社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－DD－58 (20)
2006－EIP－34 (20)
2006／12／1

Wikipediaマイニングによる信頼性情報を考慮した記事関係の抽出

中山浩太郎 † 原 隆浩 † 西尾 章治郎 †

Wikipediaは，WWW上に構築された百科事典であり，誰もが簡単にWebブラウザを通じて編集可能であるために，膨大な数の記事が投稿，公開されている．しかし，2005年末から2006年初頭にかけて，虚偽の記事が投稿されるような事件が発端となり，コンテンツの信頼性が大きな問題となってきた．筆者らは，これまでの研究でWikipediaにおける記事同士の関係性を抽出する手法について提案し，その有効性を証明してきたが，このように不特定多数のユーザがコンテンツを管理するような環境においては，信頼性を考慮した解析手法が重要となる．本研究では，Wikipediaのダイナミクスと信頼性の問題を分析するとともに，リンク構造解析アルゴリズムについて検討し，記事関連性抽出における信頼性の高い情報抽出方法を模索する．
キーワード Wikipedia，Webマイニング，Webの信頼性

Wikipedia Mining for Associated Document Extraction Based on Trust Information

KOTARO NAKAYAMA ,† TAKAHIRO HARA †
and SHOJIRO NISHIO†

Wikipedia is a Web-based dictionary that can easily be edited through Web browsers by any Internet user. Thus huge amounts of articles are published and managed on it. However, after a number of article reliability issues, the trust problem on Wikipedia is still in controversy. In previous works, we proved the effectiveness and potential of the article association extraction based on Wikipedia mining. In this paper, we first analyze the link structure of Wikipedia and dynamics of Wikipedia. Then, we present an effective method for link structure mining for Wikipedia and describe how link structure mining for Wikipedia is helpful for extracting trusted information.
Keywords Wikipedia, Web mining, Trust on the WWW

1. まえがき

近年，インターネットユーザの爆発的な増加と多様化に伴い，WWW上でブラウザを通じてコンテンツを管理するタイプのCMS（Content Management System）が注目を集めている．BlogやWiki，Xoopsなどはその代表例である．この中でも最近ではユーザ同士が協調してコンテンツを管理するタイプのWiki[10]が注目を集めている．Wikiは，ユーザがWebブラウザを通じて追加・削除・編集などのコンテンツ管理が可能なCMSであり，容易にコンテンツ編集が可能なため，個人ユーザから企業ユーザまで多くの場面で利用されている．実際に運用されているWikiのほとんどは，ログイン機能によってコンテンツ管理者を制限しているが，Wikipedia[16]のように訪問者が自由に編集できるコンテンツも存在する．

Wikipediaは，Wikiを利用して構築された百科辞典であり，文化，歴史，数学，科学，社会，テクノロジなどの幅広い分野の語（記事）をカバーしている．Wikipediaでは，Webブラウザを通じて，他のユーザと議論しながら自由に記事を投稿できることが最大の特徴である．この結果，Wikipediaのコンテンツ量はここ数年で爆発的に増大し，一年前の2005年9月には75万件程度であった記事数は，2006年9月の段階で2倍以上の168万記事（英語のみカウント）を超え，さらに日に日にその量を急速に増やしている．世界最大の百科辞典であるEncyclopedia Britannicaの記事数が約6万記事であることと比較してもその数が膨大であることがわかる．筆者らは，このような膨大な記事を持つWeb辞典のリンク構造を解析するこ

† 大阪大学大学院情報科学研究科マルチメディア工学専攻
Department of Multimedia Engineering,Graduate School of Information Science and Technology,
Osaka University

とで，記事の関係性を抽出し，語彙同士の関係性を示すシソーラス辞書[2),5)]を構築することを目指した研究を行っている．

しかし，Wikipediaでは，2005年末から2006年初頭にかけて，虚偽の記事が公開や意図的な情報操作に対する脆弱性に起因する事件がいくつか発生し，そのコンテンツの信頼性に対して疑問が投げかけられている．そのため，精度高く重要な語彙関係を抽出するためには，コンテンツの信頼性を考慮した解析手法が必要であるといえる．

