情報処理学会論文誌　データベース　Vol. 3　No. 3　20–32 (Sep. 2010)

Wikipediaにおけるキーパーソン抽出による信頼度算出精度および速度の改善

鈴木　優[†1]　吉川正俊[†2]

本研究では，Wikipediaにおいて記事の信頼度を算出する際に，重要となる著者であるキーパーソンを簡易な方法で推定し，それら重要な著者の情報だけを利用して信頼度を算出することによって，すべての著者の情報を利用して信頼度を算出する方法よりも高速で精度の高い信頼度を算出する手法の提案を行う．これは，記事の大部分は少数の著者によって記述されているため，多くの著者の編集はその記事の信頼度には影響しないと考えることができ，それら多くの著者が行った編集を信頼度算出に用いないことにより，信頼度の算出にとって不要なノイズを除去することができると考えたためである．評価実験において信頼度が正しく算出できたかどうかを確かめた結果，確かに信頼度の精度が向上したことを確認することができた．さらに，提案手法を用いることにより，信頼度を計算するための計算コストを削減することも可能となった．

An Improvement Method of Calculation Cost for Assessing the Quality of Wikipedia Articles

Yu Suzuki[†1] and Masatoshi Yoshikawa[†2]

We propose a fast credibility assessment system of Wikipedia articles by identifying major contributors to reduce the calculation costs in determining the degree of credibility of Wikipedia articles. In our proposed system, similar to existing credibility degree measuring systems, the first calculates each editor's credibility values using the lifetime of versions, which is a number of versions includes the edits. Next, our system calculates the credibility values of articles by combining those of the article's editors. In this process, when the system identifies a small number of major contributors who have large effects to credibility degrees of articles, we can reduce calculation costs. Therefore, we propose three methods for identifying major contributors, such as number of versions based method, number of distinct document based method, and combined method of number of versions based method and number of distinct document based method. In our experimental evaluation, we unveil that our proposed system can reduce the calculation costs and increase the accuracy of credibility values of articles.

1. はじめに

Wikipedia[⋆1]は，現在最も成功している百科事典の1つであり，Wikiにより作成されている．Wikipediaの1つの特徴として，誰もがページの編集を行うことができるという点があげられる．ところが，これらの記述はつねに正しいとは限らない．つまり，もしある編集者が誤った記述を追加した場合でもWikipediaはその記述をそのままページに反映させるため，Wikipedia上の記事は必ずしも信頼できる情報だけで構成されていない[1)]．さらに，Wikipediaの記述量は年々増加傾向にある[2)]ため，人手によって信頼度を判定することは困難な作業となりつつある．

Wikipediaの信頼性に関する問題点を解決するために，多くの編集者によって信頼度の低い情報や不適切な情報の削除，編集が行われている．ところが，Wikipediaに含まれる記事の記述量が多くなるに従って，すべての記事に対して信頼度を保つことが困難になりつつある．なぜなら，記事の記述量と質は必ずしも比例しないと考えられるためである[3)]．この問題を解決するために，記事の信頼度を自動的に算出し，編集者がどの記事に対して改善を行う必要があるかを示すアプリケーションの必要性が高まりつつある．

記事の信頼度は，Wikipediaを信頼できる百科事典として利用している利用者にとっても必要である．もし利用者が信頼できない記事を信頼できる記事であると誤認したとき，その誤認が社会的な問題を引き起こす可能性もある．利用者は一般に，Wikipediaに記述されている記事のうち未知の情報を検索することも多い．そのとき，利用者はWikipediaの記述を信頼できるかどうかを判断すること自体が困難であり，既知の利用者にとって容易に信頼できないと判定できる記述を既知でない利用者が信頼してしまう可能性はきわめて高い．

この問題を解決するために，利用者による投票，投稿を利用した情報評価手法が利用されている．たとえば動画投稿サイトであるYouTube[⋆2]では，アップロードされた動画が良いかどうかを利用者が5段階で評価するシステムを導入しており，通信販売サイトである

†1 名古屋大学情報基盤センター
Information Technology Center, Nagoya University
†2 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
⋆1 http://ja.wikipedia.org/
⋆2 http://www.youtube.com/

Amazon.com[⋆1]においても同様のシステムを導入している．これら利用者による投票を利用したシステムでは，利用者が的確に記事に対して評価を行うことが，的確な信頼度判定のために必要である．ところが，YouTube に関する調査[4] では，ほとんどの利用者が検索対象に対して最高点である 5 点を投票しており，他の点数をほとんど付与していないことが分かっている．YouTube で公開されている動画がすべて良質な動画であるとは限らないことから，これらの点数は実際の動画の質を反映しているとはいえないことが分かる．そのため，これらの手法では正確な質の評価を行うことが困難である．

そこで我々は，著者相互で評価を行うことによる信頼度算出手法を利用する．Adler ら[5)–7)] や Hu ら[8] は，著者の信頼度を算出することによって記事の信頼度を算出するための手法を提案している．我々の提案手法はこれらの手法を基準としている．これらの手法では，記事の残存率に着目した信頼度の算出を行っている．つまり，ある著者が記述した部分が長い間残存しているとき，その記述の信頼度は高く，その記述を行った著者の信頼度が高いという仮定である．この仮定を利用することによって，記事の残存度から著者の信頼度を算出し，記事の信頼度を算出することができる．

この手法における問題点として，貢献度の小さな著者の取扱い方があげられる．Wikipedia では，多くの著者は少量の編集を少ない回数だけ行っているため，それらの著者の Wikipedia に対する貢献度は小さいと考えられる．ところが，それらの著者の信頼度を記事の信頼度に反映させたとき，貢献度の大きな著者数よりも小さな著者数のほうが多いため，貢献度の大きな著者の信頼度が記事の信頼度に相対的に反映されにくいという問題がある．つまり，貢献度の小さな著者の信頼度がノイズとなってしまう．

もう 1 つの問題は計算コストである．現在までに提案されている信頼度算出手法では，信頼度を算出するために多くの時間がかかる．なぜなら，Wikipedia の編集履歴はしばしば長大となることがあり，1 人の著者が記述する記事の数も大量となる場合があるため，システムがこれら大量の編集履歴をさかのぼって調査する必要があるためである．

これら 2 つの問題点を解決するために，我々は信頼度を算出する著者数を削減する手法を提案する．なぜならば，記事の信頼度に大きな影響を与える著者の数は全体の著者数と比較して非常に小さいと考えたためである．つまり，もし我々が記事の信頼度に影響を与える著者を簡単な方法で特定することが可能となれば，貢献度の小さな著者に対する扱いを改善することができ，しかも記事の信頼度算出に必要な計算コストを下げることができる．本論文では以後，このような記事の信頼度に大きな影響を与える著者を**キーパーソン**と呼ぶ．

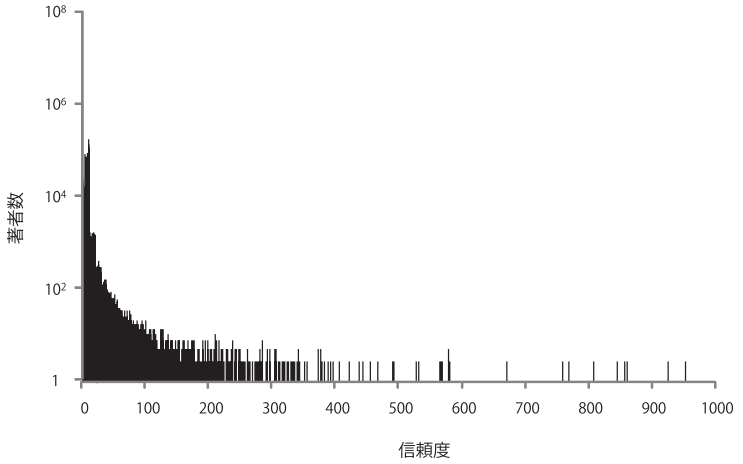

図 1　信頼度と著者数との関係
Fig. 1　Number of authors and the authors' credibility values.

