CredibilityRank: 編集履歴と著者情報を用いたWikipedia の記事信頼度算出手法

鈴木　優† 吉川　正俊††

† 名古屋大学情報基盤センター　〒 450–0002 愛知県名古屋市千種区不老町
†† 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: †suzuki@db.itc.nagoya-u.ac.jp, ††yoshikawa@i.kyoto-u.ac.jp

あらまし　本研究では Wikipedia の記事に対して，編集履歴と著者情報を利用して信頼度を高精度に算出するための手法を提案する．我々は現在までに，情報の編集履歴を用いることによって情報の信頼度を算出する方法を提案していた．この手法では，記事に書き加えられた記述が多くの編集を経て残存している場合，その記述の信頼度が高いと仮定していた．つまり，ある記述が他の著者によって削除された場合，その記述の信頼度が低くなる．ところが，もし記述を削除した著者の信頼度が低い場合，その著者は適切な記述を削除する可能性が比較的高くなると考えられる．そのため，本研究では，記述を削除した著者の信頼度を考慮した記述の信頼度算出手法を提案する．評価実験により，算出された信頼度が利用者の直感に近いことが分かった．
キーワード　信頼度，Wikipedia，編集履歴

CredibilityRank: Credibility Values of Wikipedia Articles using Edit History and Editor Information

Yu SUZUKI† and Masatoshi YOSHIKAWA††

† Information Technology Center, Nagoya University　Furo, Chikusa, Nagoya, Aichi 450–0002, Japan
†† Graduate School of Informatice, Kyoto University　Yoshida-Honmachi, Sakyo, Kyoto 606–8501, Japan
E-mail: †suzuki@db.itc.nagoya-u.ac.jp, ††yoshikawa@i.kyoto-u.ac.jp

Abstract　In this study, we propose a method to calculate credibility degrees to Wikipedia articles, using edit history and editor information. In our previous work, we proposed a method for calculating credibility values to Wikipedia articles using edit history. In that study, we calculate credibility values using only remain ratio of description. However, if low credible edutors deletes descriptions, even if the deletions are not appropriate, these remain ratio decrease, and the credibility degrees of the descriptions decrease. In this paper, we improve the credibility degree calculation ratio using not only remain ratio but also editors' credibility degrees. Using our proposed method, we can improve the accuracy of credibility values for editors and descriptions.
Key words　Credibility, Wikipedia, Edit History

1. はじめに

Web2.0 の発達により，インターネット上の多くの利用者が互いに対話することによって，多くのコンテンツが生成されるようになった．これらのコンテンツは UGC (User Generated Contents) と呼ばれ，例えば Wikipedia や blog，SNS などが挙げられる．UGC の利点は，多くの利用者による意見が結集しているため，一般の書籍や辞書と比較して情報の量が多いこと，新しい用語や事象に関する情報が即座に提供されることが挙げられる．UGC を生成している利用者には，それぞれ専門的な知識を持つ場合があるため，少数の利用者が生成するコンテンツよりも専門的な知識を入力することができる場合が多い．これらの利点から，情報を得るための手段としての UGC はますます重要なものとなりつつある．

Wikipedia などの UGC は，誰でも記事を編集することが可能であるため，収録されている情報の網羅度が高いという長所を持つ一方で，信頼度の高い情報だけではなく，信頼度の低い情報が含まれているという欠点も存在する．そのため，この問題を解消するために，UGC 上の情報に対して信頼度を付与する手法が提案されている．ところが，UGC に含まれる情報量

は非常に多い．たとえば日本語版 Wikipedia には 約 180 万件の記事が収録されており，この記事数は日々増加しているため，それらの記事をすべて閲覧することは困難である．一般的に，情報の信頼性を検証することは人手であっても極めて困難である．そのため，自動的に情報の信頼度を算出する手法は重要となっている．

本研究では，情報の信頼度を算出するための方法として，記事自体の残存率に着目した．この方法では，ある記事が投稿されたとき，その記事が他の著者によって削除されなければ，その記事の信頼度が高いと判定する方法である．投稿された記事を編集する著者は，編集前の記事を読み，不要であると判断した部分を削除する．もし信頼度の高い記事が投稿されたとき，その記事は多くの著者が消去すべきでないと判定すると考えられる．その結果，信頼度の高い記事は他の著者の編集により残存する可能性が高い．我々はこの性質に着目し，以前の研究 [1] において記事の信頼度算出を行った．