本論文では，Wikipediaのダイナミクス，リンク構造，信頼性の問題を分析した後，記事同士の関係性を抽出する方法について検討する．さらに，共起性解析，tfidfなどの従来手法および新手法によるシソーラス辞書構築の方法を提案・適用し，比較することによって，信頼性の高い関係性抽出方法を検討する．

本稿の以下では，2章でWikipediaのダイナミクスについて述べ，3章でWikipediaから有用な知識を抽出する手法について記述する．また，4章では3章で提案する手法を比較し，信頼性の高い情報抽出方法について検討する．最後に，5章でまとめと今後の展開を記述する．

2. Wikipediaのダイナミクス

Wikipediaは，Webブラウザを通じて複数のユーザがコンテンツを共有する百科事典である．Wikipediaの特徴は，その管理体制にあり，Webブラウザを通じて誰でも自由に閲覧できるだけでなく，記事の修正が可能である．その結果，1）膨大な量のコンテンツ，2）即時性の高いコンテンツ管理体制，3）密なリンク構造，4）URLによる語彙の一意性確立，5）質の高いリンクテキストなど，知識獲得のWebコーパスとして見たときに，他のWebコンテンツには見られない多くの特徴的な利点が存在する（図1）．本章では，これらWikipediaの特徴を解説する．

2.0.1 即時性の高いコンテンツ管理体制

従来の辞書では，一般的な語からトップダウン的に追加されていくのが通常であり，一般的でない語や専門的な語は辞書に追加されるのが遅れる，もしくはいつまでも登録されないのが一般的である．しかし，Wikipediaでは，インターネットを通じてリアルタイムに記事が公開・アップロードされ，リンクが構築されていくため，極めて即時性が高い．例えば，ある企業から最新の技術の発表があった数時間後には，エントリが生成され，その説明や詳細なスペック，画像などが他の語へのリンク付きで公開されたというケースもある．このような新しい概念に対する網羅性の高さはWebコーパスとしてみたときの重要な特徴の一つである．

2.1 密なリンク構造

筆者らは，予備実験として2006年9月の段階におけるWikipedia内におけるリンクの数をカウントした．約168万ページ（英語のみ）を解析したところ，約4,998万の内部リンク（リダイレクトリンクを除く）を抽出した．これは，一ページあたり平均29.62のリンクを持つ計算となる．また，Backwardリンクの分散を調査したところ，1万以上のBackwardリンクを持つ記事は196件，1,000以上のBackwardリンクを持つページは3,198件，100以上のBackwardリンクを持つページにいたっては67,515件も存在することを確認した．しかも，これらのリンクはサイト内に対するリンクのみをカウントしたものであり，サイト外へのリンクは含まれていない．これは，Wikipediaでは閉じられた語彙空間の中で密なリンク構造を持っており，リンク構造を解析することで有用な情報を抽出できる可能性を示している．

2.2 URLによる語彙の一意性

URLによって語彙の一意性が確立されている点は，Wikipediaの大きな特徴の一つである．電子辞書では，通常一つの見出し語が一つのページに割り当てられており，その中で複数の意味について詳述される（図2左）．一方，Wikipediaでは一つのURL（ページ）に一つの概念が割り当てられており，多義性がURLによって解決されている点が大きな特徴である．たとえば，「Football」は強いコンテキスト依存を持つ単語であり，アメリカンフットボールを示す場合もサッカーを示す場合もある．Wikipediaでは，これら二つの概念は別のページで管理されており，それぞれ「http://en.wikipedia.org/wiki/American_Football」「http://en.wikipedia.org/wiki/Football_(soccer)」（図2右）という別々のURLが割り当てられている．

