情報の質を考慮した多言語Wikipedia記事の差異抽出手法の提案

藤原　裕也† 　鈴木　優†† 　小西　幸男††† 　灘本　明代††††

† 甲南大学大学院　自然科学研究科〒 658-8501 兵庫県神戸市東灘区岡本 8-9-1
†† 名古屋大学　情報基盤センター〒 464-8601 愛知県名古屋市千種区不老町
††† 甲南大学　国際交流センター〒 658-8501 兵庫県神戸市東灘区岡本 8-9-1
†††† 甲南大学　知能情報学部〒 658-8501 兵庫県神戸市東灘区岡本 8-9-1
E-mail: †life.of.151a@gmail.co.jp, ††suzuki@db.itc.nagoya-u.ac.jp, †††ykonishi@center.konan-u.ac.jp,
††††nadamoto@konan-u.ac.jp

あらまし　Wikipedia は，日本語版や英語版などの各言語版によって記述されている情報の量が異なる．そのため，一つの言語版だけでは情報量が不足する記事が多く存在する．そこで本研究では母国語と母国語以外の Wikipedia の記事を比較し，母国語版に不足している情報を母国語以外の言語版から補足する手法を提案する．具体的には，ユーザの入力した事柄に関する母国語の記事と母国語以外の記事を比較し，その差分情報を提示する．この時，言語間の情報の粒度が異なるため，一つの母国語の記事が一つの母国語以外の記事に対応するとは限らない．そこで，比較対象となる複数の母国語以外の記事を，リンクグラフを用いて抽出し，抽出された複数の母国語以外の記事と母国語の記事を比較対象とし，その差分情報を抽出する．このとき補足する母国語以外の記事の情報の質を計ることにより，より質の高い情報をユーザに提示する．

キーワード　Wikipedia，多言語，差異抽出，情報の質

Extracting Difference Information from Multilingual Wikipedia in Consideration of the Quality of Information

Yuya FUJIWARA†, Yu SUZUKI††, Yukio KONISHI†††, and Akiyo NADAMOTO††††

† Graduate School of natural science graduate course, Konan University Graduate School8-9-1 Okamoto Higashinada , Kobe, Hyogo, 6588501 Japan
†† Information Technology Center, Nagoya University Furotyo, Chikusa, Nagoya, Aichi, 4640814 Japan
††† Konan International Exchange Center,Konan University8-9-1 Okamoto Higashinada , Kobe, Hyogo, 6588501 Japan
†††† Dept.of Intelligene and Informatics, Konan University8-9-1 Okamoto Higashinada , Kobe, Hyogo, 6588501 Japan
E-mail: †life.of.151a@gmail.co.jp, ††suzuki@db.itc.nagoya-u.ac.jp, †††ykonishi@center.konan-u.ac.jp,
††††nadamoto@konan-u.ac.jp

Key words　Wikipedia，Multilingual，Difference Information，Quality of Information

1. はじめに

Wikipedia(注1)はオンライン上の百科事典の一つであり，誰もがコンテンツを作成できる点や，一つの話題に対し複数の言語版で書かれている点などを特徴として挙げることができる．Wikipedia の各言語版はそれぞれ独立して管理されている場合が多く，各々の言語版を閲覧しているユーザが記事の作成，追記，修正を行っている．例えば，イギリスの伝統的なスポーツである「ローンボウルズ」の英語版の Wikipedia の記事には目次項目も多く，コンテンツが詳細に記述されているが，日本語版の記事は目次項目が少なく，コンテンツ量も少ない．なぜなら，日本語版の記事を作成しているユーザが主に日本人であり,「ローンボウルズ」のことについて十分な知識を持っていない点や，興味がイギリス人と比較して比較的少ない点が原因