ところが，この方法では悪意のある利用者によって意図的に信頼度を操作できてしまうという問題がある．たとえば，もしある特定の著者における信頼度を低下させようと考えたとき，その著者が記述した記事をすべて削除する方法がある．この問題は，記事の信頼度を算出する際に記事を削除する利用者自身の信頼度を考慮していないことに起因する．信頼度の高い著者は信頼度の低い記述だけを削除すると考えられることに対して，信頼度の低い著者は信頼度が高い記述も低い記述も両方削除してしまう可能性が比較的高いと仮定する．そこで，記事を削除する著者の信頼度が高いかどうかを記述の信頼度に反映させることができれば，意図的に信頼度を低下してしまう問題を解決することができる．

そこで本研究では，記事を削除する著者の信頼度を考慮した記述の信頼度算出手法の提案を行う．信頼度が低い著者は，他の著者の信頼度算出への影響を小さくし，信頼度の高い著者は，他の著者の信頼度算出への影響を大きくする．この方法によって，より精度が高い信頼度を算出することができる．このとき，記述の信頼度を算出するために著者の信頼度を用いることとなる．ところが，著者の信頼度は記述の信頼度から算出される．つまり，一方の信頼度を算出しなければ他方の信頼度を算出することができず，このままでは著者の信頼度を考慮した記事の信頼度を算出することができない．

この問題を解決するために，本研究ではまず著者の信頼度を定数としてあらかじめ設定しておき，この仮決めされている著者の信頼度と記述の残存率を利用して記述の信頼度を算出する．次に，記述の信頼度を利用して著者の信頼度を算出する．そして，算出された著者の信頼度を利用して記述の信頼度を算出する．このようにして，記述の信頼度と著者の信頼度を交互に繰り返し求めることによって，著者の信頼度を考慮した記述の信頼度を算出する．

評価実験において，提案手法によって記述の信頼度が正しく算出できているかどうかを確かめる実験を行った．また，実際にどのような記述に対してどのような信頼度が算出されているかを確かめることができるように，Web 上のサービスとして提案手法の実装を行った．日本語版 Wikipedia の編集履歴データを利用して評価実験を行った結果，確かに提案手法によって記述の信頼度の算出精度が向上していることを確認した．

2. 関 連 研 究

商品や人物など，ある対象や事象に対して信頼度や質を測定することを目的とした研究は数多く行われている [2] が，これらの研究を明示的な評価による方法，暗示的な評価による方法の二つに分類することができる．明示的な対象の評価とは利用者が明示的に信頼度を示す方法であり，例えば投票のような方法が挙げられる．暗示的な対象の評価とは，利用者が明示的に信頼度を示さない方法であり，利用者の行動や入力などにより信頼度を測定する方法である．以下にそれぞれの方法を概観した上で，提案手法との差異について述べる．

2.1 明示的な評価による信頼度算出

情報の信頼度や質を算出するために，現在最も実用的に利用されている方法は，利用者の評価を利用する方法である．例えば Amazon.com (注1) では，商品の購入者が商品に対して 5 段階の評価を付与することによって，その商品の信頼度，質を客観的に評価している．この手法は，利用者にとって非常に明快な方法であり，簡易な方法で実装することが可能であることから，多くのシステムで利用されている．

この方法では，利用者が明示的に信頼度や質などをシステムに入力する機能をあらかじめシステムに実装していなければならない．ところが，現在の Wikipedia ではこのような機能を実装しておらず，明示的な評価による信頼度を算出することができない．そこで，このような機能を実装することによって利用者による明示的な評価によって信頼度を算出しようという試みがある．Kramer ら [3] は Wikipedia とは別に MediaWiki (注2) に対して利用者による記事への評価投票システムを付加することによって，明示的な評価を入力する機能を実装した．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における問題の一つに，全ての利用者が的確に記事の質を判定することは困難である点が挙げられる．映像投稿サイトである YouTube おける調査 (注3) において，ほとんど全ての利用者が最も高い評価である 5 つ星を映像に付与していることからも明らかである．その後，YouTube では利用者の映像に対して星による評価を行うことを廃止していることからも，利用者による評価が有用でなかったことが分かる．

一般に，人手によって信頼度が高いかどうかを判定することは極めて困難な作業であり，付与された信頼度は必ずしも精度が高いとは限らない．また，算出された信頼度を少数の利用者によって恣意的に高く，もしくは低く誘導することも極めて容易な方法である．明示的な評価による信頼度算出はその算出方