このように，概念とURLが一対一で対応していることは，概念の関連性を解析する際に多義性やコンテキストの依存性の影響を受けずに解析できることを示している．

2.3 質の高いリンクテキスト

リンクテキストは，多くの場合リンク先のページ内容の要約であることは，最近の研究において証明されている．しかし，通常のWebページにおけるリンクは，「最新情報はこちらをクリック」といったようにリンク先の概念とは無関係の情報が多く含まれる場合が多い．このようにノイズの多いリンクテキストを解析

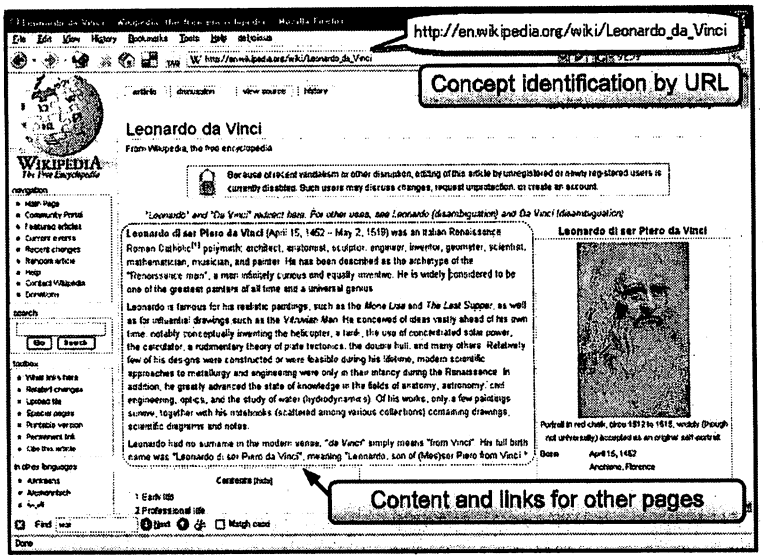

図 1 Web コーパスとしての Wikipedia
Fig. 1 Wikipedia as a Web corpus.

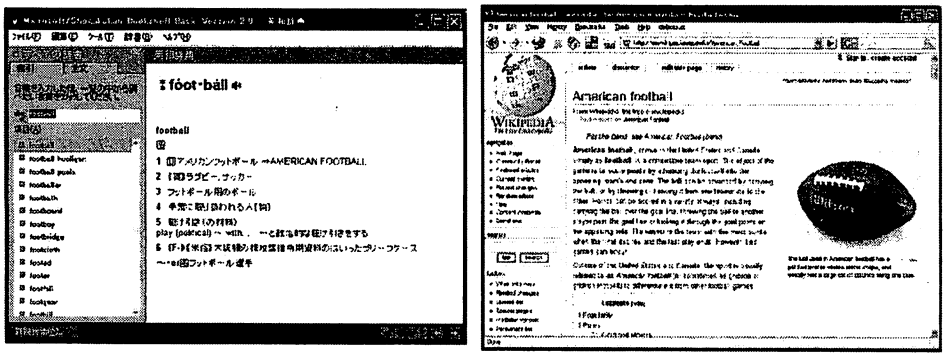

図 2 電子辞典と Wikipedia の語彙の一意性の違い
Fig. 2 The difference of identification of concept with ordinary dictionaries.

し，リンク先のページの要約とする場合には，Chen らの手法[3] のように，自然言語処理ツールを利用することにより，リンクテキストを解析し，統計的にリンク先の概念ラベルを抽出するのが一般的である．しかし，このように自然言語処理を利用してリンクテキストを解析した場合，シソーラス辞書の精度が低下することが予備実験によって判明している．

Wikipedia においては，図 1 に示すとおり，リンク内のテキストはリンク先の概念の要約を端的に表す語であり，多くの場合は慣例的に記事のタイトルが利用される．

2.4 Wikipedia の信頼性

以上のとおり，Wikipedia は知識獲得のための Web コーパスとして見たときに多くの魅力的な特徴を持つ．また，Web ブラウザを利用して誰でも更新が可能であるため，間違いが迅速に修正され，その結果信頼性が高いコンテンツが実現できていると主張するユーザも少なくない．これを裏付けるように，2005 年 12 月に公開された英 Nature 誌の調査によれば，Wikipedia は世界最大の百科事典のブリタニカと同等の規模と精度を持つと報告されている[8]．しかし，最近では各種の虚偽の書き込み事件をきっかけにその信頼性におけ

