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

A Fast and Accurate Method of Calculating Wikipedia Article Credibility Degrees using Identification Techniques of Key Persons

鈴木 優* 吉川 正俊
Yu Suzuki Masatoshi Yoshikawa

京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

Abstract:
本研究では，Wikipedia において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特定し，それら重要な著者の情報だけを利用して記事の信頼度を算出することによって，全ての著者を利用して信頼度を算出する方法よりも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．

1 はじめに

Wikipedia [*1] は，現在最も成功している百科事典の一つであり，Wiki により作成されている．Wikipedia の記述量は年々増加傾向にある [1]．Wikipedia の一つの特徴として，誰もがページの編集を行うことができる点が挙げられる．ところが，これらの記述は常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事を追加，削除した場合でも Wikipedia はそのままページに反映する．そのため，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るわけではない [2]．

Wikipedia の信頼性の問題点を解決するために，現在 Wikipedia では多くの編集者によって信頼性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 に含まれるページの記述量が多くなるに従って，全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，ページの記述量と質は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，ページ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．

ページの信頼度は， 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．Wikipedia の利用者は一般に，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 が信頼できるかどうかを判断することが困難であり，信頼できない記述を信頼してしまう可能性は極めて高い．

この問題を解決するために，集合知を利用した情報評価手法が数多くの場面で利用されてきた．例えば 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五段階で評価するシステムを導入しており，Amazon [*3] でも類似した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全て良質な動

* 連絡先：京都大学大学院情報学研究科
〒606-8501 京都府京都市左京区吉田本町
E-mail: {ysuzuki, yoshikawa}@i.kyoto-u.ac.jp
[*1] http://ja.wikipedia.org/
[*2] http://www.youtube.com
[*3] http://www.amazon.co.jp/

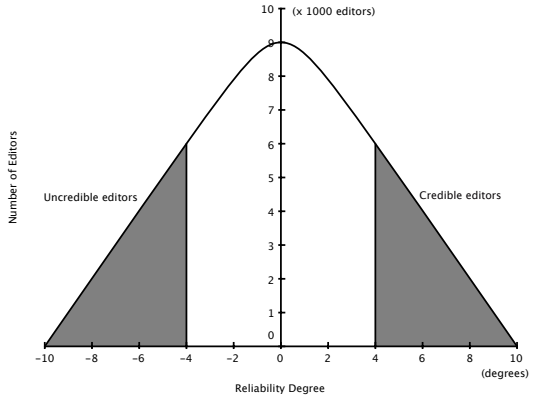

図 1　信頼度と著者数との関係

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といえる．

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の提案を行う．Adler ら [5, 6, 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考える．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．

この手法の問題点として，全ての著者が同等に扱われている点が挙げられる．著者の Wikipedia 全体への貢献度合いはそれぞれ異なっており，貢献度合いの高い著者を優先的に記事の信頼度へ反映させるべきであると考えている．これは，貢献度の小さい著者は記事に対して与える影響が小さく，貢献度の大きな著者は記事に対して与える影響が大きいと考えたためである．

もう一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，一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．そこで我々は，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．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，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．本稿では以後，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．

本稿では，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．ここでキーパーソンとは信頼度が比較的高い著者と比較的低い著者であると考えている．図 1 において，信頼度と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．ここで，グラフのうち信頼度の低い部分，高い部分が灰色で塗られているが，この部分がキーパーソンの含まれる領域である．つまり，全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．そのため，本稿で提案する，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，高速に著者の信頼度を計算することができる．

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 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，ほぼ全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．

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，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，キーパーソンに対してだけ著者の信頼度を算出する．最後に全ての記事に対して，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 (2) 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二つの要件があり，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が必要である．そこで本研究では著者の記述量，著者が記述した記事数，それら二つの値の組み合わせである三つの特徴量算出手法を提案した．算出に必要な計算コスト，算出された記事の信頼度に関する精度の二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]．例えば，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