(注1)：http://www.amazon.com
(注2)：MediaWiki は Wikipedia で利用されている Wiki システムである．http://www.mediawiki.org/
(注3)：http://youtube-global.blogspot.com/2009/09/five-stars-dominate-ratings.html

法の明快さ，透明性という利点がある反面，算出された信頼度の精度に問題があるという欠点がある．

2.2 暗黙的な評価による信頼度算出

次に，利用者が暗黙的に対象に対して評価を行うことによって信頼度を算出する方法について述べる．この手法では，利用者にとって明示的な評価を行うことはせず，それに代わる評価を暗黙的に利用者から得るという方法である．本提案は，この手法を利用している．この手法を用いるときに重要な点は，利用者による明示的な評価に代わる利用者からの評価を，どのような方法で得るかという点である．

Wöhner ら [4] 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ことが問題である．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 [9] 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，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，彼らの考え方は記事の残存率だけに基づく方法であり，新しく作成された記事や記述に対して信頼度を算出することができない．つまり，本研究の 3.3.2 節に述べる方法と類似した方法である．本研究では，著者の信頼度を算出することによって，同一の著者による記述に同一の信頼度を算出するため，新しく作成された記述に対する信頼度を算出することができる．

3. 提案手法

本論文では，Wikipedia の編集履歴から記事の信頼度を算出する方法について述べる．まず，本研究における信頼度算出の考え方を述べ，次に本論文で利用する変数，定数の定義を行い，最後に信頼度算出手法の定式化を行う．

3.1 基本的な考え方

本研究では，記事の信頼度を基準を利用して事象間の関連に対して信頼度を求める．ここで基本となる考え方として，**長い編集を経て残留している記述は信頼度が高い**という仮定を行う．著者が編集を行うとき，その記事に誤った記事や不正確な記事が記述されているとき，その記述を消すことが多い．また，記事を編集する回数が多いということは，それだけ多くの著者によって閲覧されていると考えることができ，それらの著者が削除，編集を行う必要が無いと判断した記述は信頼度が高いと考えることができる．この考え方を利用することによって，記事の部分に対して信頼度を算出することができる．

記述の信頼度だけを利用することによる問題として，あまり編集されない記事に対して信頼度を算出することができないことや，記事にとって最後の編集に対して信頼度を算出することができないという点がある．そこで，このような問題を解決するために，著者の信頼度という概念を導入する．信頼度の高い著者とは，高い信頼度を持つ記事を多く書く著者のことである．平均して信頼度の高い記述を記事に対して行う著者は，ほかの記事に対しても高い信頼度の記述を行う可能性が高いと考えられる．

ここでもう一度，記事の信頼度について考える．長い編集を経て残留している記述は信頼度が高いと仮定したが，この仮定は必ずしも正しいとは限らない．たとえば，悪質な利用者が Wikipedia を編集するとき，信頼度の高い記述を意図的に削除する場合がある．また，最初に定義した基本的な考え方では，記述を削除することによってその記述を行った著者の信頼度を下げることができる．そのため，悪質な利用者によって意図的に信頼度を低下もしくは向上させることが可能である点は問題である．そこで，**記述の削除を行った著者の信頼度によって，記述の信頼度を補正する**ことを考える．つまり，記述の削除を行った著者の信頼度が高いとき，その記述の削除は妥当であると考え，記述の信頼度を低下させる．また，記述の削除を行った著者の信頼度が低いとき，その記述の削除は妥当ではない可能性が高いと考え，記述の信頼度を低下させない．このように記述の残留率だけではなく，その記述を削除した著者の信頼度も考慮することによって，より精度の高い信頼度を算出することができる．

3.2 Wikipedia の記事のモデル化

本論文で利用する変数および定数について，次のように定義する．Wikipedia に含まれる文書集合 $D = \{d_i | i = 1, 2, \cdots N\}$ に含まれている任意の文書 d_i を考える．この文書には M_i 個 $(M_i > 0)$ のバージョン集合 $V_i = \{v_{i,j} | j = 0, 1, \cdots, M_i\}$ がある．ここで，$j = 0$ のとき，$v_{i,0}$ は文書の内容が空白であると定義する．つまり，著者が文書 d_i に対して初めてバージョンを作成したとき，そのバージョンは $v_{i,1}$ として保存される．ここでバージョンとは，j 回目に作成されたテキスト全体のことであるため，j 回目にコンテンツを編集した著者だけではなく 1〜$j-1$ 回目に編集した著者が作成したコンテンツも含まれる．