る潜在的な問題点が明らかにされてきた．例えば，ケネディ兄弟暗殺に元ジャーナリストの John Seigenthaler が関与しているという虚偽の書き込みが数か月放置されていた「ジョン・シーゲンソーラー・ウィキペディア経歴論争」や，ポッドキャスティングの先駆け的存在である Adam Curry が競合他社の情報を削除しようとしたことを告発された事例などが有名な例である．前者の例では，記事の掲載から数か月後に記事内容は修正されているが，その間，間違いが放置されていたままであった．追跡調査の結果，書き込みを行った当人はいたずら目的で記事を書き込んだことを謝罪するとともに，事件は解決の方向へ向かっていることが判明しているが，この事件は Wikipedia の信頼性に大きな疑問を投げかけた結果となった．

Wikipedia のページを変更するのは困難ではなく，Web ブラウザを通じて編集した後は保存するだけで修正された記事が WWW 上に公開される．この編集の容易さが Wikipedia の最大の特徴であり，昨今問題になっている点であるといえる．このような信頼性の問題に対し，Wikipedia ではユーザ認証や記事のプロテクト機能などいくつかの対策が講じられている．

ユーザ認証機能とは，ログインしていないユーザや，新規参入ユーザの記事への編集権限を制限する機能である．たとえば，匿名ユーザはディスカッションページにおいて記事内容に対する議論に参加することはできるが，編集はできない．この結果，悪意のあるユーザや広告目的のユーザによる情報操作を回避できると考えられている．

記事のプロテクト機能とは，ある程度コンテンツの内容が固まった場合に編集者が記事内容をそれ以上変更できないように記事内容を固定する機能である．プロテクトされた記事に対しては，匿名ユーザは編集ができず，記事を変更したい場合はその旨を編集者に依頼することしかできない．

しかし，ユーザアカウントの作成は容易であるため，登録ユーザによる情報操作などには比較的弱いといえる．そのため，利便性とコンテンツの信頼性は未だトレードオフの関係にあり，その解決は困難な問題として残っている．

3. Wikipedia マイニング

「Wikipedia マイニング」とは，筆者らの造語で，Wikipedia に対して Web マイニングを行うことで，有益な情報を抽出する手法の総称である．筆者らは，これまでの研究で Wikipedia が膨大なコンテンツ量を持っていながら，サイト内部で密なリンク構造ができていることに着目し，そのリンク構造を解析することで概念（記事）同士の関係を定義したシソーラス辞書を抽出できることを示した[13]．

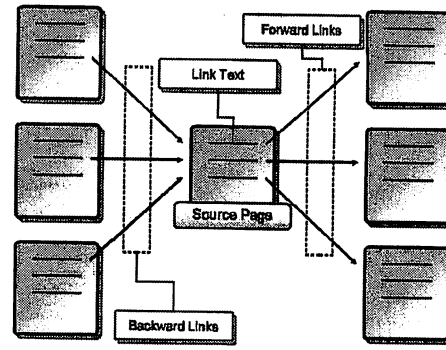

図 3　リンクの持つ情報
Fig. 3　Various information on hyperlinks.