(注1)：http://ja.wikipedia.org/wiki/

であると考えられる．一方，日本の伝統的な建物である「平等院」について日本語版と英語版の記事で比較すると，日本語版の記事には目次項目が多く，コンテンツが詳細に記述されているが，英語版の記事は目次項目とコンテンツが少ないといったように,「ローンボウルズ」と逆の現象が起こっている．このように，Wikipedia の記事の言語間で情報の量が異なり，自国の言語版だけでは得られる情報が不足する場合があることがわかる．一般的にユーザは，母国語版の Wikipedia を閲覧する場合が多い．たとえ母国語以外の言語を読むことが可能であっても，母国語の記事を読んで理解することと比較し時間が掛かり困難な作業である．例えば，ある程度英語を読むことができる日本人ユーザが英語版の「Bowls」を全て読んで理解することは，日本語の記事を読んで理解することと比べて困難である．同様に，日本語をある程度読むことができる英語圏のユーザが日本語版の「平等院」を全て読むことは困難である．そこで我々はユーザが閲覧している母国語版の記事に対して，母国語版では不足している情報を他の言語版から取得し挿入することによって，よりユーザの理解度が上がると考えた．本論文では，多言語版 Wikipedia 間の差分情報を抽出し，提示するシステムを提案する．このとき，Wikipedia は誰もが編集を行うことができるため，記述されている情報が常に正しいとは限らない．つまり，誤った記述が投稿された場合でも Wikipedia はその記述をそのまま記事に反映させるため，Wikipedia の記事には誤った情報が含まれる場合があり，記事の全部が必ずしも信頼できる情報だけで構成されているわけではない．最近ではある口コミサイトで嘘のクチコミよる店の評価やランキングの操作(注2)など誤った情報を提示することは社会問題に発展することがある．このように質のある情報を提示することは有益であると考えられる．そこで本研究では，多言語 Wikipedia を比較し取得した差分情報の質を求め，ある程度質の高い情報のみを提示することを行う．これによりユーザは，より有益な情報を容易に取得することが可能となる．

我々の想定しているユーザは，母国語ではない言語をある程度理解することができる人である．本論文ではこのユーザがある程度理解できる言語を比較言語と呼ぶ．例えば，ある程度英語がわかる母国語を日本語とする人の場合，母国語は日本語，比較言語は英語となる．本研究では，ユーザの入力したキーワードに関するユーザの母国語版の Wikipedia と比較言語版の Wikipedia の記事をそれぞれ抽出し，記事の内容を比較した上でその差分情報をユーザに提示する．この時，一つの母国語の記事が一つの比較言語の記事だけに対応するわけではなく，複数の比較言語の記事に対応する場合がある．例えば，日英を比較するときの「ローンボウルズ」の場合．日本語版はローンボウルズの世界大会の説明がローンボウルズという記事１つに書いてあるのに対し，英語版ではローンボウルズの記事だけでなくその世界大会の記事が存在し，複数ページまたがっている．このように言語が異なると Wikipedia の記事の構成も異なる．

(注2)：http://sankei.jp.msn.com/life/news/120117/trd12011718250013-n1.htm

そこで我々は，比較対象となる複数の比較言語の記事を，記事間のリンクグラフと我々の提案する関連度を用いて抽出する．さらに，抽出された複数の比較言語の記事と母国語の記事を各々の記事の目次構造に基づいて比較し，その差分情報を提示する．そして得られた差分情報の質を測るため，記事の残存率に基づく質の算出手法を用いて算出する．なお，本論文ではユーザの母国語で書かれた Wikipedia の記事を母国語記事と呼び，比較言語で書かれた Wikipedia の記事を比較言語記事と呼ぶ．以下に処理の流れを示す．