Wikipedia のコンテンツを作成した著者集合 $E = \{e_k | k = 1, 2, \cdots K\}$ について述べる．任意の著者 e_k は少なくとも 1 回以上のバージョンを作成しており，e_k が作成したバージョン集合を $V(e_k) = \{v_{i,j} | i = 1, 2, \cdots, N, j = 0, 1, \cdots, M_i, v_{i,j} \text{ is edited by } e_k\}$ と定義し，各バージョンの著者を $e(v_{i,j})$ とする．ある著者が同一の文書 d_i に対して連続して 2 回以上のバージョンを作成したとき，その著者が作成した最後のバージョンだけを残し，それ以外のバージョンを削除する．つまり $i = 1, 2, \cdots, N$，$j = 0, 1, \cdots, M_i - 1$ であるとき，常に $e(v_{i,j}) \neq e(v_{i,j+1})$ が成り立つ．

バージョン $v_{i,j}$ に含まれる部分文書について述べる．一つのバージョンには $M_{i,j}$ 個の部分文書 $p^x_{i,j}$ $(x = 1, 2, \cdots, M_{i,j})$ があり，一つの部分文書は同一の著者 $e(p^x_{i,j})$ が記述している．また，部分文書群 $p^x_{i,j}$ について x が小さい順に $M_{i,j}$ 個の部分文書を並べると，バージョン $v_{i,j}$ となる．

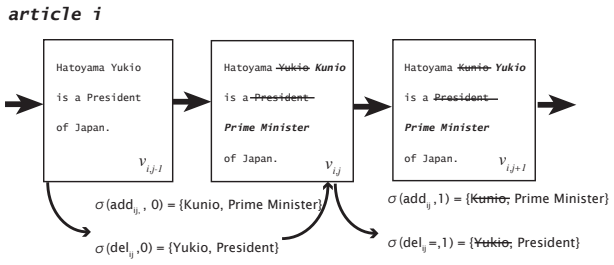

図 1　編集履歴における追加と削除

Fig. 1　An example of added and deleted contents in page edit history.

3.3　信頼度の算出

3.3.1　編集履歴から部分文書の抽出

まず，Wikipedia からダウンロードされる記事の編集履歴から，部分文書とその部分文書を記述した著者を抽出する．

記事の編集履歴には，その記事のタイトル，投稿時間，投稿した著者，その時間での記事の内容などが XML 形式で記述されている．そこで，まず投稿時間と著者，記事内容を抽出する．そして，その記事と一つ前の記事との差分をとり，増加した部分はその記事のバージョンを投稿した著者が記述したと考える．この考え方を利用して，記事に含まれる全てのバージョンを取り出し，バージョン間の差分をとることによって，どの部分をどの著者が記述したかを特定する．

ここで revert と呼ばれる，以前のバージョンへ戻すような編集が記事に対して行われた場合について考える．このとき，差分情報だけをとって記述を行った著者を特定すると，revert を行った著者は，バージョンを戻す前に書かれていた部分を全て書いたと判定されてしまう．ところが，このように判定されてしまうと，もし編集合戦のように revert が繰り返し行われたとき，編集合戦の最中に記述を行った著者の信頼度が低下してしまう．この問題を解決するために，以前のバージョンにおける記述部分の著者をそのまま revert 後の記述部分の著者とし，revert を行った著者は長さ 0 の記述を行ったとする．このような処理を行うことにより，編集合戦など不適当な編集を行うことによって意図的に信頼度を低下させることを防ぐことができる．

3.3.2　記事の変更に対する信頼度 (1)

次に，$v_{i,j}$ 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 $\tau(v_{i,j})$ 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 1 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，$add_{i,j}$ が残存している割合を算出する．ここで，$p = 0$ のときは $\delta(add_{i,j}, 0) = add_{i,j}$ とする．まず，$v_{i,j+p}$ の中から $add_{i,j}$ に相当する部分 $\delta(add_{i,j}, p)$ を抽出する．次に，追加部の残存率である追加残存率 $R(i, j, p)$ を (1) 式によって求める．

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

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

図 1 における例を利用して，$p = 1$ の時の追加残存率 $R(i, j, 1)$ 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．

そして，(2) 式によって追加残存率から記述の信頼度 $\tau(v_{i,j})$ を求める．

$$\tau(v_{i,j}) = \sum_{p=1}^{N_i - j} \log_2 R(i, j, p) \tag{2}$$

ここで，追加残存率に対して log をとっている．これは，非常に長い文章を編集することによって，信頼度に非常に大きな影響を与えることを防ぐためである．また，記述の信頼度を算出するときに，記述を行ったときの残存率，つまり $p = 0$ のときの記述量を信頼度に反映しない．