ハイパーリンクは単にあるページから別のページへジャンプする「Forward リンク」の機能だけでなく，どのページからリンクされているかを示す「Backward リンク」やリンク先の情報の要約として利用可能な「リンクテキスト」など多くの情報（図 3）を含むことが知られている[4]．

リンクテキストも Web マイニングによるシソーラス辞書構築において重要な役割を果たす．リンクテキストとは，ハイパーリンク（A タグ）における内部テキスト部分を示す．例えば，以下のようなハイパーテキストを考えた場合，テキスト部分「Apple」がリンクテキストに相当する．リンクテキストは一般的に被リンクページの内容（要約）を表現していることが多い．

```
<a href="http://en.wikipedia.com/wiki/Apple_Computer">
Apple
</a>
```

また，サイトのトポロジ情報やトピックの局所性など，リンク構造には多くの有用な情報が含まれる．トピックの局所性とは，ハイパーリンクで繋がっているページ同士は，繋がっていないページ同士に比べて同じトピックに関する記述である場合が多いという性質である．Davison らの研究[6]は，このトピックの局所性が多くの場合に正しいことを示している．上記のような Web コーパスの特徴を活かし，リンク構造を解析することで，シソーラス辞書を自動的に生成する研究が最近注目を集めている．Web マイニングによるシソーラス辞書構築では，Web コンテンツの増加・更

新に従い，新しい語や他の語との関係などの情報を更新することができることが大きな特徴である．例えば，Chen ら[3] は，Web ページ同士のリンク構造を解析することで Web シソーラス辞書を自動的に構築する新しい手法を提案している．

しかし，前述のとおり Wikipedia においてはその信頼性が大きな問題となっている．そのため，信頼性を考慮した記事関係の抽出方法が必要となる．このような目的には，統計的な手法によって信頼性を確保する方法が有効だと考えられる．Google の PageRank アルゴリズム[11] や HITS[9] アルゴリズムなどでも被リンク数を統計的に解析することで客観的なデータを抽出することの重要性が示されている．

たとえば，虚偽の情報や関係ないリンクがドキュメント内に挿入されたと想定する．このとき，一つ一つの情報の真偽は判断することが極めて難しい一方で，複数の記事にわたって，リンク構造を解析することで客観的なデータを抽出できることが予想される．

本論文では，このような前提で Web 辞典のリンク構造を解析し，記事同士の関係性を抽出する問題を対象とし，いくつかの解析アルゴリズムを提案する．以降，各種法の詳細について述べる．

3.1 tfidf

tfidf は，自然言語で記述された文章の中から特徴的な語を抽出するアルゴリズムである．tfidf では，文章中に現れる語の頻度（tf: Term Frequency）と逆文書頻度（idf: Inversed Document Frequency）を利用した特徴語抽出アルゴリズムであり，以下の式によって算出される．

$$tfidf(t,d) = tf_{t,d} \cdot \log \frac{N}{df(t)}. \qquad (1)$$

$tf_{t,d}$ は，ドキュメント d における単語 t の出現頻度，N は総ドキュメント数，$df(t)$ は，単語 t が出現するドキュメント数を表す．これをリンク構造解析に適用するには，ドキュメント内の単語 t を Web ページ内のハイパーリンクと置き換え，特徴的な概念へのリンクを抽出するアルゴリズムを以下の式によって定義することができる．

$$tfidf(l,w) = tf_{l,w} \cdot \log \frac{N}{wf(l)}. \qquad (2)$$

$tf_{l,w}$ は，Web ページ w におけるリンク l の出現頻度，N は総ページ数数，$wf(l)$ は，リンク l が出現するページ数を表す．

3.2 lfibf

$lfibf$[14] は，筆者らの提案するリンク構造解析手法であり，グラフ $G = \{V, E\}$ 内において n ホップ以内のノード同士の関係性を数値化することを目的としている．このとき，2 記事間（v_i，v_j）の関係の強さを計測する問題を考えた場合，関係の強さは以下の二つの要素に依存すると考えられる．