本論文では，キーパーソンを簡易な方法で特定することによって，記事の信頼度算出に必要な計算コストを削減する方法について述べる．**図 1** に，3 章に述べている著者相互で評価を行うことによる信頼度算出手法で計算した著者の信頼度と，5 章に述べている評価実験で用いたテストセットにおける著者数との関係を表したグラフを示す．この図では，x 軸に信頼度，y 軸に x 軸で示された信頼度を持つ著者の数を示している．このグラフにおいて，信頼度が比較的高い著者の数は非常に少ないことが分かる．このように，すべての著者に対して信頼度を求めた後であればキーパーソンを抽出することは容易である．ところが，信頼度を計算するためには大量の計算が必要である．そこで，4 章で提案する，信頼度を計算する前にキーパーソンを抽出する方法を利用することによって，計算量の削減を行う．

さらに本提案手法では，キーパーソンだけを記事の信頼度算出に利用することによって，記事の信頼度に関する精度が向上すると考えている．我々の調査によると，日本語版 Wikipedia において約 20%の著者によって約 80%の記述が行われていることが分かった．そのため，20%の著者に対してだけ信頼度を付与した場合であっても，80%の記述に対して信頼度を算出することができるため，理論的にはほぼすべての記事に対して約 20%の計算量で信頼度を算出することができると考えられる．

⋆1 http://www.amazon.com/

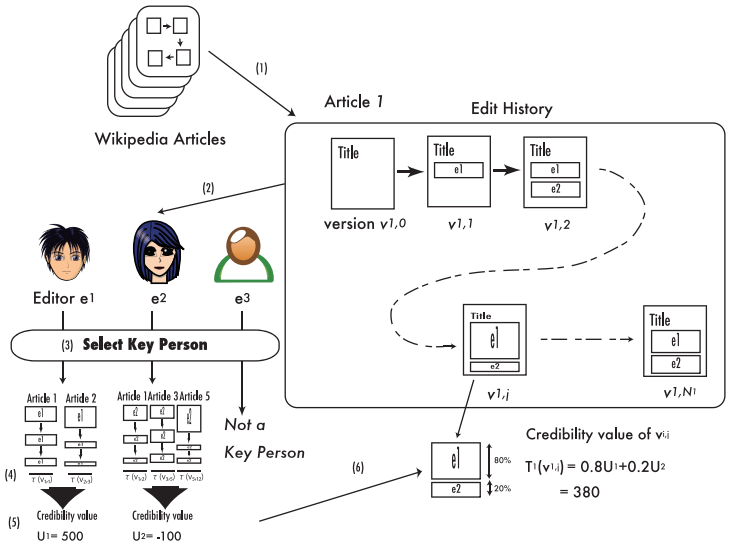

図 2　提案手法の概要
Fig. 2　Overview of our proposed method.

提案手法の概要を**図 2** に示す．まず，システムは Wikipedia の編集履歴データからそれぞれの記事を取り出す．次に，4 章で述べるキーパーソン抽出手法を利用してこれらの記事群から著者の特徴量を抽出し，キーパーソンを特定する．そして，キーパーソンが記述した記述量から，3 章で述べている，従来手法である著者相互の評価により信頼度を算出する手法を利用して，キーパーソンに対してだけ著者の信頼度を算出する．最後にすべての記事に対して，それぞれの記事を編集した著者の信頼度から記事の信頼度を算出する．

図 2 の (3) の部分でキーパーソンを特定する際に，著者の特徴量を抽出しなければならない．ここで抽出する特徴量は 2 つの要件があり，算出される信頼度と相関関係が高い値であること，信頼度を算出するよりも低い計算コストで得られる値であることの 2 つである．そこで本研究では著者の記述量，著者が記述した記事数，それら 2 つの値の組合せである 3 つの特徴量算出手法を提案した．5 章では，特徴量算出に必要な計算コスト，算出された記事の信頼度の精度の 2 つを評価実験において評価し，提案手法が有効であることを示す．

2. 関 連 研 究

Wikipedia や特定分野の文書に対して信頼度を算出する手法は，現在までに数多く提案されている[9]．たとえば，論文誌や国際会議に投稿された論文に対して査読を行う作業は，論文に対して信頼度を人手で算出する作業であるといえる．ここでは，自動的もしくは半自動的に Wikipedia に対して信頼度を算出する方法に関する研究について述べる．

Wikipedia に対して信頼度を算出する方法として利用者の投票を利用した方法，自然言語処理による方法，編集履歴を利用する方法の 3 つが主に考えられる．以下，それぞれの方法について述べる．

投票を利用した Wikipedia の信頼度算出方法として，Kramer らの方法[10] がある．この手法では，MediaWiki[⋆1] に対して利用者による記事への投票システムを付加している．このシステムでは，利用者はどの記事の質が高いかを利用者自身で判定し，システムに入力することによって，どの記事の質が高いかを閲覧者が容易に知ることができる．ところが，このシステムの問題は十分な数の利用者が記事の質を判定しなければ十分な精度の信頼性を求めることが困難であること，すべての利用者が的確に記事の質を判定することが困難であることの 2 つである．我々の提案システムでは，利用者は記事に対して質を判定する必要がないため，利用者の手間を軽減することができ，精度の高い信頼度算出を行うことができる．

現在一般的に利用されている，学術論文の査読システムにおける半自動査読システムを Mizzaro ら[11] が提案しており，このシステムを Wikipedia に対して応用した方法を Cusinato ら[12] が提案している．この手法では，著者の編集を査読の 1 つであると考え，ある著者が記述を削除したときに，その記述を述べた著者に対して負の評価を行ったと考える．逆に，その著者が記事を削除せずそのままにしたときには，その記述を述べた著者に対して正の評価を行ったと考える．この方法は我々の提案手法に近い方法である．ところが，この手法では非常に多くの計算時間がかかる．さらに，削除した文字数などは判定していないため，我々の提案手法と比べて記事から抽出する特徴が少ないため，信頼度の精度が低下する可能性がある．