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 1 における例では，$\tau(v_{i,j}) = log_2 13 \approx 3.7$ となる．そのため，もしある著者が 100 字の記述を行い，次の編集において全て削除してしまった場合，$R(i, j, 0) = 100$ となるが $R(i, j, 1) = 0$ となるため，$\tau(v_{i,j}) = R(i, j, 1) = 0$ となる．

3.3.3　著者の信頼度 (1)

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，3.2 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 (3) 式で計算される．

$$U_e = \frac{\sum_{v_{i,j} \in A_e} \tau(v_{i,j})}{|A_e|} \tag{3}$$

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．

ここで，著者の信頼度を算出する際に特殊な記事，例えばノートや Wikipedia の規約に関する記事など特殊な用途に利用される記事は，著者の信頼度の際に考慮しないこととする．なぜなら，これら特殊な用途に利用される記事は著者の意見が書かれることが多く，信頼度が高いかどうかに関係無く消去されない．そのため，これらの記事に記述した著者の信頼度が非常に高くなってしまうためである．

3.3.4　記事の変更に対する信頼度 (2)

次に，著者の信頼度を利用して記事の信頼度を算出する．3.3.2 節の (2) 式に，著者の信頼度による信頼度を追加したも

ので，次のように計算する．

$$\tau_k(v_{i,j}) = \alpha \sum_{p=1}^{N_i - j} \log_2 R(i,j,p) - (1-\alpha) \cdot ((R(i,j,0) - R(i,j,p))) \cdot u(p)) \quad (4)$$

ここで $u(p)$ は，$i = 1, 2, \cdots p$ において削除した著者 e の信頼度 U_e を平均したものである．つまり，記事を削減した著者の信頼度が低いとき，$u(p)$ の値が低くなるため，実際に削減された文字数よりも小さな文字数が削減されることになる．つまり，第 2 項の値が小さくなるため，削減されることによる信頼度の低下は小さくなる．一方，削除を行った著者の信頼度が高いとき，第 2 項の値が大きくなるため，記事の変更に対する信頼度は大きく低下する．

3.3.5 著者の信頼度 (2)

3.3.3 節と同様に，著者の信頼度 U_e は (5) 式で計算される．

$$U_e^k = \frac{\sum_{v_{i,j} \in A_e} \tau_k(v_{i,j})}{|A_e|} \quad (5)$$

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．

この処理を行った後，3.3.4 節で再度記事の変更に対する信頼度を算出する．

3.3.6 記事の信頼度

3.3.4 節と 3.3.5 節による記事の変更に対する信頼度，著者の信頼度を交互に算出した後，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，多くの部分を編集している著者の信頼度が高ければ，その記事の信頼度が高いと考える．

$T_i(v_{i,j})$ は，(6) 式で計算される．$v_{i,j}$ 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．

$$T_i(v_{i,j}) = \frac{\sum_{k=1}^{j} (U_e \cdot c_{e,k})}{\sum_{k=1}^{j} c_{e,k}} \quad (6)$$

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. 実　　装

我々は，Web インタフェースを利用して提案手法の実装を行った．Web ブラウザを利用して実装システムにアクセスすることによって，Wikipedia の記事の部分に対してどの程度の信頼度が付与されているのかを直感的に閲覧することができる．

4.1 実 装 画 面

本節では，実装したシステムの表示画面とその内容について解説する．図 2 に，提案手法によって Wikipedia の記事を閲覧する際の画面を示している．この画面では上から記事のタイトル，記事が更新された日時が表示され，記事のバージョンに対する信頼度が表示されている．ここで表示されている信頼度は 3.3.6 節の (6) 式によって算出された数値である．

その下に表示されている記事は，Wikipedia に投稿された記事の元となる記述である．Wikipedia では，例えば表や画像などテキストだけでは表現できない様々な記述の表現を行うために，MediaWiki 記法と呼ばれる独自の記法が存在する．また，Wikipedia を実際に閲覧する場合には，これら MediaWiki 記法を一般的な HTML に変換して表示を行っているため，可読性が高い表示となっている．ところが，MediaWiki 記法を HTML に変換して表示を行う際に，様々な情報が欠落してしまう場合がある．たとえばある文字列を他の Web 文書へリンクする文字列へ変更を行うときや，表のデザインを変更するときなどの，記事の文書構造を変更する場合などである．そこで，このような変更をすべて表示するために，まず MediaWiki 記法によって書かれている文字列そのものを利用者に提示するような実装を行った．この部分は可読性が低下しているという問題があることから，今後一般的な HTML による情報提示も行うことができるように改良を行う予定である．