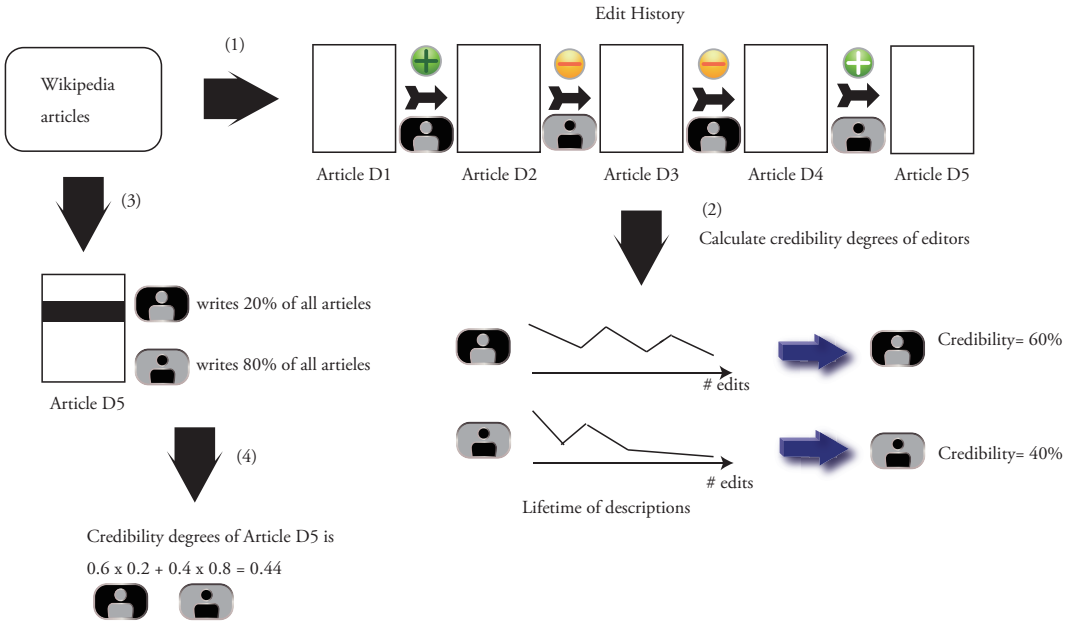

図 2　提案手法の概要

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には主に三つが考えられる．一つは利用者の投票を利用した方法，一つは自然言語処理による方法，そして編集履歴を利用する方法である．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 [*4] 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，全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二つである．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無いため，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 Stefano 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一つであると考え，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時には，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．

Adler ら [5, 6, 7] や Hu ら [8]， Wilkinson ら [13] 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．Adler らはこの手法を MediaWiki のプラグインと FireFox のプラグインとして実装し，WikiTrust Module [14] [*5] として公開している．これらのシステムでは，全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

本章では，まず信頼度算出手法について述べる．この手法は 2 章で述べた Adler らの手法に基づいている．この手法は，記述の残存率に基づいて計算される方法である．つまり，ある記述が記事に追加されたときに，その記事が多くの編集の上で残存している場合，その記述は信頼度が高いと判定する手法である．なぜならば，著者

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

[*5] http://en.wikipedia.org/wiki/WikiTrust

は一般的に，記事のうち問題であると考えている部分を削除し，問題の無い部分は削除しないと考えられるためである．

この手法では，記事の信頼度を算出するために，その記事を記述した著者の信頼度を利用する．そのため，まず著者の信頼度を求める必要がある．

3.1 節において提案手法の概要について述べる．次に，3.2 節で著者の信頼度算出手法について述べ，最後に 3.3 節で記事の信頼度算出手法について述べる．

3.1 提案手法の概要

この手法では，次の三つの手順により記事の信頼度を算出する．

1. **編集履歴から特徴量の抽出**
   編集履歴から特徴量を抽出する．ここで特徴量として，記事のタイトル，著者，記述の三つを利用した．
2. **著者の信頼度算出**
   編集履歴を解析することによって，著者の信頼度の計算を行う．
3. **記事の信頼度算出**
   記事を記述した著者の信頼度を利用して，記事の信頼度を算出する．

まず，変数の定義を行う．記事 D には一つまたは複数の履歴が存在し，それぞれ D_i $(i = 1, 2, \cdots, N)$ とする．D_i は i 回目に編集された後の記事の記述である．次に，編集履歴 D_i と D_{i+1} が与えられた時にの著者の信頼度を算出する方法を述べる．