Wöhner ら[13] は，記事編集の周期的な変化に着目することによって，典型的な記事編集の周期に対して信頼度を算出する方法を提案している．この方法では，著者の編集量の変化と信頼度には関連があることに着目している．ところが，この方法では記事の量そのものだ

⋆1 MediaWiki は Wikipedia で利用されている Wiki システムである．http://www.mediawiki.org/

けに着目しており，記事を記述した著者は考慮されていないため，新しい記事に対して信頼度を算出することができないこと，編集合戦が行われたときに信頼度が低下してしまうという問題点がある．我々の手法では著者を考慮した信頼度の算出を行っているため，新しい記事に対して信頼度を算出することができ，しかも編集合戦が行われたときにも適切な信頼度を算出することができる．

Adler ら[5)–7)] や Hu ら[8)]，Wilkinson ら[14)] は，編集履歴を利用することによって信頼度の算出を行っている．これらのシステムでは，すべての著者に対して信頼度を算出している．我々の提案手法と Adler らの手法は，信頼度算出手法の観点からは類似した方法である．ところが，我々の手法ではキーパーソンだけに対して信頼度を算出する点が大きく異なる点である．つまり，これら既存研究における手法と比較して，信頼度の精度が向上する点と計算コストが削減された点が大きく異なる点である．

3. 信頼度の算出手法

Wikipedia の記事におけるそれぞれのバージョンに対して信頼度を算出するために，まず変更の妥当性を示す信頼度を算出する．次に，著者に対して信頼度を算出し，最後に記事のバージョンに対して信頼度を算出する．まず本章では，Wikipedia の記事のモデル化を行い，信頼度算出の定式化を行う際に必要な変数を定義する．定義したモデルを基に，記事のバージョンに対して信頼度を算出する方法を述べる．

3.1 Wikipedia のモデル化

まず，本論文で利用する変数の定義を行う．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 はすべての記事のすべ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ない状態を表す．Wikipedia の記事を記述した著者 $e = 1, 2, \cdots, K$ は，1 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，1 つのバージョンを作成した著者は 1 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．

3.2 記事の変更に対する信頼度

まず，$v_{i,j}$ が妥当な編集であったかどうかを調べ，$v_{i,j}$ における記事変更における信頼度である記事変更信頼度 $r(v_{i,j})$ を算出する．ここで妥当な編集の定義として，2 章において示した Adler らの定義を利用している．つまり，妥当な編集とは他の著者による編集後の

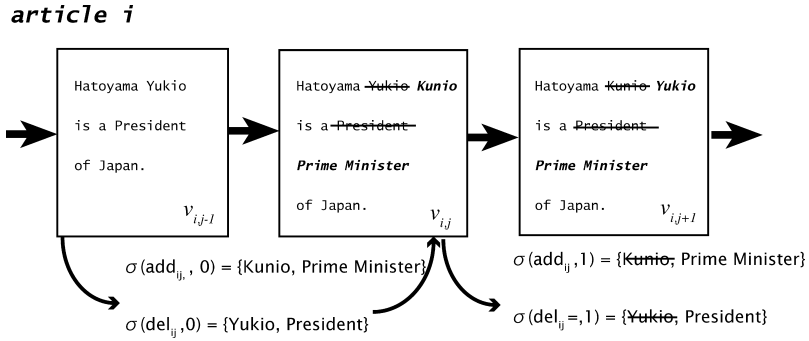

図 3　編集履歴における追加と削除
Fig. 3　An example of added and deleted contents in page edit history.

残存文字数，削除文字数が小さな編集である．これは，著者がもし妥当な文字の追加を行った場合には，他の著者はその追加した文字を削除する可能性が低いためである．同様に，妥当な文字の削除を行った場合には，他の著者はその削除した文字を再び追加しないとする．

図 3 に示す例を利用して，追加と削除に関する信頼度算出手法を説明する．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．この例の場合，“Kunio”，“Prime Minister” は $add_{i,j}$ に含まれ，“Yukio”，“President” は $del_{i,j}$ に含まれる．

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 $R^{add}(i, j, p)$，$R^{del}(i, j, p)$ を式 (1)，(2) によって求める．

$$R^{add}(i, j, p) = \frac{|\delta(add_{i,j}, p)|}{|add_{i,j}|} \tag{1}$$

$$R^{del}(i, j, p) = \frac{|\delta(del_{i,j}, p)|}{|del_{i,j}|} \tag{2}$$

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．

図 3 における例では，$p = 1$ のときの追加残存率 $R^{add}(i, j, 1)$ と削除残存率 $R^{del}(i, j, 1)$

を算出する．この場合，$add_{i,j}$ には空白文字を除くと 18 文字含まれており，$\delta(add_{i,j}, p)$ には 13 文字含まれているため，$R^{add}(i,j,1) = \frac{13}{18} = 0.72$ となる．同様に，$del_{i,j}$ には 14 文字含まれており，$\delta(add_{i,j}, p)$ には 9 文字含まれているため，$R^{add}(i,j,1) = \frac{9}{14} = 0.64$ となる．

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，削除されたとき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あげられる．正規化された残存率 $\overline{R^{add}(i,j,p)}$，$\overline{R^{del}(i,j,p)}$ を式 (3)，(4) によって求める．

$$\overline{R^{add}(i,j,p)} = \frac{R^{add}(i,j,p)}{S^{add}(p)} \tag{3}$$

$$\overline{R^{del}(i,j,p)} = \frac{R^{del}(i,j,p)}{S^{del}(p)} \tag{4}$$

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．図 3 における例では，予備実験の結果 $S^{add}(1) = 0.93$，$S^{del}(1) = 0.99$ であるため，それぞれ $\overline{R^{add}(i,j,1)} = \frac{0.72}{0.93} = 0.77$，$\overline{R^{del}(i,j,1)} = \frac{0.64}{0.99} = 0.64$ となる．

そして，追加残存率と削除残存率を組み合わせて，記事変更信頼度 $\tau(v_{i,j})$ を求める．記事変更信頼度は，追加残存率と削除残存率の総和であり，式 (5) で求める．

$$\tau(v_{i,j}) = \sum_{q=1}^{N_i - j} R^{add}(i,j,p) + \sum_{q=1}^{N_i - j} R^{del}(i,j,p) \tag{5}$$

記事変更信頼度を算出するとき，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信頼度は高くなるべきであると考えたためである．図 3 における例では，$\tau(v_{i,j}) = \overline{R^{add}(i,j,1)} + \overline{R^{del}(i,j,1)} = 0.77 + 0.64 = 1.41$ となる．

最後に，記事変更信頼度の平均を 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，記事自体の信頼度は変化しないと考えられる．それらの記事変更信頼度の変化よりも低い記事変更信頼度があったとき，その変更は記事の信頼度を低下させていると考えられるためである．最終的な記事変更信頼度 $\overline{\tau(v_{i,j})}$ は式 (6) で求める．

$$\overline{\tau(v_{i,j})} = \tau(v_{i,j}) - \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i}\tau(v_{i,j})}{\sum_{i=1}^{M} N_i} \tag{6}$$