（1） ユーザは調べたい比較言語の文化に関する事柄に関する用語をクエリとして入力する．

（2） 入力されたクエリをタイトルとする母国語記事と，比較言語記事をそれぞれ取得する．

（3） (2) で取得した比較言語記事のリンク構造を解析し，比較対象となる比較言語記事群を取得する．

（4） (3) で取得した比較言語事群と，(2) で取得した母国語記事を各々比較し差分情報を取得する．

（5） (4) で取得した差分情報に対して，記事の残留度を利用した質の計算を行う．

（6） (5) で得られた，質の高い差分情報をユーザに提示する．

2. 関 連 研 究

Wikipedia における記事間の関連を抽出する研究は，現在までに数多く存在する．Michael [1] や Milne ら [2] は Wikipedia の記事が属するカテゴリ情報を利用して，概念間の関連度を算出している．二つの概念が与えられたときに，カテゴリの中でその二つの概念がどれほど近いかを基準に関連度を算出している．これに対し，我々の提案している関連度は記事間のリンクを利用している点が異なる．中山ら [3] は Wikipedia のリンク構造を解析しシソーラス辞書を構築する手法を提案している．その際に，pfibf と呼ばれる手法を用いて Wikipedia の記事同士の関連度を計測している．pfibf とは記事から記事へのパスの多さ，各パスの長さに着目し記事同士の関連度を計る手法である．それに対し，本研究ではリンク数，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．高橋ら [4] は Wikipedia のリンク構造を用いて，歴史エンティティの重要度を算出している．具体的には，ある歴史エンティティから他の歴史エンティティへのリンクが張られている場合，前者は後者から影響を受けたと見なしている．また，Wikipedia の記事には時間や場所に関する情報が含まれることが多い．これらの情報に Brin ら [5] の PageRank に類似した反復計算アルゴリズムを用いて計算を行っている． Jaap ら [6] は Web のインリンクを重要と考え，Wikipedia と Web のリンク構造の違いについて研究を行っている．インリンクとは他サイトから自サイトへ貼られたリンクのことを表す．本研究では Wikipedia の記事同士の内容の違いを抽出しているため，ある話題に対して載っていない情報や，ユーザが知らない情報を知ることができる点が異なる．Wikipedia のリンク構造を用いてシソーラスを構築している研究も行われている．David [7] は Wikipedia のリン

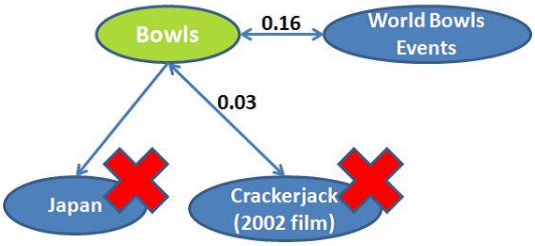

図 1　リンク構造の解析
Fig. 1　Analysis of the link structure

ク構造を用いて，語の意味的関係を抽出している．Chen ら [8] は Web ページ同士のリンク構造を解析し、Web シソーラス辞書を自動的に構築する手法を提案している．具体的には Web サイトの階層構造からサブツリーと呼ばれる多数の木を形成し，木の中での語の共起性を利用することで語の関連度を計測している．本研究との違いはリンク構造を用いて Wikipedia の記事同士の関連する度合い抽出する点で異なる．多言語 Wikipedia を用いて言語版同士を比較している研究も存在する．Eytan ら [9] は Wikipedia の infobox の違いを抽出し他の言語版へ補完を行っている．infobox とは Wikipedia の記事に含まれる，データが記述されている表のことである．しかし我々は目次構造に着目し Wikipedia の記事を比較しているため，片方の言語版記事にない目次や記事の内容を知ることができる点が異なる．Wikipedia 記事の質に関する研究も多く存在する．Wöhner ら [10] は記事の周期的に起こる変化に着目することによって，記事の質を測定している．著者の編集量の変化と質には関連があることに着目している．Hu [11] らは記事と編集者の相互依存関係に基づき，記事の質の高さによって順位付けを行っている．その際，編集履歴と記事の長さに着目し質の算出を行っている．これに対して我々は，記事の残留度に注目して質を算出することによって，小さな計算量で十分な精度の質を算出することができる点が異なる．

3.　比較対象記事群の抽出

3.1　リンクグラフの作成

多言語 Wikipedia は各言語版により言語や文化の違いから情報の粒度がことなり，対応する記事が複数にまたがる場合がある．例えばイギリスに関する記事の場合，日本語版 1 記事に対し，英語版が複数になっていることが多い．「ローンボウルズ」という記事は日本語版だとローンボウルズの競技用の芝生であるボウリンググリーンが説明されているが，それに対し英語版は同じ用に記事の中にボウリンググリーンの説明があるのはもちろんのこと，さらにそれを詳しく説明した「Bowling green」という記事が存在しこのように複数にまたがっている．そこで我々はこのような複数にまたがっている記事を抽出するため，リンク構造の解析を行い抽出する．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比較言語記事から比較言語記事をノード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．そしてユーザの入力したクエリと同じタイトルを持つ記事のノードを基準ノードと呼ぶ．