3.2 著者の信頼度算出

まず，二つの編集における追加分 $A(i-1, i)$ を次のように定義する．

$$A(i-1, i) = F(D_i) - (F(D_{i-1}) \cap F(D_i)) \tag{1}$$

ここで $F(D_i)$ は D_i に含まれる記述であり，$F(D_{i-1}) \cap F(D_i)$ は D_i，D_{i-1} の両方に含まれる記述を示す．

そして，求められた追加分による著者の信頼度を計算する．

$$I_{add}(D_{i-1}, D_i) = \frac{|\, A(i-1, i) \cap F(D_i) \,|}{|\, A(i-1, i) \,|} \tag{2}$$

次に，二つの編集における削除分 $J(i-1, i)$ を次のように定義する．

$$J(i-1, i) = (F(D_{i-1}) \cap F(D_i)) - F(D_i) \tag{3}$$

そして，求められた削除分により信頼度の計算を行う

$$I_{del}(D_{i-1}i, D_i) = 1 - \frac{|\, J(i-1, i) \cap F(D_i) \,|}{|\, J(i-1, i) \,|} \tag{4}$$

追加と削除によって計算された信頼度を統合する．一つの編集には複数の追加および削除が含まれている．そのため，それぞれの方法で算出された信頼度を統合し，一つの編集における著者の信頼度を算出する．信頼度の変化量 $E(D_i, D_{i+1})$ は次のように計算する．

$$E(D_i, D_{i+1}) = \sum_{i=1}^{K-1} I_{add}(D_i, D_{i+1}) + \sum_{i=1}^{K-1} I_{del}(D_i, D_{i+1}) \tag{5}$$

最後に，著者 u_i の信頼度 $P(u_i)$ を次の方法によって求める．

$$P(u_i) = \frac{\sum_{i=1}^{K} \sum_{j=i+1}^{K} E(D_i, D_j)}{K} \tag{6}$$

ここで K は記事 D の編集回数である．

3.3 記事の信頼度算出

最後に，記事の信頼度 $T(D)$ を求める．記事の信頼度は，著者の信頼度を平均したものであり，次の式で求められる．

$$T(D) = \frac{\sum_{i=1}^{M} r(u_i) \cdot P(u_i)}{M} \tag{7}$$

ここで M は記事 D を記述した著者の量であり，$r(u_i)$ は著者 u_i が記述した割合であり，0 と 1 との間の値である．

4 キーパーソンの抽出

3 章では，Wikipedia の記事に対して信頼度を算出する手法について述べた．ところがこの手法では，計算コストが大きいため，膨大な編集履歴群に対して高速に，的確に記事の信頼度を算出することができない．そこで本章では，キーパーソンを算出する方法についての提案を行う．

まず，Wikipedia の編集履歴の解析を行う．日本語版 Wikipedia [*6] の 2009 年 9 月 20 日版の編集履歴[*7] を利用して，著者と記事数の相関関係を調査した．図 3 に結果を示す．x 軸は記事数の多い順に並べた著者を示し，y 軸は記事数を示している．この結果から，一部の著者が大部分の記事を記述していること，大部分の著者は少数の記事だけを編集していることが分かる．

次に，3 章で述べた方法によって，無作為に抽出した 5,000 人の著者に対して信頼度を算出し，信頼度と著者との相関関係を調査した．図 4 に実験結果を示す．x 軸は信頼度の低い順に著者を並べたときの著者であり，y