3.3　著者の信頼度

著者 e の信頼度 U_e を，e の編集した記事の割合から算出する．まず，3.1 節で述べたように，e の編集したバージョンの集合を A_e と定義している．U_e は式 (7) で計算される．

$$U_e = \frac{\sum_{v_{i,j} \in A_e} \overline{\tau(v_{i,j})}}{|A_e|} \tag{7}$$

ここで，$|A_e|$ は V_e に含まれるバージョンの数であり，利用者が編集した記事の総数である．

3.4　記事の信頼度

最後に，記事のバージョン $v_{i,j}$ における信頼度 $T_i(v_{i,j})$ を求める．記事の信頼度は，その記事を記述した著者の信頼度を，その記述量による重み付き平均によって算出する．つまり，$T_i(v_{i,j})$ は，式 (8) で計算される．$v_{i,j}$ を記述している著者の集合 $E(v_{i,j}) = \{e | e \in E\}$ を利用して計算する．

$$T_i(v_{i,j}) = \frac{\sum_{k=1}^{j} U_e \cdot c_{e,k}}{\sum_{k=1}^{j} c_{e,k}} \tag{8}$$

ここで $c_{e,j}$ は著者 e の記述が，バージョン j において残存している文字数である．

4. キーパーソンの特定

3 章における記事の信頼度算出において最も計算時間がかかる処理は，3.2 節で述べた記事変更信頼度算出である．そこで，もしすべての記事変更ではなく一部の記事変更だけに対してだけ記事変更信頼度を算出し，記事の信頼度に大きな差がなければ，記事の信頼度を高速に算出することができると考えた．そこで，3.4 節における式 (8) において，U_e の値が 0 に近い値を持つ著者 e を特定する．そして，U_e の値が十分大きな，もしくは小さな著者が編集した記事変更だけに対して記事変更信頼度を算出する．

本章で述べる処理は，3 章で述べた手順のうち，3.2 節の処理の前に行う処理である．つまり，$\overline{\tau(v_{i,j})}$ を算出するかどうかを判定するための方法を述べる．ここで，$v_{i,j}$ はどの著者が編集を行ったバージョンであるかは，容易に分かるものとする．

4.1　著者の信頼度順推定

4.1.1　手法 1：記述量によるキーパーソンの特定

まず，記述量によりキーパーソンを特定するための方法について述べる．記述量が多い著者は，Wikipedia に対して多くの影響を与えていると考えられるため，著者の信頼度にも

影響があると考えられる．

利用者 e が記事 i に対して $f(i,e)$ バイトの記述を行ったとき，その著者 e の重要度 $I_1(e)$ を次の式で計算する．

$$I_1(e) = \sum_{i=1}^{M} f(i,e) \tag{9}$$

この手法では，最も新しいバージョンに多く記述をしている著者をキーパーソンとして抽出する．ところがこの手法では，過去に多くの記述を行っていたが新しいバージョンではあまり記述をしない著者は，その著者が信頼度の多い記述を行っていてもキーパーソンとは判断されない点が問題である．

4.1.2　手法 2：記述した記事の数

この手法は，著者が記述した記事の数を著者の重要度であると考える手法である．この手法では，記述量は考慮しない．

まず，記事 i に著者 e が記述を行っているかどうかを示す変数 $P(i,e)$ を用意する．

$$P(i,e) = \begin{cases} 1 & \text{if } e \text{ edits } i \text{ more than once} \\ 0 & \text{else} \end{cases} \tag{10}$$

この変数 $P(i,e)$ を利用して，著者の重要度 $I_2(e)$ を算出する．

$$I_2(e) = \sum_{i=1}^{M} P(i,e) \tag{11}$$

この手法では手法 1 と異なり，最新バージョンの記事だけではなくすべてのバージョンを対象とした計算が必要となるため，計算量が手法 1 よりも上昇する．ただし，すべての著者に対して信頼度を算出するために必要な計算コストと比較すると，手法 2 で必要な計算コストはきわめて小さい．

手法 1 と比較すると，この手法では過去に信頼度の高い記述を行った著者も正しくキーパーソンとして評価することができる．ところが，この手法では非常に多くの微細な修正を行った著者も誤ってキーパーソンとして扱ってしまうという問題点がある．

4.1.3　手法 3：記述量と記述した記事の数の組合せ

この手法では，記述した記事の数が少なく，記事の量が多い著者がキーパーソンであると考える方法である．この手法は TFIDF による重み付け手法と似た考え方である．つまり，一つの記事に対して多くの記事量を投稿している場合には，その著者はある分野における記事に対して高い知識を持っていると考えられる．そのため，このような著者は Wikipedia の信頼度に対して影響を与えていると考えられる．

利用者 e の重要度 $I_3(e)$ は式 (12) で算出する．

$$I_3(e) = \sum_{i=1}^{M} f(i,e) \cdot -\log \frac{\sum_{i=1}^{K} P(i,e)}{K} \tag{12}$$

この手法は手法 1 と 2 を組み合わせたものである．そのため，この手法では過去に信頼度の高い記述を行った著者も，微細な修正を数多く行う著者も正しく判断を行うことができる．ところが，この手法によって $I_3(e)$ を計算するために必要な計算コストは手法 1 と 2 の和となるため，比較的多くの計算が必要となる．ところが，手法 1 と手法 2 における計算は並列に行うことができるため，実際には手法 1，2 のうち計算コストの大きいものと同じ計算時間で $I_3(e)$ を算出することができる．

4.2　キーパーソンの特定

4.1 節で示した手法を利用し，キーパーソンの特定を行う．ここで，まず 3 つの手法により計算された指標 $I_1(e)$，$I_2(e)$，$I_3(e)$ のうち 1 つを選択し，$I(e)$ とする．

$I(e)$ の値が高い著者から順に並べた著者のリストを作成する．次に，最も $I(e)$ が高い著者から順に t%の著者を特定する．ここで，これらの著者をキーパーソンとして指定する．そして，3 章に述べた方法によって，キーパーソンだけに対して信頼度を算出し，最後に記事の信頼度を算出する．

我々は，$I(e)$ を降順に並べた著者順と 3 章の手法によって算出した著者の信頼度によって降順に並べた著者順には相関関係が高いと考えている．そこで，次の章においてこの仮定が実際に正しいかどうかを検証した．

4.3　予 備 実 験

予備実験では，信頼度とこれら 3 つの指標との相関関係が実際に存在することを示す．信頼度とそれぞれの指標との相関関係が高いときには，信頼度の代わりにこれらの指標を用いて特定の著者を選択することが可能となるために，計算量の削減が可能となることを確かめることができる．

予備実験は次のような方法で行った．

(1)　編集履歴データを利用してすべての著者の信頼度を 3 章で述べた手法によって算出し，信頼度の高い著者から順に著者 ID を並べる．

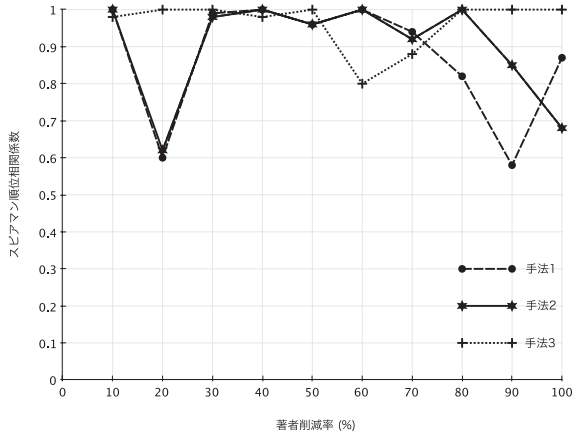

図 4　3 つの方法によって算出された著者の順位と Adler らの方法による著者の順位とのスピアマン順位相関係数
Fig. 4　Spearman's rank correlation coefficient for three methods vs. the method using all editors' rank.

