信頼度を考慮した知識の構造化

鈴木　優† 石川　佳治†,††

† 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
†† 国立情報学研究所　〒 101–0003 東京都千代田区一ツ橋 2 丁目 1-2
E-mail: †suzuki@db.itc.nagoya-u.ac.jp, ††ishikawa@itc.nagoya-u.ac.jp

あらまし　本稿では，情報の信頼度を考慮したオントロジーの作成手法について述べる．現在，Wikipedia のように不特定多数の利用者によって作成されるコンテンツ (User Generated Contents; UGC) が普及している．また，これらの情報を知識，オントロジーとして自動的に体系化する試みが数多く行われており，例えば YAGO や DBPedia などが挙げられる．ところが，これら UGC は必ずしも正確な情報であるとは限らない．そのため，これら UGC を基準に作成されたオントロジーは不正確な情報が混在する．本研究ではこの問題を解決するために，Wikipedia の編集履歴からコンテンツの信頼度を算出し，オントロジーを作成する際のメタデータとして信頼度の付与を行う．利用者は信頼度メタデータが付与されたオントロジーを利用することによって，大規模なオントロジーを高速に，高精度に作成することができる．

キーワード　信頼度，オントロジー，User Generated Contents

Knowledge Construction using Credibility

Yu SUZUKI† and Yoshiharu ISHIKAWA†,††

† Information Technology Center, Nagoya University　Furo, Chikusa, Nagoya, Aichi 450–0002, Japan
†† National Institute of Informatics　2-1-2, Hitotsubashi, Chiyoda, Tokyo 101–0003, Japan
E-mail: †suzuki@db.itc.nagoya-u.ac.jp, ††ishikawa@itc.nagoya-u.ac.jp

Abstract　In this paper, we propose a method to construct ontology with information credibility. UGC (User Generated Contents), such as Wikipedia and SNSs are widely used as knowledge bases. Several related methods, such as YAGO and DBPedia automatically construct ontologies using these UGCs. However, these ontologies usually contains not credible information, because UGCs do not always consist of credible information. In this paper, we propose a method to construct ontology with information credibility using credibility degree of UGCs. Using ontology with credibility information, users can easily understand which relation in the ontology is credible or not.

Key words　Credibility, ontology, User Generated Contents

1. はじめに

現在，様々な事象に対してそれぞれの関係を表現することによって，知識を体系化する必要性が高まっている．例えば “コンピュータ” と “計算機” は同意語であるが，Web 情報検索において二つの単語が同意語であるという情報，つまり知識を利用しなければ異なる語として認識されるため，“コンピュータ” で検索を行った際に “計算機” が含まれ “コンピュータ” が含まれない文書は検索されない．その他にも，上位語や下位語に関する知識を利用することによって，より高度で精度の高い情報検索を行うことができる．

このような体系化された知識はオントロジと呼ばれ，様々な方法で作成されている．例えば，WordNet (注1) や日本語 WordNet (注2) は人手で作成された大規模なオントロジであり，情報検索やセマンティックウェブなどの分野で活用されている．人手で作成されたオントロジは，精度が高いという特徴がある一方，収録されている語，概念数が少ないという問題点がある．そこで，このような問題を解消するために，既存の情報源から語や概念を抽出することによって，大規模なオントロジを自動的に構築するための手法が必要となってきている．

そのような試みの一つが YAGO [1] である．YAGO では，

（注1）：Princeton University “About WordNet.” WordNet. Princeton University. 2010. ¡http://wordnet.princeton.edu¿
（注2）：NICT, “日本語 WordNet” http://nlpwww.nict.go.jp/wn-ja/

Wikipedia において記述されている様々な情報から語や語義を抽出し，それらの関連も自動的に抽出している．この手法によって，収録されている語や概念数が WordNet では約 20,000 個であったところが YAGO では 約 600,000 個となり，およそ 30 倍となっている．一方で，収録されている語や概念の抽出精度は WordNet と比較して低下している．この問題は二つの原因によって引き起こされると考えられ，一つは Wikipedia から単語や概念を抽出する際の手法に起因するもの，もう一つは情報源となる Wikipedia 自身の信頼度に起因するものである．本研究では，後者のほうの原因である，Wikipedia 自身の信頼度に起因する精度低下が起こらないようにするため，Wikipedia の記述に対する信頼度を自動生成されたオントロジに反映させる方法を提案する．

事象間の関連を算出する手法として，記事の信頼度，著者の信頼度，そして記述の信頼度を利用する方法の三つの方法を提案する．信頼度は様々なレイヤで算出されるため，それぞれの方法を利用することによって異なる信頼度を得ることができる．記事全体に着目する方法と，記事においてリンクの部分だけに着目した方法という二つの観点から考案した方法である．そこで，これらの手法を利用して実際にオントロジに対して信頼度を構築する．