図 2　実装システムによる記事の閲覧画面

また，記事の元となる記述に対して，色によって信頼度を表示している．文字の背景色が青いものは信頼度の高い記述，赤いものは信頼度の低い記述であることを示している．ここで，信頼度に対応する色を選択するために 3.3.3 節で述べた，著者の信頼度を算出するための式である (3) 式を利用した．著者に対する信頼度によって著者の順位付けを行い，著者の信頼度が上位 10% の著者が記述した部分は青い字とし，上位 30% の著者が記述した部分は薄い青を文字の背景とした．また，下位 30% の著者が記述した部分は薄い赤を，下位 10% の著者が記述した部分は字を濃い赤とした．図 2 では，多くの記述は信頼度が高く，信頼度が低い記述が少し書かれている様子が分かる．

次に，図 3 においてバージョン一覧を表示している部分を示している．まず，画面の上部では編集を行う過程で編集がどのように行われているかをグラフによって図示している．グラフの x 軸は編集回数であり，y 軸は文字数を示している．このグラフから，どの著者が記事における主要な著者であるかを示している．さらに，グラフには信頼度の編集回数による遷移を

図 3　実装システムによるバージョン一覧の画面

図 4　実装システムによる著者情報の画面

示している．ここで示している信頼度は記事の信頼度であり，3.3.6 節の (6) 式によって算出した数値である．

グラフの下部には，記事のバージョン一覧を示している．一つの行には，バージョンが作成された日時，および作成した著者の名前，そのバージョンの信頼度を示している．この信頼度の数値はバージョン一覧の画面における信頼度と同じ値である．日時と著者の名前はリンクが作成されており，日時をクリックするとそのバージョンの記事閲覧画面，著者をクリックすると次に示す著者情報の画面を閲覧することができる．

図 4 において著者情報を表示している部分を示している．画面の上部では，著者の信頼度がどのように遷移しているかを示すグラフを示している．グラフの x 軸は編集回数であり，y 軸は信頼度を表している．このグラフを閲覧することによって，著者の信頼度がどのように変化しているかを知ることができる．

4.2　実験データ

本実装では，2010 年 11 月 2 日現在の日本語版 Wikipedia における編集履歴(注4) を利用した．このデータでは，1,889,129 件の記事，24,054,128 個のバージョンが含まれており，2,178,003 人の著者によって記述されている．このデータは bzip2 によって圧縮された XML データであり，約 20.1 GB ある．

(注4)：http://download.wikimedia.org/jawiki/20101102/jawiki-20101102-pages-meta-history.xml.bz2

図 5　実装システムによる記事一覧の画面

4.3　データ解析，表示

本実装は二つの部分によって構成されており，一つはデータ解析部分，もう一つは表示部分である．二つの部分はデータベースを共有しているため，データ解析が行われると即座に表示部分に反映されるようになっている．

データ解析部分は，3.3 節で述べているダウンロードデータからのデータ抽出部分と，信頼度算出部分に分けられている．Sun Java Standard Edition JDK 1.6.0_22-b04 によって実装し，データを格納するためのデータベースとして，mongoDB 1.7.5 を利用した．バッチ処理によってダウンロードデータから信頼度解析までを一度に処理している．データ解析に必要な時間は，約 10 日程度である．これは，CPU として Intel Core i7 920，メモリを 10GB に設定していた場合である．データベースによるデータ格納に必要な容量は 約 1,800 GB であった．

Wikipedia は日々更新されており，それに伴って新たに編集履歴も公開されている．ところが現在はその差分だけを利用して，システム上のデータを更新するプログラムをまだ実装していない．そのため，今後新たな編集履歴データを入力として利用するときには同程度の時間がかかるが，差分情報だけを利用したデータ更新を行うことが可能となれば，短時間で新しいデータへ差し替えを行うことが可能となると考えられる．

表示部分は，Ruby 1.9.2 と Ruby on Rails 3.0.3 を利用して実装をした．データ格納部分は，データ解析部分と同様に mongoDB 1.7.5 を利用した．また，グラフ表示を行うためには HighCharts (注5) を利用した．Rails 上では信頼度算出やグラフ生成などの処理を行わず，データベースに格納されている数値を加工して表示することだけに利用しているため，高速に表示を行うことができる．