(2)　すべての著者に対して，3 つの指標によって順位を算出し，それぞれの順に著者 ID を並べる．

(3)　手順 (1) において算出された信頼度と手順 (2) において算出された 3 つの指標による順位との相関関係を，それぞれスピアマン順位相関係数によって求める．

評価実験のための編集履歴データとして，5 章で利用しているデータである 2010 年 3 月 28 日の日本語版 Wikipedia データ[⋆1]すべてを利用した．

信頼度が上位である部分の順位相関係数を調べるために，信頼度に関する上位 k%の著者を抽出し，その著者が手順 (2) によって順位付けされた場合の順位を調べ，スピアマンの順位相関係数を算出した．実験結果を**図 4** に示す．

この図から，3 つの特徴量と信頼度には相関関係があることが分かり，提案手法が有効に機能する可能性があることが分かった．特に著者の記述量と記述文書数の組合せによる著者順は他の著者順と比較して相対的に相関関係が高いことが分かった．この理由として，3 章で示した信頼度算出手法が，記述量や記事数と相関関係があることがあげられる．また，過去に信頼度の高い記述を行った著者，多くの微細な修正を行った著者による相関係数の低下は小さいことが分かった．これは，今回扱った編集履歴データには，このような著者があまり存在しなかったことが原因としてあげられる．

また，図 4 において著者削減率を 20%にした際に，手法 1 と 2 においてスピアマン順位相関係数が 0.6 程度まで低下した．つまり，著者の順位の差異が著者順の上位 10%から 20%の部分で大きく異なっていたことが分かる．これは，手法 1，2 ではこの著者順の区間で異なる順を算出してしまったためである．つまり，著者順の上位 10%から 20%の区間では，信頼度の高い著者のうち記述量もしくは記述記事数のどちらかが大きく，Adler らの方法により信頼度が高い著者が多かった．

5 章では，実際に評価実験を行い提案手法が有効であることを示す．

5. 評 価 実 験

提案手法の有効性を調べるために，信頼度を計算するための時間と計算された信頼度の精度を測定した．まず評価実験に関する実験条件について示し，次に計算時間における評価を示し，最後に精度に関する評価について述べる．

5.1 実 験 準 備

我々は，以下の手順で評価実験を行った．

(1)　データセットからの特徴量抽出
我々は次の 3 つの特徴量を編集履歴データから取り出した．
- タイトル ID
- 著者 ID
- 記述

(2)　上の特徴量から，3 つの著者順位表を作成した．
- 著者の記述量による著者順位表．4.1.1 項で述べた．
- 著者の記述した記事数による著者順位表．4.1.2 項で述べた．
- 著者の記述量と記事数の組合せによる著者順位表．4.1.3 項で述べた．

(3)　求められた 3 つの著者順位表から，それぞれ上位 k 件の著者を求める．これらの著者をキーパーソンとする．

(4)　それぞれの方法で求められたキーパーソンに対して，著者の信頼度を算出し，記事の信頼度を算出する．

(5)　信頼度の高い記事から順に利用者へ提示する．

⋆1 http://download.wikipedia.org/jawiki/20100328/pages-meta-history.xml.bz2

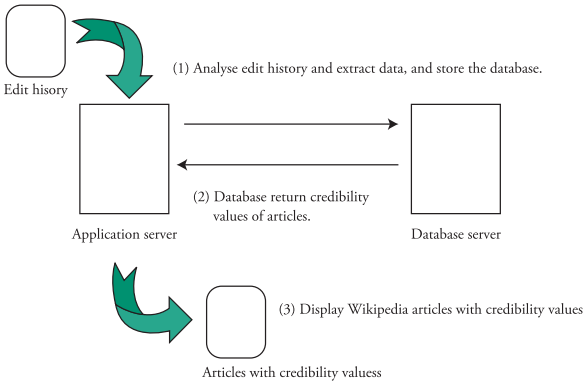

図 5　提案手法の実装
Fig. 5　Implementation of our proposed system.

図 6　提案手法の実装イメージ
Fig. 6　User interface of our proposed method.

5.2 実　　装

本論文で提案した手法を基に，実装を行った．実装は 2 つの部分からなり，一つは信頼度算出であり，Wikipedia の編集履歴データから特徴量を抽出し，信頼度を計算する部分である．もう 1 つはユーザインタフェースであり，計算された信頼度を記事とともに表示する部分である．

5.2.1 信頼度算出

図 5 に示すように，我々は 2 つの計算機を利用して提案手法の実装を行った．1 つの計算機はデータを格納するためのデータベースサーバとして利用した．この計算機は 2 つの Intel Xeon 3.06 GHz プロセッサ，4 GByte メモリ，500 GByte のハードディスクが実装されている．この計算機には Mac OS 10.6.3（Snow Leopard）上に MySQL 5.1.38 の MyISAM データベースエンジンが実装されている．もう 1 つの計算機は，編集履歴データから特徴量を抽出するために利用した．この計算機は 1 つの Intel Core i7 プロセッサ，16 GByte のメモリ，80 GByte のハードディスクが実装されている．この計算機には Microsoft Windows 7 オペレーティングシステム上の Java JDK 1.6.0_14 を利用して実装した．

5.2.2 ユーザインタフェース

図 6 に提案手法の実装画面を示す．利用者は Wikipedia の記事内容とともに，その記事の信頼度を表示している．記述部分は赤色，青色で表示されている．青色の部分は信頼度が高い部分であることを示し，赤色の部分は信頼度が低いことを示し，黄色の部分は信頼度が計算されていない部分であることを示している．

画面の右上には，表示されている記事全体の信頼度と，記事における信頼度の高い部分，低い部分，不明な部分の割合を表示している．利用者はこの部分を閲覧することにより，直感的に記事の信頼度を確かめることができる．

このユーザインタフェースは，Ruby on Rails 2.3.5，MySQL 5.1.38，Apache 2.2.14 を利用して実装されている．利用者はキーワード検索インタフェースによってキーワードを入力すると，利用者はそのキーワードに関連する記事の一覧を取得することができる．利用者は一覧の中から必要な記事をクリックすると，システムはそのページの最新バージョンを，記述の信頼できる部分を色付けして表示することができる．利用者は，記事のバージョンを選択することによって，異なるバージョンの記述とその信頼度を閲覧することもできる．