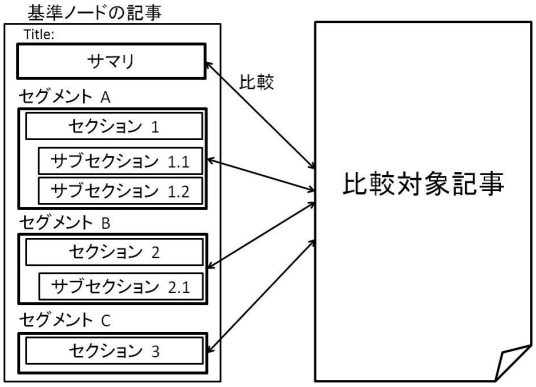

図 2　関　連　度
Fig. 2　relevance degreee

（2）　基準ノードと双方向にリンクしているノードは強連結であり，基準ノードの記事に深く関連すると考え，双方向リンクされているノード以外のノードを削除する．

（3）　(2) の基準リンクグラフから，母国語記事を持つノードを削除する．(図 1)

（4）　(3) の基準リンクグラフにおいて，基準ノードとその他のノード間の関連度を求める．

（5）　その関連度が事前に設定した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比較言語記事群とする．

3.2　関　連　度

我々の言う関連度とは Wikipedia のある記事と双方向リンクされている記事との関わりがどれだけ深いかを調べるための尺度である．今回我々は新たに基準ノードの記事の部分的な情報と比較対象ノードの類似性、アンカー文字列の出現位置，アンカー文字列の出現回数に注目し関連度を求めている．以下に関連度を抽出する手順を示す．

（1）　基準ノードとなる Wikipedia の記事を分割する．この時，Wikipedia の一番最初の説明部分をサマリとしそれ以外のコンテンツをセグメントと呼ぶ．

（2）　基準リンクグラフ内の各ノードが，基準ノードのサマリ，セグメントのどの部分からリンクを貼っているかを全て抽出する．

（3）　そのサマリまたはセグメントと比較対象ノードとを対象にし，以下の式 (1) (2) を用いて関連度を求める (図 2)．

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

$$W_{kl} = R_{kl} / \max(R_{\mathrm{km}}) \qquad (2)$$

ここでの tf_{sum} は基準ノード k のサマリに出現するアンカー文字列の出現回数を指す．S_{sum} は基準ノード k のサマリと比較対象ノード l の名詞の出現頻度を重みとしたコサイン類似度である．tf_i はそれ以外のあるノードすなわちある部分グラフに出現するアンカー文字列の出現回数を，そして S_i は基準ノード k におけるその部分グラフと比較対象ノード l の名詞の出現頻度を重みとしたコサイン類似度である．尚，我々は基

準ノードのサマリにリンクを貼っている比較対象の記事は関連性が高いと考え α を掛けている．なぜならサマリの情報はその記事全体の大まかな概要を示してあり，そこに出現するアンカー文字列はこの記事を知る上で必要不可欠な情報であると判断したためである．例えば,「Bowls」の記事と双方向リンクされている「Bowling green」の場合は「Bowling green」のアンカー文字列が「Bowls」の ”Game ”に存在することから tf_i=2, S_i=0.41 となりこれを上記の (1) 式に当てはめ 0.8 となる．さらに (2) 比較対象ノード全ての関連度の最大値で割ることにより 0 から 1 の範囲に正規化を行なっている．このようにして「Bowling green」の関連度は 0.6 となる．この関連度が事前に設定した閾値 β 以上であれば比較対象の記事とし抽出する．

4. コンテンツの比較

言語にかかわらず，Wikipedia の記事は目次構造に基づきセグメントに分かれていることが多い．このとき，Wikipedia のセグメント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 を比較する際にセグメントに注目し，セグメントに基づくコンテンツの比較を行い (図 3)，類似していないセグメントを抽出し差分情報とする．まず初めに母国語記事と比較言語記事をそれぞれ形態素解析を行い，名詞のみ抽出する．次に辞書を用いて比較言語記事の名詞を母国語に翻訳する．また，辞書に掲載されていない単語は Google Ajax API や Microsoft Translate API を用いて翻訳を行なっている．しかし，固有名詞や人名などは翻訳することができない．そこで Wikipedia の言語間リンクも用いて翻訳を行う． ここでは，ある単語の翻訳語を探したいとき，その単語に関する記事を検索し，その記事に含まれる言語間リンクを用いて翻訳語の言語版記事を抽出し，その記事のタイトルを翻訳語とする．なお，翻訳時に単語の多義性など意味の曖昧性が問題となるが，本論文ではこの問題は考慮せず，今後の課題とする．次に母国語記事と比較言語記事のセグメントごとに名詞の出現回数を求める．以下の式 (3) のコサイン類似度を用いて，母国語記事と比較言語記事のセグメントごとの類似度を求める．ある比較言語記事のセグメントが母国語記事の全てのセグメントに対し類似度がある閾値以下であった場合，その比較言語記事のセグメントを差分情報として抽出する．なおここでの閾値は実験より 0.3 とする．

$$\cos(x, y) = \frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}} \qquad (3)$$