(注5)：http://www.highcharts.com/

5. 評価実験

本論文で述べている提案手法によって高い精度で信頼度を算出することが可能であることを確かめるために，評価実験を行った．本評価実験の目的は，3.3.4 節および 3.3.5 節で述べた，記述の信頼度を算出する際に著者の信頼度を利用した方法が著者の信頼度を用いない方法と比較して精度が向上するかどうかを調査した．この評価実験では，3.3.4 節および 3.3.5 節で述べた手法を利用せずに 3.3.6 節で述べた記事の信頼度を算出する方法と，3.3.4 節および 3.3.5 節で述べた手法を利用して記事の信頼度を算出する方法の比較を行っている．二度以上著者の信頼度を繰り返し用いることも可能であるが，簡単のため本論文ではこれらの比較は行わない．

評価の方法は既存の情報検索における精度比較手法と同じように，再現率，精度を利用した手法に準じた手法を利用する．情報検索における再現率，精度による評価手法では，まず検索対象文書集合を準備し，被験者によって問合せを準備し，その問合せに関連する文書群を事前に選択する．次に，システムによって関連文書群を検索し，システムによる文書群と被験者によって選定された関連文書群を比較することによって再現率および精度を算出する．このとき，人手による関連文書群の精度，つまり問合せに関連する文書を人手でどれほど正しく選択できているかが，評価実験そのものの正しさに直結する．

本研究では，関連文書群に相当するものとして信頼度の高い記事を選択することによって，再現率，精度を算出することができると考えられる．ところが，人手であっても記事の信頼度が高いかどうかを判定することは困難であり，また選択された信頼度が高いかどうかを検証する手段も存在しない．さらに，Wikipedia では非常に信頼度の低い記事は少ない．そこで，信頼度の高い記事として本研究では秀逸な記事(注6)および良質な記事(注7)を利用して評価を行う（**評価実験 1**）．さらに人手による評価として，システムにより信頼度を算出し，信頼度が高いものから順に並べた記事リストのうち上位 100 件のうち最新の記事を第一著者によって閲覧し，信頼度が高いと判定できる記事であるかどうかを判定した（**評価実験 2**）．

まず，算出された著者の信頼度の分布を図 6 に示す．この図では，x 軸には著者 ID を表し，y 軸にはその著者の信頼度を表している．著者が 20 万人近く存在するため，このグラフでは無作為に 5000 人の著者を取り出し，グラフに表示している．この図から分かるように，非常に少ない著者が非常に高い信頼度であることに対して，非常に多くの著者が低い信頼度であることが分かる．

5.1 評価実験 1: 秀逸な記事，良質な記事による評価

秀逸な記事，良質な記事を正解集合としたときの 11 点再現率-精度グラフを示した．ここで，秀逸な記事は 87 本，良質な記事は 499 本であり，合わせて 586 本が正解集合である．再現率-精度曲線を図 7 に示す．

(注6)：http://ja.wikipedia.org/wiki/Wikipedia:秀逸な記事
(注7)：http://ja.wikipedia.org/wiki/Wikipedia:良質な記事

