認できるようになっている (図 8).

6. 実験と考察

6.1 評 価 実 験

作成したプロトタイプシステムを用いて提案の差分抽出手法の有用性を示すために評価実験を行った．実験データは入力 Query として日本の文化である花札，下駄，俳句，漫才，流鏑馬の 6 コの単語を用いた．尚，本実験における比較対象のページは各 Query1 ページとした実験内容はプロトタイプシステムで得られた差分情報の適合率，再現率，F 値を求める．ここでの再現率は人手により抽出した日本語版だけにしかない目次コンテンツと英語だけにしかない目次コンテンツを正解データとした．再現率，適合率，F 値の式を以下に示す．なお，今回の評価実験は筆者 1 人で行った．

表 1 実験の結果
Table 1 Result of experiment

クエリ	適合率	再現率	F 値
花札	60 %	46 %	52 %
下駄	100 %	42 %	59 %
俳句	83 %	62 %	71 %
漫才	100 %	50 %	67 %
流鏑馬	72 %	93 %	81 %
平均	83 %	59 %	66 %

$$\text{再現率} = \frac{\text{正解データ} \wedge \text{抽出した差分情報}}{\text{正解データ}} \tag{3}$$

$$\text{適合率} = \frac{\text{正解データ} \wedge \text{抽出した差分情報}}{\text{抽出した差分情報}} \tag{4}$$

$$\text{F 値} = \frac{2 * \text{適合率*再現率}}{\text{適合率+再現率}} \tag{5}$$

6.2 実験の結果

結果を以下の**表??**に示す．

差分情報の例として俳句という記事であれば日本語版だけの目次コンテンツとして「句またがり」,「本歌取り」といった俳句における技法や「客観写生」といった俳句における文学理論が抽出された．逆に英語版だけの目次コンテンツとして「インターネット」といったオンラインで俳句を公開しているサイトや雑誌などが紹介されているなどの差分情報が抽出された．

6.3 考　　察

表 1 からわかるように結果として F 値の平均が 66%となり良い結果を得ることができた．結果が悪い例として類似度が閾値以下であるが英語版に存在する記事が差分として抽出された．その原因として人名や和歌や発句などの和声単語の翻訳ができなかった，そして花札

⋆1 Tree Tagger http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
⋆2 Mysql http://www-jp.mysql.com/

Vol.2011-DBS-152 No.3
Vol.2011-IFAT-103 No.3
2011/8/2

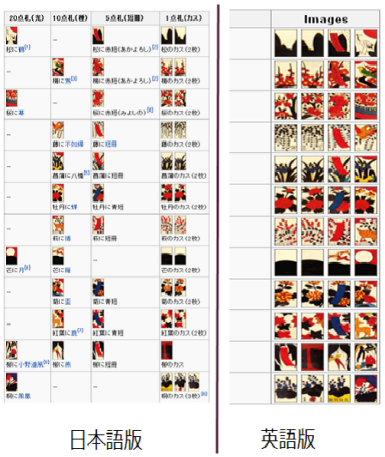

図 9 結果の悪い例
Fig. 9 Bad example of the result

などの画像情報は今回提案した手法では類似度計算ができない故に起こったというものがあげられる (**図 9**).

7. まとめと今後の課題

本研究では, ユーザの入力したキーワードの日語版と英語版の Wikipedia 上での差分情報を取得する手法の提案および作成したプロトタイプシステムの性能の評価実験の結果についての報告を行った．具体的には，比較する日本語の Wikipedia の記事を決定する際に記事のリンク構造を解析した．そして，差分抽出では Wikipedia の目次構造に着目し，目次構造を構成する記事の最小単位であるセグメントごとに比較をし差分情報を抽出した．実験の結果，適合率の平均が 83%，再現率の平均 59%，そして F 値の平均が 66%となる良い結果を得ることができた．今後の課題は以下の通りである．

- 重要度，信頼度の計算
  本論文では日本の伝統文化を対象とした日語版と英語版の Wikipedia 上での差分情報を取得する手法の提案を行った．今後は得られた差分情報の重要度，信頼度を計り他方の言語版に加えるために，重要度，信頼度の算出する手法を提案し，Wikipedia の情報の網羅性，信頼性を向上することに取り組んでいきたい．
- 単語の多義性
  本研究では，多言語間の差分情報抽出手法を提案した．しかし，今回は単語の多義性は考慮していない．故に今後は単語の多義性を解消させることで差分情報抽出手法の精度を向上させ，再現率を向上させることに取り組んでいきたい．
- 差分情報の提示方法
  本研究では，Wikipedia の差分情報の提示を行った．今後は，差分情報を他方の言語版に加え提示する予定である．故に今後はその提示する方法を考えていく予定である．
- 差分抽出方法の性能向上
  本研究では，多言語間の差分情報抽出手法を提案した．しかし，今回は目次のみが差分情報として抽出されている．故に今後は目次の中身のコンテンツを文や文章単位で差分情報を抽出する手法の提案に取り組んでいきたい．
- 評価実験
  本研究では，評価実験として提案手法を用いて得られた差分情報の適合率，再現率，F 値を求めた．しかし，この評価実験は筆者 1 人で行ったものである．今後は筆者以外の被験者に評価実験を行い，得られた差分情報の精度を調べる予定である．そして比較対象ページについての実験も行っていないので今後の課題とする．

参 考 文 献

1) asahi.com: 悩むウィキペディア　少ない管理人　芸能系ばかり人気，朝日新聞（オンライン），
入手先⟨http://www.asahi.com/national/update/0303/TKY201003030157.html⟩
(参照 2010-03-04).
2) 森　竜也，増田英孝，中川裕志，清田陽司：Wikipedia における言語間の差異マイニング，情報処理学会創立 50 周年記念 (第 72 回) 全国大会， No.5, pp.181–182 (2010).
3) 斎藤雄介，山田剛一，絹川博之，中川裕志：日中英ニュース記事比較のための収集と検索，情報処理学会 第 71 回 全国大会， No.2, pp.269–270 (2009).
4) 中崎寛之，川場真理子，山崎小有里，宇津呂武仁，福原知宏：同一トピックの日英ブログにおける文化間差異の発見支援，*DEIM Forum 2009*.
5) 立床雅司，高橋　慧，斎藤昭則，吉川正俊：Wikipedia のリンク構造に基づく関連度を利用したコンテンツ検索手法と地球科学データへの応用，*DEIM Forum 2009*, pp. 1–8.

Vol.2011-DBS-152 No.3
Vol.2011-IFAT-103 No.3
2011/8/2

6) K. Nakayama, T. Hara, S. Nishio：Wikipedia Mining for An Association Web Thesaurus Construction, *WISE 2007*, pp.1–11.
7) 藤原裕也，灘本明代：言語 Wikipedia を用いた伝統文化の差異情報抽出の提案，情報処理学会 第 73 回 全国大会， No.1, pp.1.575–1.576.