[*6] http://ja.wikipedia.org/
[*7] http://download.wikipedia.org/

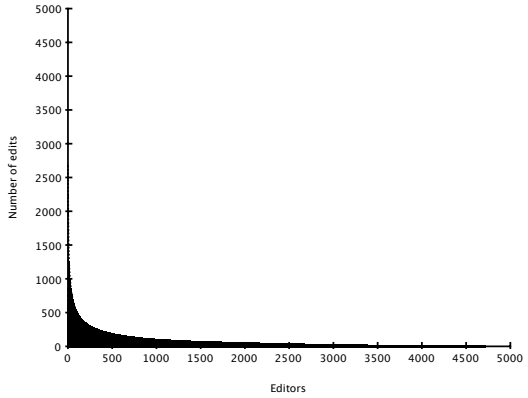

図 3　著者と記述数との相関関係

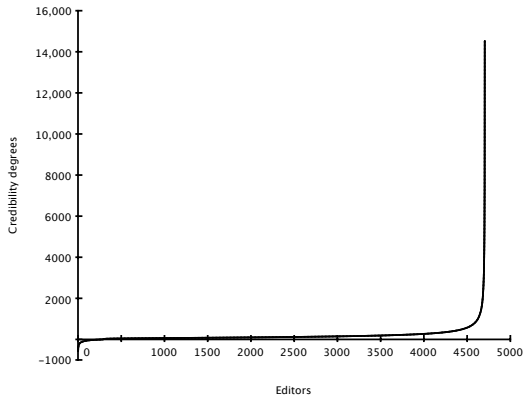

図 4　著者とその信頼度との相関関係

軸には著者の信頼度を表している．この図からも分かるように，90% の著者は信頼度がほぼ 0 となるため，記事の信頼度に影響を与えないことが分かる．そのため，これらの著者に対して信頼度を算出する前にキーパーソンを特定することが可能となれば，記事の信頼度に対する精度に影響を与えることなく計算コストを削減することが可能となると考えられる．

ここで，信頼度を算出することなく著者を信頼度順に並べ替える必要がある．そのため，信頼度と相関関係が高く，信頼度の算出よりも簡単な方法で算出可能な値が必要である．本論文ではこのような値を算出する方法として三つの方法を提案し，それぞれ著者の記事記述量と記事数，およびその組合せを提案する．

この章ではまず，提案する三つの手法によるキーパーソン抽出手法についてそれぞれ述べる．次に，これらの数値が実際に利用できるかどうかを確かめるために，予備実験として実際に算出した信頼度による著者順と，三つの提案手法による著者順をスピアマンの順位相関係数によって確かめた．

4.1　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 u_i が記事 a_j に対して $f(u_i, a_j)$ バイトの記述を行ったとき，その著者 u_i の重要度 $l_1(u_i)$ を次の式で計算する．

$$I_1(u_i) = \sum_{j=0}^{N} f(u_i, a_j) \tag{8}$$

4.2　記述した記事の数によるキーパーソンの特定

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

$$P(u_i, a_j) = \begin{cases} 1 & u_i \text{ が } a_j \text{ に一度以上記述したとき} \\ 0 & \text{それ以外} \end{cases} \tag{9}$$

この変数 $P(u_i, a_j)$ を利用して，著者の重要度 $l_2(u_i)$ を算出する．

$$I_2(u_i) = \sum_{j=0}^{N} P(u_i, a_j) \tag{10}$$

4.3　記述量と記述した記事の数の組合せによるキーパーソンの特定

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 u_i の重要度 $l_3(u_i)$ は次の方法で算出される．

$$I_3(u_i) = \sum_{j=0}^{N} f(u_i, a_j) \cdot - \log \frac{\sum_{j=0}^{N} P(u_i, a_j)}{N} \tag{11}$$

4.4 キーパーソンの抽出

4.1 節，4.2 節，および 4.3 節で述べられている方法によって，著者の順位を算出する．次に，著者の順位を決定した上で上位 k% の著者を特定し，それらの著者をキーパーソンであるとする．

4.5 予備実験

予備実験では，信頼度とこれら三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時には，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全ての著者の信頼度を算出し，信頼度の高い著者から順に著者 ID を算出する
2. 全ての著者に対して，三つの指標によって順位を算出し，それぞれの順に著者 ID を算出する
3. Phase 1. において算出された信頼度と Phase 2. において算出されたそれぞれの順位との相関関係を，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 2009 年 9 月 20 日の Wikipedia data in Japanese[*8] 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 Phase 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 5 に示す．