- 記事 v_i から記事 v_j へのパスの多さ
- 記事 v_i から記事 v_j への最短距離

つまり，記事 v_i から記事 v_j へのパスが多ければ多いほど（共通のリンク先や共通の参照元が多いほど），記事間の関係性は強く，またそのパスの長さが短ければ短いほど強く関係すると考えられる．v_i から v_j への n ホップ先の全経路 $T = \{t_1, t_2, ..., t_n\}$ が与えられたとき，記事 v_i から記事 v_j の関係性 $lfibf$（Link Frequency Inversed Backward link Frequency）を以下の式により表現する．

$$lfibf(v_i, v_j) = \sum_{k=1}^{n} \frac{1}{d(|t_k|)} \cdot \log \frac{N}{bf(v_j)}. \qquad (3)$$

d は経路 t_k の経路長に応じて増加する関数であり，単調増加関数や指数関数を利用する．N は全記事数，$bf(v_j)$ は記事 v_j が持つ他の記事からのリンク数とする．つまり，$lf\text{-}ibf$ は多くのリンク先を共有するが，他の記事とはリンク先を共有しない記事により高い値を示す．また，同じ距離（例えば距離 1，直接リンク関係にある）の記事であっても，より多くリンク先を共有する記事に対して高い値を示す．

3.3 FB 法

Web 辞典のリンク構造を解析した結果，一般的な語，有名な事件，人気のなど，一部のページに極端に多くのリンクが集中する Zipf 分布に従うことが分かっている．これは，論文の参照頻度や Web ページの被リンク頻度などにもみられる分布であり，このような状況の下では，Backward リンクの多い記事に対する解析結果が収束しないことが予備実験によってわかっている．これは，Backward リンクの多い語は，記事内容が信頼できるため，Forward リンク解析の重みづけを重視することが望ましい一方で，様々な分野からリンクされているために Backward リンクの解析結果が収束しないことに起因する．そのため，筆者らは $lfibf$ の拡張手法として，記事の Backward リンクの数に応じて Forward リンク解析と Backward リンク解析の重みづけを変える Forward/Backward リンク

重みづけ法[14]（以降FB法）を提案した．FB法では，以下の式によりBackwardリンク解析の重みづけW_bとForwardリンク解析の重みづけW_fを行う．

$$W_b(|B_d|) = 0.5/(|B_d|^{\alpha}). \quad (4)$$
$$W_f(|B_d|) = 1 - W_b(|B_d|). \quad (5)$$

$W_b()$では，記事dの持つBackwardリンク数に応じてBackwardリンクの重みを変更する．αはパラメータであり，これまでの予備実験から，平均して数十から最大数百のリンクを持つWikipediaにおいては，0.05程度が妥当な値であるという知見が得られている．

3.4 リンク共起性解析

共起性解析は，任意の2つの語が同じ文章中に出現する頻度のことであり，自然言語処理や情報検索の分野で広くその有用性が実証されてきた[15]．通常のテキスト文章を語の集合としてとらえたとき，語と語の同時出現頻度を計測するのが共起性解析であるが，Web文章を関連記事へのリンク集合としてとらえたとき，リンクの共起性を計測することでリンク先の記事同士の関係性を抽出できることが予測できる．そのため，Web辞典からリンク共起性を示す関数lc（Link Cooccurrence）を以下のとおり提案する．

$$lc(l_1, l_2) = \frac{|D_{l_1} \cap D_{l_2}|}{|D_{l_1} \cup D_{l_2}|}. \quad (6)$$

l_1, l_2は任意のリンクを示し，D_{l_1}はリンクl_1が含まれるドキュメント集合を示す．ドキュメントによっては数千のリンクを持つものも存在するため，ドキュメント全体を解析対象としてリンクの共起性を計測するには，解析の際に多量の記憶容量を消費してしまう．そのため，解析対象のリンクを起点に，前後何個までのリンクとの共起性を解析するかを定めるウィンドウサイズを導入するのが一般的である．ここでは，ウィンドウサイズを5とし，前後5つのリンクペアに対して共起性を解析した．このウィンドウサイズは，解析用ワークステーションの限界値である．解析用ワークステーションの環境を表1に示す．

表1 シソーラス構築のための環境
Table 1 Environment for thesaurus construction.