ここでの x は母国語記事の 1 つのセグメントであり，y は比較言語記事の 1 つのセグメントである．x_i は母国語記事の 1 つのセグメント内に存在する名詞 i の出現回数のベクトルであり，y_i は比較言語記事の 1 つのセグメント内に存在する母国語に翻訳した名詞 i の出現回数ベクトルを表す．

5. 情報の質の計算

コンテンツの比較によって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しているため，得られた差分情報の中には質の低い情報が存在することがある．そこで我々は鈴木ら [12] によって提案された，記事の残存率に基づく質の算出手法を用いて記事の情報の質を測定する．

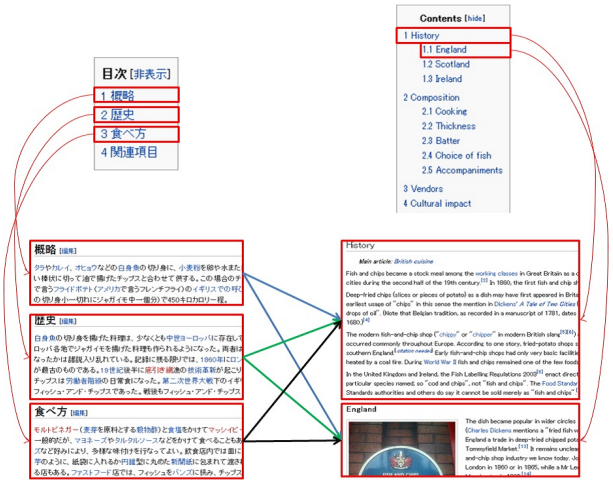

図 3　コンテンツの比較
Fig. 3　Compare contents

具体的な質の算出手法について述べる．Wikipedia には記事 $i = 1, 2, \cdots, M$ が存在し，それぞれの記事には $v_{i,j} \in V$ のバージョンが存在する．V は全ての記事の全てのバージョンの集合である．ここで $j = 0, 1, \cdots, N_i$ は記事のバージョン番号を表す．$j = 1$ のとき，つまり $v_{i,1}$ は記事 i が最初に作成されたバージョンを表す．$j = 0$ のとき，つまり $v_{i,0}$ は記事 i が作成されているが内容が無い状態を表す．Wikipedia の記事を記述した著者 $e = 1, 2, \cdots, K$ は，一つ以上の記事のバージョンを作成している．著者 e が作成した記事は $A_e = \{v_{i,j} | v_{i,j} \in V \text{ and } v_{i,j} \text{ is written by } e\}$ であり，一つのバージョンを作成した著者は一人である．ただし，$j = 0$ のバージョンの著者は存在しないと仮定する．

まず，j 回目の編集においてどの部分を追加・削除したかを特定するために，$v_{i,j-1}$ と $v_{i,j}$ との増加部分 $add_{i,j}$ および削除部分 $del_{i,j}$ を求める．

次に，p $(p = 0, 1, \cdots, N_i - j)$ 回後に編集されたバージョン $v_{i,j+p}$ において，$add_{i,j}$ と $del_{i,j}$ が残存している割合を算出する．ここで，$p = 0$ のときは $\delta(add_{i,j}, 0) = add_{i,j}$ ，$\delta(del_{i,j}, 0) = del_{i,j}$ とする．まず，$v_{i,j+p}$ の中から $add_{i,j}$ ，$del_{i,j}$ に相当する部分 $\delta(add_{i,j}, p)$, $\delta(del_{i,j}, p)$ を抽出する．次に，追加部分，削除部分の残存率である追加残存率，削除残存率 $R^{add}(i, j, p)$, $R^{del}(i, j, p)$ を (4) ，(5) 式によって求める．

$$R^{add}(i, j, p) = \frac{|\delta(add_{i,j}, p)|}{|add_{i,j}|} \qquad (4)$$

$$R^{del}(i, j, p) = \frac{|\delta(del_{i,j}, p)|}{|del_{i,j}|} \qquad (5)$$

ここで，$|\delta(add_{i,j}, p)|$, $|\delta(del_{i,j}, p)|$, $|add_{i,j}|$, $|del_{i,j}|$ はそれぞれ，$\delta(add_{i,j}, p)$, $\delta(del_{i,j}, p)$, $add_{i,j}$, $del_{i,j}$ に含まれる文字数である．