この図から，三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．

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

[*8] http://download.wikipedia.org/jawiki/20090920/pages-meta-history.xml.bz2

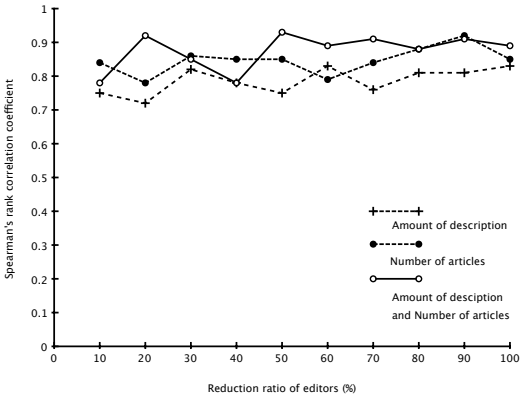

図 5 三つの方法によって算出された著者の順位と Adler らの方法による著者の順位とのスピアマン順位相関係数

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
   我々は次の三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，三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 節で述べた．
   - 著者の記述した記事数による著者順位表．4.2 節で述べた．
   - 著者の記述量と記事数の組合せによる著者順位表．4.3 節で述べた．
3. 求められた三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．

5.1.1 実装

図 6 に示すように，我々は二つの計算機を利用して提案手法の実装を行った．一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．この計算機は二つの Intel Xeon 3.06GHz プロセッサ，4GByte

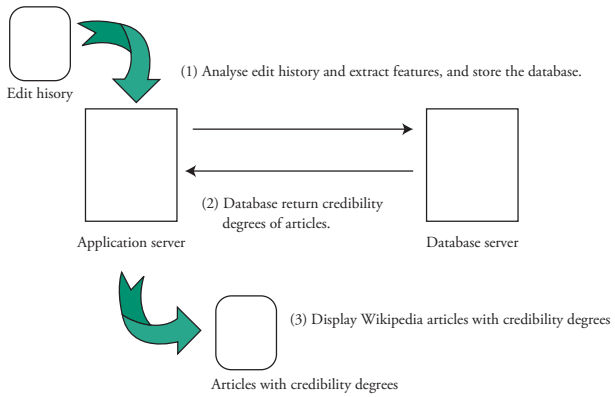

図 6　提案手法の実装

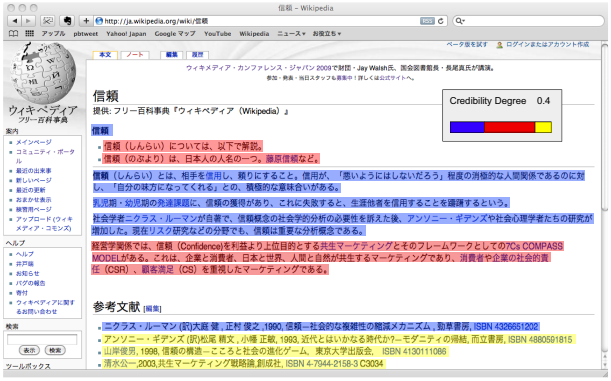

図 7　提案手法の実装イメージ

メモリ，500GByte 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 10.6.1 (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．もう一つの計算機は，編集履歴データから特徴量を抽出するために利用した．この計算機は一つの Intel Core i7 プロセッサ，16GByte のメモリ，80GByte 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 を利用して実装がなされている．二つの文字列を比較するために我々は Diff, Match, Patch ライブラリ[*9] を利用している．

5.1.2　評価実験に利用したデータ群

我々は日本語版 Wikipedia の編集履歴データを利用した．このデータには 627,110 件の記事，29,264,823 件の編集，および 358,561 人の登録利用者が含まれている．この登録利用者のうち 11,332 人は少なくとも 1 ヶ月間に 1 回以上の編集を行った著者である．ところがこの編集履歴データはあまりにも大きく，全てのデータを処理するために膨大な時間がかかる．そこで，我々は 85,028 件の記事 (全体の約 13.6%) で，二回以上の編集が行われた記事だけを対象とした．著者数は 705,713 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．

我々の提案手法は，日本語版以外の Wikipedia に適用することは可能である．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ている．

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 [*10] を利用した．ところが，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 Adler らのシステムを利用した．このシステムは，全ての著者に対して信頼度を計算するシステムである．