5.2.3 評価実験に利用したデータ

我々は 2010 年 3 月 28 日における日本語版 Wikipedia の編集履歴データを利用した．このデータには 1,138,433 件の記事，4,456,066 件の編集が含まれている．この登録利用者のうち 246,540 人は少なくとも 1 カ月間に 1 回以上の編集を行った著者である．著者数は 1,179,867 人であり，この中には登録利用者ではないために IP アドレスで表現された著者も含まれている．また，3 月 28 日時点では削除されたため閲覧することができないが，過去に閲覧可能であった記事も存在する．

我々の提案手法は字句解析を行う手法ではなく，文字数だけを利用しているため日本語版以外の Wikipedia に適用することが可能である．ところが，英語版の Wikipedia では編集履歴が公開されていなかったことや，他の言語では信頼性が存在する記事かどうかを人手で判定することが困難であることから，日本語版の Wikipedia データを評価実験に利用した．もし英語版のデータが公開された場合には，英語版の Wikipedia データを利用して実験を行う予定である．

抽出された記事の中から，特殊な目的で作成された記事を対象から除外した．これら除外した記事は，“Wikipedia:”，“Help:”，“Template”，“利用者:”，“ファイル:”，“MediaWiki:”，“Category:”，“Portal” などがタイトルに含まれる記事である．他に，曖昧さ回避のために利用されているページや，記事へのリンクが列挙されているページ，テンプレートなども対象から除外した．

さらに，我々は著者の中からボットと呼ばれる，Wikipedia の記事を自動的もしくは半自動的に編集するプログラムを示す利用者を，著者リストから除外した．これは，ボットは記事の内容を判定しているわけではなく機械的な作業しか行っていないためである．これらボットのリストは Wikipedia に示されたリスト[⋆1]を利用した．一方，我々は匿名の利用者である IP アドレスで表示された利用者は，除外しなかった．なぜならば，匿名の利用者であっても有用な編集を行う可能性は高いと考えたためである．

本実験では性能比較の基準となるシステムとして，3 章で述べた方法だけを利用した，すべての著者に対して信頼度を算出する手法を利用した．

5.3　評価実験 1：計算コスト

計算コストがどの程度削減されるかどうかを確かめるために，評価実験を行った．5.1 節において，評価実験の手順を述べた．この手順の中で，手順 (1) で示されている特徴量の抽出部分は，本提案手法で削減できる手順ではない．そのため，手順 (2) から (4) までに示されている部分について比較を行った．**図 7** に実験結果を示す．

まず，手順 (1) における計算時間を測定した．計測の結果，約 1,280 日の時間が必要であった．この時間は主に，圧縮された gzip データを解凍するための時間，および XML データを解析するための時間であった．本実験では 62 並列で解析を行ったため，実際には約 20 日の時間がかかった．

次に，図 7 において，特徴量から記事の信頼度を算出するための時間を示している．この図では，キーパーソンの著者全体における割合を 10%から 90%まで 10%刻みで変化させてゆき，それぞれ計算時間を計測した．100%はキーパーソンを算出しなかった方法であり，既存手法における計算時間を示している．

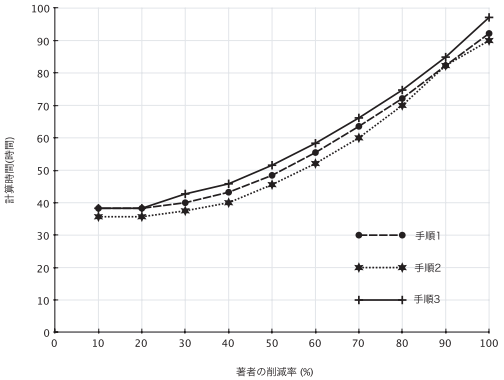

図 7　著者の削減率と計算時間との関係
Fig. 7　Editor reduction ratio and calculation time.

この図から，信頼度を計算する著者の割合と計算量は，著者の削減率が 40%から 100%の間では比例することが分かった．ところが，著者の削減率を 30%以下に削減しても，それほど計算量を削減することができなかった．原因として，貢献度の大きな著者は信頼度算出に時間がかかるため，このような著者だけがキーパーソンとして算出されてしまったとき，著者 1 人あたりに必要な平均計算量が相対的に増加するためであるといえる．

また，計算方法によって大きく計算量に差が出ないことが分かった．このことから，提案手法はどの手法であっても十分に小さな計算量で計算ができることが分かった．以上の結果から，提案手法を用いることにより確かに計算コストが削減されたことが分かった．

5.4　評価実験 2：信頼度の精度

最後に，提案した信頼度が Adler らの手法と比較して大きく低下することはないことを確かめるために，信頼度の精度を比較した．まず，著者への信頼度算出率を低下させたときの精度を調査するために，Wikipedia の著者による投票で決められる「秀逸な記事」および「良質な記事」を信頼度の高い記事であると仮定した．そして，それらの記事群を記事の信頼度順の上位に順位付けすることができるかどうかを確かめた．次に，信頼度を算出した記事群の中から無作為に 100 件の記事に対して実際に人手による評価を行い，利用者が判断

⋆1 http://ja.wikipedia.org/w/index.php?title=特別：登録利用者一覧&group=bot

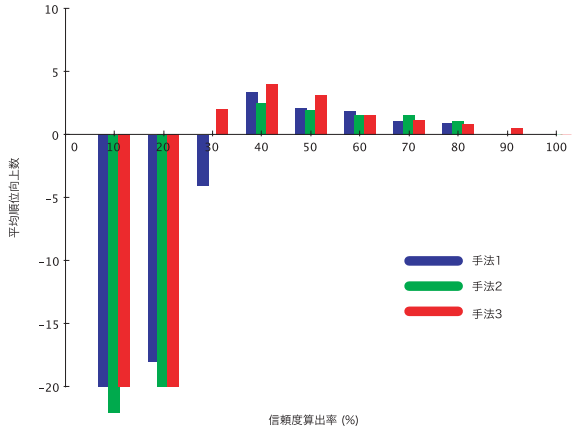

図 8　キーパーソンの割合を変化させたときの 3 つの手法による平均順位向上数
Fig. 8　Average order improvement degree vs. calculation ratio of key persons.