次に，標準残存率を利用して残存率を正規化する．標準残存

率とは p 回後に追加，削除された時の残存率の平均値である．標準残存率を利用して正規化を行う理由として，事前に行った予備実験において，編集回数が増加するごとに残存率が低下することが分かったことが挙げられる．正規化された残存率 $\overline{R^{add}(i,j,p)}$，$\overline{R^{del}(i,j,p)}$ を (6), (7) 式によって求める．

$$\overline{R^{add}(i,j,p)} = \frac{R^{add}(i,j,p)}{S^{add}(p)} \tag{6}$$

$$\overline{R^{del}(i,j,p)} = \frac{R^{del}(i,j,p)}{S^{del}(p)} \tag{7}$$

ここで $S^{add}(p)$, $S^{del}(p)$ はそれぞれ p 回後の編集における残存率の平均値である．

そして，追加残存率と削除残存率を組み合わせて，記事変更質度 $\tau(v_{i,j})$ を求める．記事変更質は，追加残存率と削除残存率の総和であり，(8) 式で求める．

$$\tau(v_{i,j}) = \sum_{q=1}^{N_i-j} R^{add}(i,j,p) + \sum_{q=1}^{N_i-j} R^{del}(i,j,p) \tag{8}$$

記事変更質を算出するとき，編集回数による正規化を行わない．なぜならば，編集回数が多いとき編集の記事変更質は高くなるべきであると考えたためである．

最後に，記事変更質の平均を 0 とする．なぜならば，Wikipedia における記事変更の大半は小さな変更であり，記事自体の質は変化しないと考えられる．それらの記事変更質の変化よりも低い記事変更質があったとき，その変更は記事の質を低下させていると考えられるためである．最終的な記事変更質 $\overline{\tau(v_{i,j})}$ は (9) 式で求める．

$$\overline{\tau(v_{i,j})} = \tau(v_{i,j}) - \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i}\tau(v_{i,j})}{\sum_{i=1}^{M} N_i} \tag{9}$$

6. 実験:重み α と閾値 β の設定

ここでは 3.2 で述べた関連度の式 1 の重み α と閾値 β の設定をするために行った実験について述べる．我々は関連度の式に対して α を 1 から 10 の 1 刻みに設定し，閾値を 0 から 1 の範囲を 0.05 刻みにし各々の F 値を比較した．尚，この実験に我々は母国語記事として日本語版 Wikipedia，比較言語記事として英語版 Wikipedia を採用し実験データとして 13 個のクエリを使用した．そのクエリと正解データの数を表 1 に示す．

表 1　実験で用いたクエリとその正解データの数

#	クエリ	# 正解データの数
1	Bannock (food)	2
2	Warwick Castle	2
3	Black dog (ghost)	7
4	Fish and chips	4
5	Goodwood Festival of Speed	2
6	Bowls	2
7	Blue plaque	1
8	Burlesque	3
9	Flag of Scotland	6
10	Gaelic handball	4
11	Kipper	3
12	National Gallery of Scotland	12
13	Lipton	1
	Total	49