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．一つは利用者が明示的に対象を評価する方法，もう一つは利用者が暗黙的に評価を行う方法である．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3) 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4) 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5) において，ほとんど全ての利用者が 5 つ星を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 [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．

(注3)：http://www.amazon.com

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

(注5)：http://youtube-global.blogspot.com/2009/09/five-stars-dominate-ratings.html

3. Wikipedia の信頼度算出手法

本研究の目的は，事象間の関係とその信頼度を DAG （Directed acyclic graph; 閉路を持たない有向グラフ）によって表現することである．ここで事象を Wikipedia の記事のタイトルで表現される文字列とする．これによって，同義語は異なる事象として表現される．

本研究では，次のような手順で信頼度を算出する．

（1） Wikipedia から編集履歴データを取り出す．

（2） 編集履歴から，記述の残存率を計算する．

（3） 残存率を利用して，記述の信頼度を計算する．

次に，ここで得られた信頼度を利用して，事象間の関連とその信頼度を算出する．

（1） 記事のリンク構造を利用して事象間の関連を抽出する．

（2） リンク元とリンク先の記事全体の信頼度から，事象間の信頼度メタデータを付与する．

本研究ではまず，Wikipedia の編集履歴から信頼度を算出する方法について述べ，次に事象のリンク間の信頼度についての算出手法について述べる．

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と考えられる．

3.2 Wikipedia の記事信頼度算出

3.2.1 Wikipedia の記事のモデル化

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．

3.2.2 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 $\tau(v_{i,j})$ を算出する．ここで妥当な編集の定義として，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，$add_{i,j}$ と $del_{i,j}$ が残存している割合を算出する．

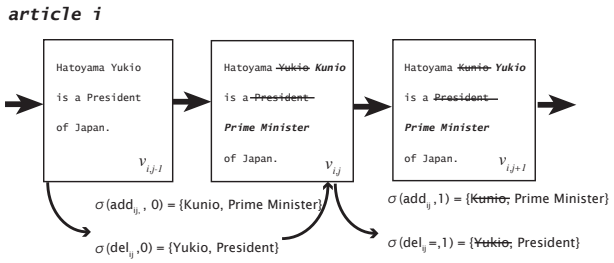

図 1　編集履歴における追加と削除

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

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 (1) ，(2) 式によって求める．

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

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

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 1 における例では，$p = 1$ の時の追加残存率 $R^{add}(i, j, 1)$ 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 (3), (4) 式によって求める．

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

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

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 1 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，(5) 式で求める．

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

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 1 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 (6) 式で求める．

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

3.2.3　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，3.2.1 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 (7) 式で計算される．

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

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．

3.2.4　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，(8) 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．

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

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.　事象間の関連算出

Wikipedia における記事の間には，リンクが張られている．それらのリンクを事象間の関連であると考え，リンクとして抽出する．オントロジの構築においては通常，事象間の関連として上位語や下位語，同位語などの関連を付与することが多いが，本研究ではそれら関連の種類については扱わない．

事象間の信頼度を算出するとき，算出に利用する信頼度として記事の信頼度，著者の信頼度，そして記述の信頼度の三つを考えることができる．そこで，以下ではこれら三つの方法について述べる．

4.1　方法 1: 記事の信頼度を利用する方法

まず，記事の信頼度を利用して事象間関連の信頼度を算出する．ここでは，事象間の信頼度は，二つの事象それぞれの信頼度の平均であるとする考え方である．二つの事象に関する記事の信頼度が高ければ，二つの事象を繋ぐリンクも正確に作成されていると考えられるため，事象間の信頼度は高くなるといえる．また，記事の信頼度が低いときには，二つの事象は実際にはあまり関係が無い場合もあり，信頼度が低いといえる．

記事 d_i から d_j に対してリンクが張られているとき，関連 $r_{i,j}$ を d_i, d_j の間に作成する．このとき，リンクが何回張られていても，1 個の関連だけしか作成しない．関連には方向があり，$r_{i,j}$ は d_i から d_j の方向を表している．

次に，それらのリンクに対して信頼度を算出する．ここで関連の信頼度は，記事 d_i と d_j との信頼度の平均とする．つまり，d_i と d_j 両方の信頼度が高いとき，信頼度の高い関連であるとする．まず，記事 i の信頼度 T_i を求める．3.2.4 節では，バージョンごとに信頼度を算出していたが，事象間の関連を算出する際には，最も新しいバージョンを対象にして信頼度を算出する．つまり，T_i は次のように求められる．

$$T_i = T_i(v_{i,N_i}) \tag{9}$$

次に，関連の信頼度を算出する．記事 d_i から d_j への関連の信頼度 $\tau^1_{i,j}$ を次のように定義する．

$$\tau^1_{i,j} = \frac{T_i + T_j}{2} \tag{10}$$

以上の式で，事象に対して信頼度を算出することができる．

この手法は，記事全体の信頼度を事象の信頼度に反映させるため，リンクを作成した著者以外の信頼度を関連の信頼度としている．そのため，一人の著者の信頼度だけに関連の信頼度が左右されず，極端に高いもしくは低い信頼度とはならないという利点がある．ところが，記事の信頼度と関連の信頼度には関係があるとは必ずしもいえないため，不適切な信頼度が得られてしまう可能性が高いという問題点もある．