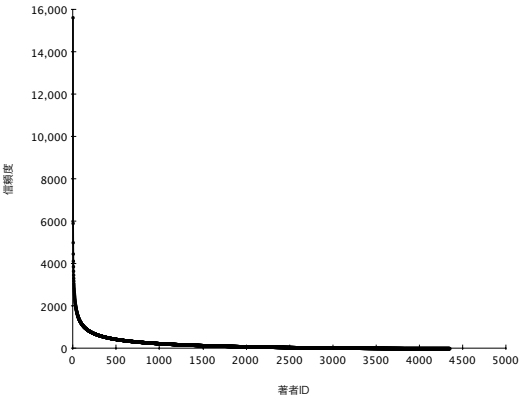

図 6　著者とその信頼度

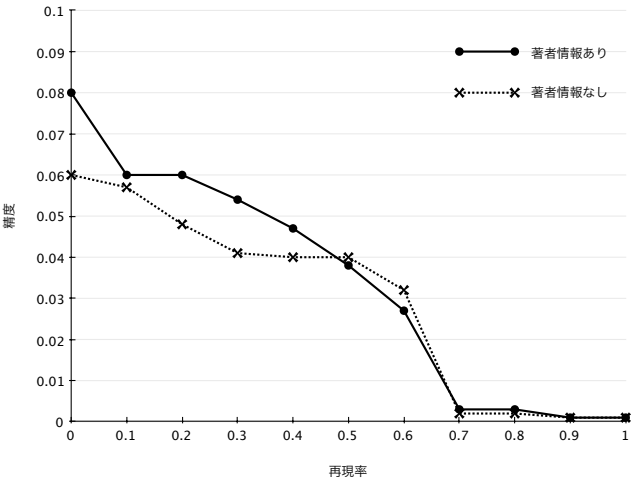

図 7　11 点再現率-精度グラフ

この評価実験から，確かに信頼度算出精度が向上したことが分かった．実験データを詳しく見ると，編集合戦のように不適切な編集が行われたときに信頼度が低下していた著者について，その信頼度の低下が軽減されていることが分かった．そのため，編集過程において不適切な編集が行われているとき，それらの記事の信頼度を上げることができた．一方，不適切な編集が行われていない記事や，編集があまり行われていないにも関わらず良質な記事が存在している場合であっても，それらの記事の信頼度は変化しなかった．そのため，不適切な編集が行われている記事に対しては本手法が有効であるが，適切な編集だけが行われている記事に対してはあまり有効ではないことが分かった．

5.2 評価実験 2: 人手による評価

次に，記事の信頼度を人手で評価することによって，精度を測定した．この評価では，著者情報を利用した方法と利用しなかった方法により全ての記事に対して信頼度を求め，信頼度が高い記事から順に並べ，記事のリストを作成する．記事のリストからそれぞれ上位 100 件を取り出し，人手によって信頼度が高い記事であるかどうかを分別する．このとき，10 回以下の編集によって行われている記事や，10 行以下の短い記事は記事リストに含めなかった．なぜならば，これらの記事は信頼度を判定するためにはあまりにも短いためである．また，ノートや利

用者ページなどの特殊な用途で作成された記事も除いた．

その結果，著者情報を利用しなかった場合には 48 件の記事を信頼度が高いと判定され，著者情報を利用した場合には 61 件の記事を信頼度が高いと判定された．この結果からも，提案手法が有効であることが分かった．

このとき，一覧やリストのようなページが信頼度が高いと判定される傾向にあった．これらの記事は確かに信頼度が高いと考えられるが，実際にはさらに信頼度が高い情報が Wikipedia には含まれると考えられる．そこで，記事の性質による分別，つまり記事が事象データベースのような追記型であるか，それともある一つの事柄について何度も推敲を重ねるような記事であるかによって信頼度算出手法を変更する必要があるといえる．

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，著者の信頼度を利用して記述の信頼度を算出するための手法を提案した．信頼度が低い著者は信頼度の高い記述も削除してしまう可能性が高いと仮定し，著者の信頼度によって記述の信頼度に与える影響を変化した．評価実験によって，実際に記事に対する信頼度の算出精度が向上したことを確認した．

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 [10] 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．

- **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 [11] 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．また，少量の編集によって大きく意味を変更する可能性もある．例えば，「A である」という部分を「A でない」に変更すると，文字数としては 2 文字の変更であるが，全く逆のことを言っていることになる．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**

我々は 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 [12] [13] 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．

謝　辞

本研究の一部は，文部科学省科研費 (20300036, 20500104, 21013026) による．ここに記して謝意を表す．

文　献

[1] 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および速度の改善. 情報処理学会論文誌: データベース, Vol. 3, No. 3, pp. 20–32, 2010.

[2] Besiki Stvilia, Michael Twidale, Linda Smith, and Les Gasser. Information quality work organization in wikipedia. *J. Am. Soc. Inf. Sci. Technol.*, Vol. 59, No. 6, pp. 983–1001, 2008.

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

[4] Thomas Wöhner and Ralf Peters. Assessing the quality of wikipedia articles with lifecycle based metrics. In *WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, pp. 1–10, 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*, pp. 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.* , 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 *CIKM '07: Proceedings of ACM International Conference on Information and Knowledge Management*, pp. 243–252. ACM, 2007.

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

[10] Elaine G. Toms, Tayze Mackenzie, Chris Jordan, and Sam Hall. wikisearch: enabling interactivity in search. In *SIGIR*, p. 843. ACM, 2009.

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

[12] Todd Holloway, Miran Bozicevic, and Katy Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors. *Complexity*, Vol. 12, No. 3, pp. 30–40, 2007.

[13] Benoît Otjacques, Maël Cornil, and Fernand Feltz. Visualizing cooperative activities with ellimaps: The case of wikipedia. In *CDVE '09: Proceedings of Cooperative Design Visualization and Engineering*, Vol. 5738, pp. 44–51., 2009.