した信頼度とシステムが出力した信頼度に相関関係があるかどうかを調べた．

5.4.1　評価実験 2-1：「秀逸な記事」,「良質な記事」による評価

本実験では，信頼度が高い記事に着目し，提案手法により計算された信頼度が高い順に記事を並べ，3 章において算出された方法による記事の順位と比較してどの程度順位が向上したかを調査した．信頼度が高い記事として，日本語版 Wikipedia で公開されている「秀逸な記事」および「良質な記事」を利用した．これらのリストには合計して 446 件の記事が選択されている．

我々は提案手法を評価する際に，情報検索で一般に利用されている再現率や適合率を利用することは適切ではないと考えている．もちろん，順位から精度を計算することは可能であるが，それらの精度は情報検索と比較してきわめて低く，他の手法と比較することが困難である．そこで，本実験では平均順位向上数を利用して提案手法との評価を行った．

キーパーソンの割合を 10%から 100%まで 10%刻みで設定した場合の順位向上数を**図 8** に示す．このグラフでは，すべての著者に信頼度を算出した場合と比較して，信頼度が高い記事の順位の向上もしくは下落を平均した数である．このグラフから，信頼度算出率を 40%から 70%に設定したときに，本研究で提案された 3 つの方法のどれであっても平均順位が向上したことを示している．また，著者の記述量と記述した記事数の組合せを用いた方法を利用したときに，最も精度が向上した．以上の議論より，提案手法の有効性を確認することができた．

本実験で算出された記事の信頼度が高いにもかかわらず，実際に信頼度が高いとはいえなかった記事が存在した．これらの記事は，主にある事象を追記することにより構成されている記事であり，たとえばテレビ番組で行われた言動，行動を記録したページなどである．これらの記事は非常に長いこと，数少ない著者によって記述されていることが特徴である．これらの記事は，著者が編集を行う際に必ずしも以前の版を閲覧しているとは限らず，記事の質を向上させるための編集が行われているとはいえない．今後は，このような記事を排除する新たな手法が必要であると考えられる．

5.4.2　評価実験 2-2：人手による評価

最後に，実際に得られた記事の信頼度が利用者の判断する信頼度とどの程度合致するかを確かめるために，人手により評価を行った．本実験では，提案手法を利用したとき，著者の信頼度算出率を 40%に設定したときに，信頼度の高い記事に対して高い信頼度を，信頼度の低い記事に対して低い信頼度を算出しているかを実際に確かめた．

評価実験の手順を述べる．まず，ランダムに 1,000 件の記事を選択した．このとき，10 回以上の編集が 3 人以上の著者によって行われているものだけを選択した．これは，編集回数が少ない記事や著者数が少ない記事は一般的に信頼度が低いことが多いと考えられるためである．この記事を信頼度が高いもの，低いものの 2 つに分類した．ここで，全記事の信頼度における平均値を求め，その信頼度よりも高いときに信頼度が高いと判定し，低いときには信頼度が低いと判定した．

次に，本論文の第 1 著者によって人手で信頼度が低いかどうかを確かめた．判定の基準として，Wikipedia 以外のインターネット上の情報を利用して確認することができたものを信頼度が高いとし，確認できなかったものを信頼度が低いとした．また，記事に含まれる複数の項目のうち，信頼度の高いものと低いものがどちらも含まれている場合には，割合の多い項目をその記事の信頼度とした．つまり，記事に信頼度の高い記述が多く，低い記述が少ない場合には，その記事の信頼度を高いと判定した．

そして，人手による信頼度とシステムによる信頼度との一致度を調査した．実験結果を**表 1** に示す．人手によって信頼度が高いと判定した記事数は 915 件であり，信頼度が低いと判定した記事数は 85 件であった．また，提案システムによって信頼度が高いと判定した記事数は 734 件であり，信頼度が低いと判定した記事数は 266 件であった．

この結果から，81.3%の記事で人手による判定とシステムによる判定が一致することが分かった．そのため，提案手法を利用したときに計算される精度は十分に高いと判定すること

表 1　信頼度の精度に関する人手による評価
Table 1　Manual credibillity value assessment.

		システムによる判定		
		信頼度が高いと判断	信頼度が低いと判断	合計
人手による判定	信頼度が高い	731	184	915
	信頼度が低い	3	82	85
	合計	734	266	1,000

ができる．ところが，人手による判定により信頼度が高いと判定されているにもかかわらず，システムにより信頼度が低いと判定された場合が 18.4%存在した．この原因として，信頼度が高いかどうかを判定する閾値が人手による閾値と異なっていた点があげられる．利用者によって，記事に対する信頼度は異なると考えられる．また，利用者による直感的な信頼度は人により異なる点も原因でる．本論文では信頼度を表す 1 つの指標を示したが，さらに多角的な信頼度の分析手法を今後開発しなければならないと考えられる．

6. おわりに

Wikipedia は現在 Web 上で最も成功した，集合知による百科事典の 1 つである．Wikipedia に記述された情報量は増加しているが，情報の質は情報量に比例して高まっているとはいえず，低下する傾向にある．ところが，Wikipedia の閲覧者は Wikipedia に掲載されている情報が信頼できるかどうかを判断することが困難であることが多い．また，記事の閲覧者と比較して飛躍的に記事数が増加しているため，1 つの記事を記述する著者の数は相対的に低下し，間違った情報が修正されない記事数も増加していると考えられる．本研究では，キーパーソンに対して信頼度を計算することによって，計算量を削減しつつ記事の信頼度を算出する方法についての提案を行った．提案手法を利用することによって，利用者は高速で簡単に記事の信頼度を閲覧することができるため，どの記事の信頼度を高めることが良いかを判定することが容易となる．

評価実験において，我々は約 40%の著者に対して信頼度を算出する必要があることが分かった．これは，約 40%の計算量で記事の信頼度を算出することが可能であることが分かり，さらに精度として秀逸な記事，良質な記事の順位が平均して 5 程度向上することが分かった．

本提案で利用されている信頼度とは，利用者の興味と直交する概念であると考えている．情報検索分野では，利用者の興味に適合する検索対象を高速に算出する方法について研究がなされてきた．一方，信頼度が高いからといって利用者の興味に適合するとは必ずしもいえない．我々は，利用者が必要な情報とは利用者の興味に適合することだけではなく信頼度が高いことも含まれると考えている．そのため，たとえば文献 15) に示されているように，もし我々の提案手法を検索システムに適用することによって利用者の検索システムに対する満足度を高めることができると考えている．

最後に，今後の課題について述べる．

- **一般の Web ページにおける提案手法の利用**
  我々の提案システムでは，記事の編集履歴を利用して信頼度の算出を行った．ところが，一般の Web ページでは，編集履歴が一般に公開されていることはきわめてまれで，保存されていないことも多い．さらに Web ページの数は Wikipedia の記事数と比較してもきわめて多い．そこで，我々は編集履歴以外の情報を利用した信頼度算出手法を提案しなければならない．
- **文章解析の利用**
  提案手法では，我々は文書に記述されている単語の内容の解析を行っていなかった．この利点として，どのような言語で記述されている文書にも適用することが可能であることがいえる．ところが評価実験において，丁寧な言葉で記述されている文書は信頼度が高くなりやすいという傾向を得ることができた．文献 16) では，信頼度を算出するうえで文書解析を行うことが有効であることを示している．そこで，これら文書解析による手法を我々の提案している編集履歴による手法と組み合わせることによって，より精度の高い信頼度算出システムを構築することが可能となると考えている．
- **ユーザインタフェースと可視化**
  5.2.2 項において述べたように，我々は Wikipedia の Web インタフェースを利用したユーザインタフェースを構築した．ところが，信頼できない記事に対して恐れている利用者と信頼できる記事だけを閲覧したい利用者では異なるインタフェースを利用するほうが望ましいと考えている．文献 17)，18) では，より利用者に閲覧しやすいインタフェースが利用されている．そこで，さらに利用者にとって利用しやすいインタフェースを構築することを考えている．
- **リンク構造の解析**
  提案手法では編集履歴だけを利用し，内部リンク，外部リンクを利用しなかった．リンク構造は信頼度を測定するための手段として重要であると考えている．現在までに文献 19)–21) などでリンク構造の解析が行われているため，これらの手法を利用した新