項目	仕様
CPU	Pentium4 2.4 GHz × 2
メモリ	8,192 MB
OS	Solaris 10
開発言語	C++

4. 実験と考察

本章では，上記の各種アルゴリズムによって抽出されたシソーラス辞書を比較し，信頼性情報を考慮したシソーラス辞書構築手法の在り方を検討する．表2に各手法によって構築されたシソーラス辞書と関連語リストの例を示す．

tfidfをリンク解析に適用した場合，単にWikipedia全体で出現頻度の低いリンクが特徴的なリンクとして抽出される傾向にあった，特に致命的なのは，概念「Book」に対する「BookFinder.com」であった．これは，後の更新において記事の中からは削除されているが，一時的に広告目的で作成されたリンクであったと予想される．このように，tfidfは「他の記事に現れず，その記事の中で頻出の単語を抽出する」というアルゴリズムの特性上，虚偽の情報が他のドキュメントでは低頻度で出現するようなリンクを含んでいる場合，高信頼性を実現するためのフィルタとしては動作しないことがわかった．一方，$lfibf$の2手法ではnホップ先まで解析し，統計的に関連度を抽出することが有利に働き，ランキング上位には不適切と思われる記事は出現しなかった．また，$lfibf$とその拡張方式FB法を比較すると，FB法では一般的な語のときに，より関連する語が取得できていることがわかった．これは概念「Orange」の関連語抽出の際に顕著であった．$lfibf$では第三位に「United States」といったあまり関係ない記事が抽出されているが，FB法ではより関係の深い単語が抽出できていることが分かる．さらに，リンク共起性に基づく関係記事の抽出方法でも同様に，統計的に信頼性を解析することが有効に働き，より関係の深い語が抽出されていることが分かる．

規模の比較としては，$lfibf$では各語に対するスコアの高い順に30件抽出し，それぞれ約7,879万の関連語ペアとその関連強度が抽出できたが，リンク共起性解析では約2,293万，tfidfでは956万の関連語ペアしか抽出できなかった．そのため，高信頼性の記事関係を抽出するためには，$lfibf$かリンク共起性解析，もしくはその組み合わせなどが有効であると考えられる．

5. まとめと今後の展開

本論文では，Wikipediaのダイナミクスとコンテンツ信頼性の問題を分析し，各種のリンク構造解析手法を適用し比較することで，より信頼性の高い記事関係抽出アルゴリズムの検討を行った．実験の結果，提案

表 2 各手法により構築されたシソーラス辞書と関連語リスト
Table 2 Extracted association thesaurus by various methods.

Query Term	Method	Associated terms		
Microsoft	lfibf	Microsoft Windows	Operating system	Microsoft Office
	FB	Microsoft Windows	Microsoft Office	Operating system
	tfidf	Microsoft Dynamics NAV	Microsoft Zone	La morsure du dragon
	cooccurrence	Microsoft Windows	Linux	Operating system
Thomas Edison	lfibf	Incandescent light bulb	Kinetoscope	Edison, New Jersey
	FB	Incandescent light bulb	War of Currents	Mathew Evans
	tfidf	Port Huron, MI	Edison Mem. Tower and Mus.	Pearl Street Station
	cooccurrence	Phonograph	Transformer	Telegraph
Apple Computer	lfibf	Apple Macintosh	Macintosh software	Mac OS X
	FB	Apple Macintosh	Macintosh software	Mac OS X
	tfidf	Logic Express	IReview	Philip W. Schiller
	cooccurrence	Apple Macintosh	Adobe Systems	Mac OS X
Google	lfibf	Google search	Search engine	Usenet
	FB	Search engine	PageRank	Google search
	tfidf	Google Watch	Google platform	Google Catalogs
	cooccurrence	Spamming	AdWords	DejaNews
Horse	lfibf	Rodeo	Conditions races	Cowboy
	FB	Rodeo	Cowboy	Conditions races
	tfidf	Hemionids	3-Day Eventing	Hunter Pacing
	cooccurrence	Furlong	Dog	Cattle
Football	lfibf	FIFA	Penalty area (football)	football (soccer) clubs
	FB	football (soccer) clubs	Biography	England
	tfidf	Five a side football	Football (soccer) names	Soccer (disambiguation)
	cooccurrence	England football team	FA Cup	FA Premier League
Book	lfibf	Library	Diamond Sutra	Printing
	FB	Novel	Library	Author
	tfidf	NISO	Bookmarks	BookFinder.com
	cooccurrence	Novel	Author	Fiction
Orange	lfibf	Citrus	Metre	United States
	FB	Triple sec	Rutaceae	Pummelo
	tfidf	Beehives	Blood orange	Orange
	cooccurrence	Grapefruit	Strawberry	Rutaceae