5.2　ユーザインタフェース

図 7 に提案手法の開発中画面を示す．このシステムは JSP (Java Server Pages) で実装されている．利用者は Wikipedia の記事内容と共に，その記事の信頼度を表示している．記述部分は赤色，青色，黄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．

[*9] http://code.google.com/p/google-diff-match-patch/

[*10] http://ja.wikipedia.org/w/index.php?title=特別:登録利用者一覧&group=bot

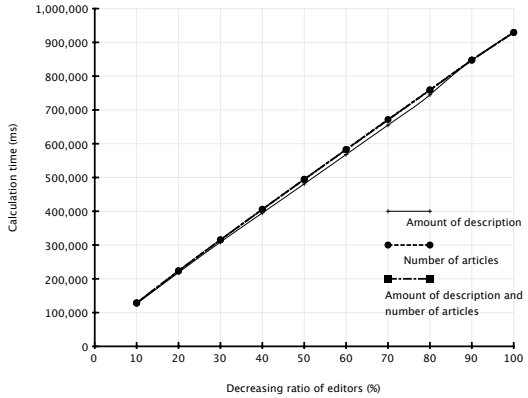

図 8　著者の削減率と計算時間との関係

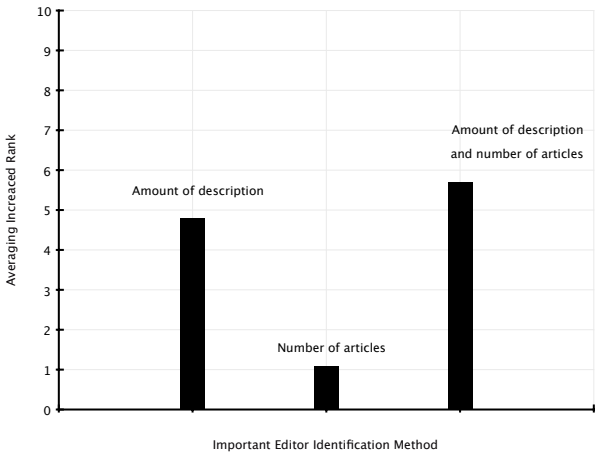

図 9　三つの手法による平均順位向上数

5.3　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無い．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 8 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 16 時間程度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．

次に，図 8 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．

この図から，信頼度を計算する著者の割合と計算量は比例することが分かる．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．つまり，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　信頼度の精度

最後に，信頼度に関する精度を算出し，提案手法の有効性を確認するための評価実験を行った．

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．

ここで，信頼度が高い記事として，日本語版 Wikipeidia 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．本実験において利用した編集履歴には 98 件の記事が含まれていたため，これらの記事を信頼度が高い記事であるとし，利用した．

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極めて低く，他の手法と比較することが困難である．

キーパーソンの割合を 40% に設定した場合の順位向上数を図 9 に示す．このグラフでは，全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，本研究で提案された三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，例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．

6　おわりに

Wikipedia は Web 上で最も成功した，集合知による百科事典の一つである．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，一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，例えば文献 [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，文書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，文書の編集履歴は一般に公開されていることは極めてまれで，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極めて多い．そこで，我々は編集履歴以外の情報を利用した，スケーラブルな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言える．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，信頼度を算出する上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**
  提案手法では，5.2 節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, 20, 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．Allan 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．

謝辞

本研究の一部は，文部科学省科学研究費補助金（課題番号 20700101, 20300036, 20500104），特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」（課題番号 21013026）および NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」によるものです．ここに記して謝意を表します．

参考文献

[1] Jim Giles. Special report: Internet encyclopedias go head to head. *Nature*, 438(15):900–901, 2005.

[2] Aniket Kittur, Bongwon Suh, and Ed H. Chi. Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia. In *CSCW '08: Proceedings of the ACM 2008 conference on Computer supported cooperative work*, pages 477–480, New York, NY,

USA, 2008. ACM.

[3] Felipe Ortega and Jesus M. Gonzalez-Barahona. Quantitative analysis of the wikipedia community of users. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 75–86, New York, NY, USA, 2007. ACM.