結果と考察

結果を図 4 に示す．尚，α=7 から 10 までは F 値が下がっていったため α=8，9 はこの図には記載しない．この図からわかるように α=3，β=0.2 のときに最も高い F 値を得ることができた．これにより重み α を 3，閾値 β を 0.2 とする．その時のコサイン類似と本提案手法の関連度を用いた時の再現率，適合率，F 値を比較した結果を表 2 に示す．抽出された記事の例として Warwick Castle であれば Warwick Castle の歴代所有者の一覧，Gaelic handball であれば GAA Handball といった Gaelic handball の理事会などが抽出することができた．表 2 からわかるようにコサイン類似度よりも本論文の提案手法の方が適合率が 35%から 57%，再現率が 49%から 59%，F 値が 41%から 58%と向上し，提案手法の有用性を示すことができた．これはコサイン類似度では比較対象のある記事と記事との類似性に注目しているため，適合率が低下したと考えられる．例えば，Gaelic handball において GAA Handball と言った記事は Gaelic handball の記事のように競技の説明をしているわけではないので類似性が低くなり抽出ができず結果が悪くなった．それに対し本提案手法では記事と記事の類似度ではなく，その複数ページの対象となる記事を説明しているコンテンツと複数ページ対象の記事とを比較しているために S_{sum} や S_i などの類似度が高くなり，さらにその複数ページの対象となる記事を説明している目次コンテンツは必然的に tf_{sum} や tf_i などのアンカー文字の出現回数が多くなるために関連度が高くなり，結果，適合率，再現率，F 値が高くなったことがわかった．Gaelic handball は正解データが主に Wikipedia の記事のサマリにアンカー文字列に貼られていたため結果が良くなった．我々が今回提案した関連度の式はサマリ部分にアンカー文字列が貼っているものは関連性が高いと考えているため値が大きくなったと考えられる．また同じく結果の良かった Burlesque は正解データが存在する部分グラフのコンテンツ量が多く対象となるアンカー文字列の出現回数が多くなり tf_{sum} や tf_i の値が大きくなったためと考えられる．結果が悪い例として，Bowls や Blue plaque などが上げられる．Bowls は Bowls の世界イベントなどが書かれてある World Bowls Events などが抽出できなかった．原因として，World Bowls Events は Bowls の ”World Bowls Events ”という目次に存在するが，中身のコンテンツの量がほぼ書かれていないため関連度の計算の際値が引くなったためと考えられる．また，Blue plaque は正解データが List of blue plaque の一つしかなく，さらに (1) 式を用いた際の値は低くなかったが，(2) 式を用いて正規化した際に値が低くなったため抽出できなかったためであると考えられる．これらは今後の課題である．