手法である $lfibf$ とリンク共起性解析が精度よく，信頼性の高い情報を抽出できていることがわかった．

今後の展開としては，記事を投稿したユーザの信頼性（記事の書き込み頻度や他ユーザからの評価など）や，記事の更新頻度，参照頻度などを考慮した信頼性の抽出アルゴリズムなどが考えられる．

謝辞 本研究の一部は，文部科学省 21 世紀 COE プログラム「ネットワーク共生環境を築く情報技術の創出」，科学技術振興調整費「先端融合領域イノベーション創出拠点の形成：ゆらぎプロジェクト」，および文部科学省特定領域研究 (18049050) の研究助成によるものである．ここに記して謝意を表す．

参 考 文 献

1) E. Brill, "A Simple Rule-based Part of Speech Tagger," Proc. of Conference on Applied Computational Linguistics (ACL), pp. 112–116, 1992.
2) H. Chen, T. Yim and D. Fye, "Automatic Thesaurus Generation for an Electronic Community System," Journal of the American Society for Information Science, Vol. 46, No. 3, pp.175–193, 1995.
3) Z. Chen, S. Liu, L Wenyin, G. Pu and W. Y. Ma, "Building a Web Thesaurus from Web Link Structure," Proc. of the ACM SIGIR, pp.48–55, 2003.
4) N. Craswell, D. Hawking, and S. Robertson, "Effective Site Finding using Link Anchor Information," Proc. of the ACM SIGIR, pp.250–257, 2001.
5) C. J. Crouch, "A Cluster Based Approach to Thesaurus Construction," Proc. of the ACM SIGIR, pp.309–320, 1988.
6) B. D. Davison, "Topical Locality in the Web," Proc. of the ACM SIGIR, pp.272–279, 2000.
7) J. Dean and M. R. Henzinger, "Finding Related Pages in the World Wide Web," Proc. of the 8th International World Wide Web Conference, pp.1467–1479, 1999.
8) J. Giles, "Internet Encyclopaedias Go Head to

Head," Nature, Vol. 438, pp.900–901, 2005.
9) J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," Journal of the ACM, Vol. 46, No. 5, pp.604–632. 1999.
10) Bo Leuf and W. Cunningham, "The Wiki Way: Collaboration and Sharing on the Internet," Addison-Wesley, 2001.
11) P. Lawrence, B. Sergey, M. Rajeev, and W. Terry "The PageRank Citation Ranking: Bringing Order to the Web," Technical Report, Stanford Digital Library Technologies Project, 1999.
12) G. A. Miller, "WordNet: A Lexical Database for English," Communications of the ACM, Vol. 38, No. 11, pp.39–41, 1995.
13) 中山浩太郎, 原隆浩, 西尾章治郎, "Wikipedia マイニングによるシソーラス辞書の構築手法," 情報処理学会論文誌, Vol. 47, No. 10, 2006.
14) 中山浩太郎, 原隆浩, 西尾章治郎, "Web 辞典からのシソーラス辞書構築手法," DBWeb シンポジウム論文集, 2006.
15) H. Schutze and Jan O. Pedersen, "A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval," International Journal of Information Processing and Management, Vol. 33, No. 3, pp.307–318, 1997.
16) Wikimedia, "Wikipedia," http://www.wikipedia.org/
17) Y. H. Tseng, "Automatic Thesaurus Generation for Chinese Documents," Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, pp. 1130–1138, 2002.