[4] MG Siegler. Youtube comes to a 5-star realization: Its ratings are useless: http://www.techcrunch.com/2009/09/22/youtube-comes-to-a-5-star-realization-its-ratings-are-useless/, September 2009.

[5] B. Thomas Adler and Luca de Alfaro. A content-driven reputation system for the wikipedia. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 261–270, New York, NY, USA, 2007. ACM.

[6] B. Thomas Adler, Krishnendu Chatterjee, Luca de Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman. Assigning trust to wikipedia content. In *WikiSym '08: Proceedings of International Symposium on Wikis*. ACM, 2008.

[7] B. Thomas Adler, B. Thomas Adler, Ian Pye, and Vishwanath Raman. Measuting author contributions to the wikipedia. In *WikiSym '08: Proceedings of International Symposium on Wikis*, 2008.

[8] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. Measuring article quality in wikipedia: models and evaluation. In Mário J. Silva, Alberto H. F. Laender, Ricardo A. Baeza-Yates, Deborah L. McGuinness, Bjørn Olstad, Øystein Haug Olsen, and André O. Falcão, editors, *CIKM*, pages 243–252. ACM, 2007.

[9] Besiki Stvilia, Michael Twidale, Linda Smith, and Les Gasser. Information quality work organization in wikipedia. *J. Am. Soc. Inf. Sci. Technol.*, 59(6):983–1001, 2008.

[10] Mark Kramer, Andy Gregorowicz, and Bala Iyer. Wiki trust metrics based on phrasal analysis. In *WikiSym '08: Proceedings of International Symposium on Wikis*. ACM, 2008.

[11] Stefano Mizzaro. Quality control in scholarly publishing:. *J. Am. Soc. Inf. Sci. Technol.*, 54:989–1005, 2003.

[12] Alberto Cusinato, Vincenzo Della Mea, Francesco Di Salvatore, and Stefano Mizzaro. Quwi: quality control in wikipedia. In *WICOW '09: Proceedings of the 3rd workshop on Information credibility on the web*, pages 27–34, New York, NY, USA, 2009. ACM.

[13] Dennis M. Wilkinson and Bernardo A. Huberman. Cooperation and quality in wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 157–164, New York, NY, USA, 2007. ACM.

[14] Krishnendu Chatterjee, Luca de Alfaro, and Ian Pye. Robust content-driven reputation. In Dirk Balfanz and Jessica Staddon, editors, *AISec*, pages 33–42. ACM, 2008.

[15] Elaine G. Toms, Tayze Mackenzie, Chris Jordan, and Sam Hall. wikisearch: enabling interactivity in search. In Allan et al. [23], page 843.

[16] Mikalai Sabel. Structuring wiki revision history. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 125–130, New York, NY, USA, 2007. ACM.

[17] Todd Holloway, Miran Bozicevic, and Katy Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors. *Complexity*, 12(3):30–40, 2007.

[18] Benoît Otjacques, Maël Cornil, and Fernand Feltz. Visualizing cooperative activities with ellimaps: The case of wikipedia. In Yuhua Luo, editor, *CDVE*, volume 5738 of *Lecture Notes in Computer Science*, pages 44–51. Springer, 2009.

[19] Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise van Raaij. Network analysis of collaboration structure in wikipedia. In Quemada et al. [24], pages 731–740.

[20] Dmitry Lizorkin, Olena Medelyan, and Maria P. Grineva. Analysis of community structure in wikipedia. In Quemada et al. [24], pages 1221–1222.

[21] Darren Wei Che Huang, Andrew Trotman, and Shlomo Geva. The importance of manual assessment in link discovery. In Allan et al. [23], pages 698–699.

[22] Eytan Adar, Michael Skinner, and Daniel S. Weld. Information arbitrage across multi-lingual wikipedia. In Ricardo A. Baeza-Yates, Paolo Boldi, Berthier A. Ribeiro-Neto, and Berkant Barla Cambazoglu, editors, *WSDM*, pages 94–103. ACM, 2009.

[23] James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors. *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*. ACM, 2009.

[24] Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl, editors. *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*. ACM, 2009.