4.2 方法 2: 著者の信頼度を利用する方法

次に，著者の信頼度を利用して事象間の信頼度を算出する方法について述べる．この方法では，事象間の信頼度は，二つの事象を繋ぐリンクを作成した著者の信頼度であるとする考え方を利用する．事象間の関係を表すリンク自身を作成した著者の信頼度が高いとき，その事象間の関連には信頼度が高いという考え方である．

記事 d_i から d_j への関連の信頼度 $\tau^2_{i,j}$ を次のように定義する．

$$\tau^2_{i,j} = U_e \tag{11}$$

ここで U_e は 3.2.3 節の (7) 式において計算された，リンクを作成した著者 e の信頼度である．

この手法は，著者の信頼度を事象の信頼度としているため，適切な著者がリンクを作成すると適切な関連が得られるという仮定がある．ところが，著者は常に同じ信頼度の記述を行っているとは考えにくく，ある記事では適切なリンクを作成しており，別の記事では不適切なリンクを作成している場合も考えられる．このような場合，不適切な信頼度が得られてしまうという問題がある．一方，新しく作成されたリンクであり記述の信頼度を算出することができない場合であっても適用できる手法であるため，新しく作成された事象間の関連に対しても信頼度を算出することができる点が利点である．

4.3 方法 3: 記述の信頼度を利用する方法

最後に，記述の信頼度を利用して事象間の信頼度を算出する方法について述べる．この方法では，二つの事象を繋ぐリンクを表現する記述自身が長い間残存しているとき，事象間の信頼度が高いと考える．

記事 d_i から d_j への関連の信頼度 $\tau^3_{i,j}$ を次のように定義する．

$$\tau^3_{i,j} = \overline{\tau(v_{i,j})} \tag{12}$$

ここで $\overline{\tau(v_{i,j})}$ は，3.2.2 節の (6) 式で計算された，リンク部分の記述の残存率である．

この手法は，リンクを作成した著者や，リンク以外の部分を記述した著者などを考慮しない手法であるため，記述自身が長い間残存しているとき，正しく信頼度を判定することができる．ところが，新たに作成されたリンクなどから生成された事象間の関連については，その記述に対して信頼度を算出することができないため，精度の高い事象間の信頼度を算出することが困難となると考えられる．

5. おわりに

本研究では，知識を体系化するために，事象間の関連の有無を調べると共に，その関連についての信頼度を算出する方法について述べた．Wikipedia の編集履歴において，ほかの編集者による編集を経てなお残存している記述は信頼度が高いという考え方から，記事に対して信頼度を算出した．そして，一つの記事を一つの事象として考え，リンク構造から事象間の関連を取り出し，それぞれの関連に対して信頼度を算出した．このとき，事象の信頼度からその関連の信頼度を算出した．

事象間の関連を算出する手法として，三つの方法を提案した．記事の信頼度，著者の信頼度，そして記述の信頼度を利用する方法である．信頼度は様々なレイヤで算出されるため，それぞれの方法を利用することによって異なる信頼度を得ることができる．記事全体に着目する方法と，記事においてリンクの部分だけに着目した方法という二つの観点から考案した方法である．そこで，これらの手法を利用して実際にオントロジに対して信頼度を構築する．

今後の課題として，信頼度算出手法に関する課題が挙げられる．事象の信頼度，つまり記事の信頼度を基準として関連を調べることができるが，他の考え方としてリンク構造の残存率から信頼度を算出する方法も考えられる．つまり，多くの編集を経てなお存在するリンクは，信頼度が高いとする考え方である．本研究で提案した手法との比較を行い，実際に信頼度が高いかどうかを調査する必要があると考えている．

もう一つの課題は，本手法の精度を計測するための方法についてである．本研究では，知識を体系化した上でその信頼度を共に算出する方法について提案している．ところが，我々の知る限りこのようなことを行っている研究は存在しない．そのため，他の関連研究との比較を行うことが困難であると共に，評価尺度としてどのようなものを用いる必要があるかについて，標準的に用いられている方法が存在しない．そのため，評価尺度について定める必要があると考えている．

謝　辞

本研究の一部は，最先端研究開発支援プログラム (FIRST) に

よる．ここに記して謝意を表す．

文　　献

[1] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *WWW*, pages 697–706. ACM, 2007.

[2] 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.

[3] 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.

[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*, pages 1–10, New York, NY, USA, 2009. ACM.

[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] 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.