たな信頼度を測定する必要があると考えている．

- **複数の言語間における記事の利用**

  Wikipedia には，異なる言語で同じ内容の記事が存在している．Ader ら[22] は，この Wikipedia の特徴に基づいて記事の信頼度に関する研究を行っている．この特徴を利用することによって，新たな信頼度提示手法を提案することが可能となると考えられる．

参 考 文 献

1) Kittur, A., Suh, B. and Chi, E.H.: Can you ever trust a wiki?: Impacting perceived trustworthiness in wikipedia, *CSCW '08: Proc. ACM 2008 Conference on Computer Supported Cooperative Work*, New York, NY, USA, pp.477–480, ACM (2008).
2) Giles, J.: Special report: Internet encyclopedias go head to head, *Nature*, Vol.438, No.15, pp.900–901 (2005).
3) Ortega, F. and Gonzalez-Barahona, J.M.: Quantitative analysis of the Wikipedia community of users, *WikiSym '07: Proc. 2007 International Symposium on Wikis*, New York, NY, USA, pp.75–86, ACM (2007).
4) Siegler, M.: YouTube Comes To A 5-Star Realization: Its Ratings Are Useless (2009). http://www.techcrunch.com/2009/09/22/youtube-comes-to-a-5-star-realization-its-ratings-are-useless/
5) Adler, B.T. and de Alfaro, L.: A content-driven reputation system for the wikipedia, *WWW '07: Proc. 16th International Conference on World Wide Web*, New York, NY, USA, pp.261–270, ACM (2007).
6) Adler, B.T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I. and Raman, V.: Assigning Trust to Wikipedia Content, *WikiSym '08: Proc. International Symposium on Wikis*, ACM (2008).
7) Adler, B.T., Adler, B.T., Pye, I. and Raman, V.: Measuting Author Contributions to the Wikipedia, *WikiSym '08: Proc. International Symposium on Wikis* (2008).
8) Hu, M., Lim, E.-P., Sun, A., Lauw, H.W. and Vuong, B.-Q.: Measuring article quality in wikipedia: Models and evaluation, *CIKM*, Silva, M.J., Laender, A.H.F., Baeza-Yates, R.A., McGuinness, D.L., Olstad, B., Olsen, Ø.H. and Falcão, A.O. (Eds.), pp.243–252, ACM (2007).
9) Stvilia, B., Twidale, M., Smith, L. and Gasser, L.: Information quality work organization in wikipedia, *J. Am. Soc. Inf. Sci. Technol.*, Vol.59, No.6, pp.983–1001 (2008).
10) Kramer, M., Gregorowicz, A. and Iyer, B.: Wiki Trust Metrics based on Phrasal Analysis, *WikiSym '08: Proc. International Symposium on Wikis*, ACM (2008).
11) Mizzaro, S.: Quality Control in Scholarly Publishing, *J. Am. Soc. Inf. Sci. Technol.*, Vol.54, pp.989–1005 (2003).
12) Cusinato, A., Della Mea, V., Di Salvatore, F. and Mizzaro, S.: QuWi: Quality control in Wikipedia, *WICOW '09: Proc. 3rd Workshop on Information Credibility on the Web*, New York, NY, USA, pp.27–34, ACM (2009).
13) Wöhner, T. and Peters, R.: Assessing the quality of Wikipedia articles with lifecycle based metrics, *WikiSym '09: Proc. 5th International Symposium on Wikis and Open Collaboration*, New York, NY, USA, pp.1–10, ACM (2009).
14) Wilkinson, D.M. and Huberman, B.A.: Cooperation and quality in wikipedia, *WikiSym '07: Proc. 2007 international symposium on Wikis*, New York, NY, USA, pp.157–164, ACM (2007).
15) Toms, E.G., Mackenzie, T., Jordan, C. and Hall, S.: Wikisearch: Enabling interactivity in search, *Proc. International Conference on Research and Development in Information Retrieval* (*SIGIR2009*), p.843.
16) Sabel, M.: Structuring wiki revision history, *WikiSym '07: Proc. 2007 International Symposium on Wikis*, New York, NY, USA, pp.125–130, ACM (2007).
17) Holloway, T., Bozicevic, M. and Börner, K.: Analyzing and visualizing the semantic coverage of Wikipedia and its authors, *Complexity*, Vol.12, No.3, pp.30–40 (2007).
18) Otjacques, B., Cornil, M. and Feltz, F.: Visualizing Cooperative Activities with Ellimaps: The Case of Wikipedia, *CDVE*, Luo, Y. (Ed.), Lecture Notes in Computer Science, Vol.5738, pp.44–51, Springer (2009).
19) Brandes, U., Kenis, P., Lerner, J. and van Raaij, D.: Network analysis of collaboration structure in Wikipedia, *Proc. International Conference on World Wide Web* (*WWW2009*), pp.731–740.
20) Lizorkin, D., Medelyan, O. and Grineva, M.P.: Analysis of community structure in Wikipedia, *Proc. International Conference on World Wide Web* (*WWW2009*), pp.1221–1222.
21) Huang, D.W.C., Trotman, A. and Geva, S.: The importance of manual assessment in link discovery, *Proc. International Conference on Research and Development in Information Retrieval* (*SIGIR2009*), pp.698–699.
22) Adar, E., Skinner, M. and Weld, D.S.: Information arbitrage across multilingual Wikipedia, *Proc. International Conference on Web Search and Data Mining* (*WSDM2009*), pp.94–103 (2009).

(平成 22 年 3 月 18 日受付)
(平成 22 年 7 月 11 日採録)

(担当編集委員　奥村 学)

鈴木　優（正会員）

奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．立命館大学，京都大学を経て 2010 年より名古屋大学情報基盤センター研究員．マルチメディア情報検索，情報の質の計測に関する研究に従事．電子情報通信学会，ACM，IEEE Computer 各会員．

吉川　正俊（正会員）

京都大学大学院工学研究科博士後期課程修了．工学博士．京都産業大学，奈良先端科学技術大学院大学，名古屋大学を経て 2006 年より京都大学大学院情報学研究科教授．この間，南カリフォルニア大学客員研究員，ウォータルー大学客員准教授．XML 情報検索，異種情報源の統合，マルチメディアデータベース等の研究に従事．電子情報通信学会，ACM 各会員．