7. まとめと今後の課題

本論文では，ユーザの入力したキーワードに関するユーザの

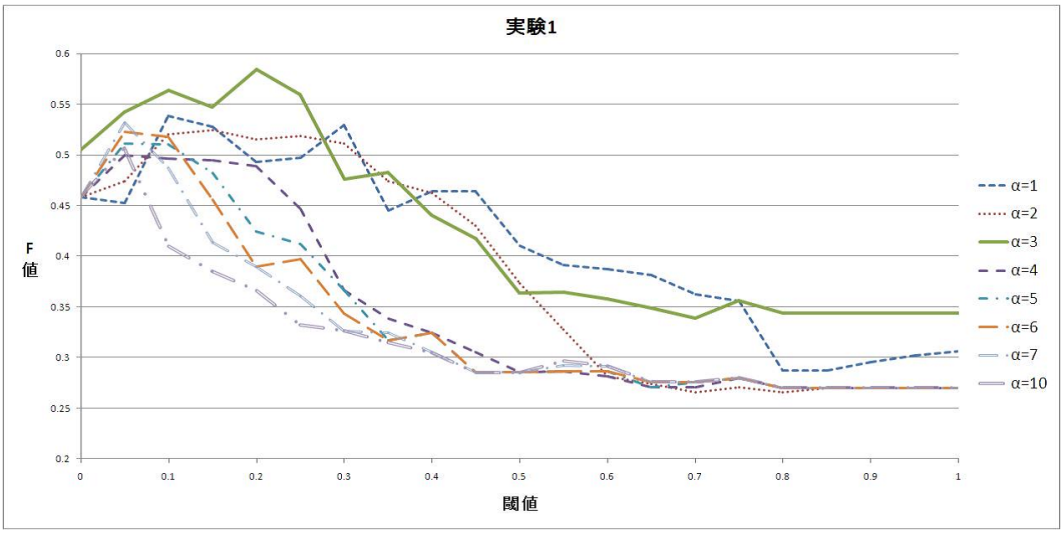

図 4　実　　験:1
Fig. 4　Experiment:1

表 2　各クエリの適合率，再現率，F 値

クエリ	コサイン類似度			提案手法		
	適合率 (%)	再現率 (%)	F 値	適合率 (%)	再現率 (%)	F 値
Bannock (food)	0%	0	0	20	50	28
Warwick Castle	15	1	27	25	100	40
Black dog (ghost)	67	29	40	100	29	44
Fish and chips	40	50	44	50	75	60
Goodwood Festival of Speed	0	0	0	50	50	50
Bowls	33	100	50	10	50	14
Blue plaque	20	100	33	0	0	0
Burlesque	60	50	55	100	67	80
Flag of Scotland	50	50	50	67	33	44
Gaelic handball	25	25	25	80	100	89
Kipper	67	67	67	1	67	80
National Gallery of Scotland	72	67	70	75	50	60
Lipton	0	0	0	25	100	40
Average	34	49	41	57	59	58

母国語版の Wikipedia と比較言語版の Wikipedia の記事をそれぞれ抽出し，記事の内容を比較した上でその差分情報をユーザに提示する手法を提案した．差分情報を抽出する際，母国語記事と比較言語記事の情報の粒度の違いを考慮し，我々の提案した関連度を用いて比較する複数の比較言語記事の抽出手法を提案した．そしてその提案手法の精度を計るため実験を行なった．また，記事をセグメントにわけコンテンツの比較をする手法の提案を行った．

- 評価実験

本論文では Wikipedia の記事の残存率に基づく質の算出手法を用いて記事の情報の質の測定について述べたが，今回それを用いた精度の実験を行わなかった．さらに提案手法の有用性を示す実験を行わなかった．今後はこれらの有用性を示す実験を行う．

- 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．そのため今後は，単語の多義性を解消させることで差分情報抽出手法の適合率を向上させることに取り組む．

- 差分情報の提示方法

本研究では，差分情報の抽出方法についての提案を行ったが，その提示方法については言及していない．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．そのため，差分情報の提示方法について考案する必要がある．

- 多言語を対象とした実験

本評価実験では日本語版記事と英語版記事を比較した．今後は比較対象を英語版のみならず中国語や韓国語等他の言語への対応を行いたい．また，二言語間の比較から三言語間の比較へと新たに取り組む．

文　　献

[1] Michael Strube and Simone Paolo Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. *American Association for Artificial Intelligence*, Vol. 2, pp. 1419–1424, July 2006.

[2] David Milne. Computing semantic relatedness using

wikipedia link structure. *New Zealand Computer Science Research Student Conference*, Vol. 7, p. 8, April 2007.

[3] Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. Wikipedia mining for an association web thesaurus construction. *Web Information Systems Engineering*, Vol. 4831, pp. 322–334, December 2007.

[4] Yuku Takahashi, Hiroaki Ohshima, Mitsuo Yamamoto, Hirotoshi Iwasaki, Satoshi Oyama, and Katsumi Tanaka. Evaluating significance of historical entities based on tempospatial impacts analysis using wikipedia link structure. *22nd ACM conference on Hypertext and hypermedia*, pp. 83–92, June 2011.

[5] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *7th international conference on World Wide Web 7*, pp. 107 – 117, April 1998.

[6] Jaap Kamps and Marijn Koolen. Is wikipedia link structure dierent? *2nd ACM International Conference on Web Search and Data Mining*, pp. 232–241, February 2009.

[7] David Milne, Olena Medelyan, and Ian H. Witten. Mining domain-specific thesauri from wikipedia: A case study. *The 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 442–448, December 2006.

[8] Zheng Chen, Shengping Liu, Liu Wenyin, Geguang Pu, and Wei-Ying Ma. Building a web thesaurus from web link structure. *26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 48–55, July 2003.

[9] Eytan Adar, Michael Skinner, and Daniel S. Weld. Information arbitrage across multi-lingual wikipedia. *2nd ACM International Conference on Web Search and Data Mining*, pp. 94–103, February 2009.

[10] Thomas Wohner and Ralf Peters. Assessing the quality of wikipedia articles with lifecycle based metrics. *5th International Symposium on Wikis and Open Colaboration*, pp. 1–10, October 2009.

[11] Meiqun Hu. Measuring article quality in wikipedia: Models and evaluation. *16th ACM conference on Conference on information and knowledge management*, pp. 243–252, November 2007.

[12] 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. 第 21 回セマンティックウェブとオントロジー研究会, pp. 20–32, 11 月 2009.