# Wikipediaのリンク共起性解析によるシソーラス辞書構築

伊藤 雅弘† 中山 浩太郎†
原 隆浩† 西尾 章治郎†

近年，知識処理の有用なコーパスとして，ユーザ同士が協調してコンテンツを編集するWeb事典である「Wikipedia」に多大な注目が集まっている．筆者らはこれまでの研究において，Wikipediaに対してリンク構造を解析することで精度の良いシソーラス辞書が構築できることを示してきた．しかし，膨大な記事数を持つWikipediaを解析するためには，高い精度を保ったままスケーラビリティのさらなる向上が技術的な課題であった．そこで，本研究ではリンクの共起性解析に着目し，スケーラビリティの高いシソーラス辞書構築手法を提案する．提案手法の性能評価のために行った実験の結果，共起性解析を用いた手法は従来手法よりも少ない計算時間で，高精度なシソーラス辞書を構築できることを確認した．さらに，共起性解析とtfidfを融合させることによって，より高い精度が実現できることを確認した．

# A Thesaurus Construction Method Based on Link Co-occurrence Analysis for Wikipedia

Masahiro Ito,† Kotaro Nakayama,† Takahiro Hara† and Shojiro Nishio†

Wikipedia, a huge scale Web based encyclopedia, attracts great attention as a valuable corpus for knowledge extraction. We have already proved how effective it is to construct a Web thesaurus. However, we still need high scalability methods to analyze the huge amount of Web pages and hyper links among articles in the encyclopedias. In this paper, we propose a scalable Web thesaurus construction method from Wikipedia by using link co-occurrence. Experimental results show that the proposed method based on link co-occurrence analysis was better on scalability and accuracy than previous methods. Moreover, the method combining tfidf with link co-occurrence analysis brought higher precision.

## 1. はじめに

一般のキーワードによる情報検索システムでは，クエリ中のキーワードを直接含まない文書を検索することができない．そのため，クエリに含まれるキーワードに関連する単語を新たにクエリに付け加える，クエリ拡張という手法が研究されてきた．クエリ拡張を行うと，クエリとして指定されたキーワードを直接含まない文書であっても検索することができる．情報検索の研究分野では，クエリ拡張を実現する技術として，シソーラス辞書を使う手法があげられる．シソーラス辞書は，語彙どうしの関係を定義した辞書であり，関係性（is-a，part-ofなど）を明確に定義した「関連シソーラス」（Relation Thesaurus）と，与えられたキーワードから連想される語を抽出するための「連想シソーラス」（Association Thesaurus）に大別される．筆者らの研究グループでは，後者の連想シソーラスの構築に関する研究を進めてきた．連想シソーラスは，各概念をノード，関連度をエッジとする一種の重み付きグラフとして表現される．関連シソーラスのような階層構造（Hierarchy）ではなく，語と語の関係がネットワーク状に配置されており，与えられた語から関連する概念のリストを高速に抽出することが可能である．

一方，WWWの爆発的な普及にともない，Wikipediaに代表されるWeb事典が公開されてきた．Wikipediaは，Wikiを利用して構築された百科事典であり，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大きな特徴である．Wikipediaには，2006年9月の段階で約137万もの膨大な数の記事（英語のみ）が公開されており，市販の

† 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

百科事典の記事数が数万から 10 万であることと比較してもその規模が膨大であることが分かる．Nature 誌の調査によると，Wikipedia の記事数および精度は，多くの専門家が集まって作成した百科事典「Britannica」と同等であると報告している[8]．また，Wikipedia などの Web 事典と通常の電子事典の最大の違いは，記事（概念）どうしがハイパーリンクで互いに参照されていることである．ハイパーリンク（以降リンク）は，インターネット上のリソースを一意に表す URL によって，そのリソースへの参照を提供する．

筆者らは，Wikipedia のこれらの特性に着目し，Wikipedia に対して Web マイニングを行い，有益な情報を抽出する Wikipedia マイニングに関する研究を行ってきた．なお，Wikipedia マイニングは筆者らによる造語である．これまでの研究において，Wikipedia のリンク構造を解析することで，語彙どうしの関係を定義した連想シソーラス辞書を高精度で構築できることを示してきた[14),15)]．クエリ拡張の際に，この研究によって構築された大規模シソーラス辞書を利用することで，広い範囲の語彙をカバーすることが可能となる．

しかし，Wikipedia のように膨大な記事数を持つデータを解析するためには，スケーラビリティのさらなる向上が技術的な課題であった．文献 14)，15) の手法では，$n$ ホップ先までのリンク構造を解析し，語彙どうしの関連度を算出している．この手法では，日に日にその数が増加するような Wikipedia においては，計算量が大きくなり，即時的に処理できなくなる可能性がある．そこで，本研究ではリンクの共起性解析により，スケーラビリティの高いシソーラス辞書構築手法を提案する．

本論文の以下では，2 章で関連研究について述べ，3 章でリンクの共起性解析について解説する．4 章では実験により本手法を評価する．最後に，5 章でまとめと今後の展望を述べる．

## 2. 関 連 研 究

### 2.1 自然言語処理によるシソーラス辞書構築

自然言語処理によるシソーラス辞書構築の研究の歴史は古く，コーパス解析により（半）自動的に構築する手法は数多く提案されてきた．たとえば，語の共起関係に基づいて構築するもの[12)]や，語のフィルタリングやクラスタリング手法を用いる研究[2),4)]などがある．しかし，自然言語処理において，語義や係り受けなどの曖昧性および多義性の解消，同義語の同定などの諸問題はいまだ残っており，シソーラス辞書構築の精度低下の主要因となっている．

また，形態素解析の問題もある．自然言語処理によりシソーラス辞書を構築する場合，前処理として，入力文を意味のある最小の言語単位である形態素に分け，品詞タグを付与する必要がある．形態素解析および品詞タグを付与するツールとしては，Brill の Tagger[1)]が有名であるが，未知語への対応や曖昧性の取扱いなどが問題となっている．

### 2.2 Web サイトからのシソーラス辞書構築

Web コーパスと通常の文書コーパスの性質の最も大きな違いは，ハイパーリンクである．ここで，あるページを $p_i$ とすると，ページ $p_i$ が持つ他のページへのリンクをページ $p_i$ のフォワードリンク（Forward Link），ページ $p_i$ が持つ他のページからのリンクをバックワードリンク（Backward Link）と呼ぶ．リンクは，単に他のドキュメントへ移動するための機能を提供するだけでなく，トピックの局所性やリンクテキスト（Link Text）など重要な情報を豊富に有している．トピックの局所性とは，リンクでつながっているページどうしは，つながっていないページどうし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 の研究[5)]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．また，リンクテキストも Web サイトからの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，リンク（A タグ）における内部テキスト部分を指す．たとえば，以下のようなハイパーリンクを考えた場合，2 行目のテキスト部分「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>
```

ここでリンクテキスト，バックワードリンク，フォワードリンクの概念を**図 1** に示す．

上記のような Web コーパスの特徴を生かし，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構築する研究が最近注目を集めている[3)]．Web サイトからの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というのが大きな特徴である．

しかし，これらの手法は，解析対象とするコーパスに関する考察がなく，依然として自然言語処理を利用

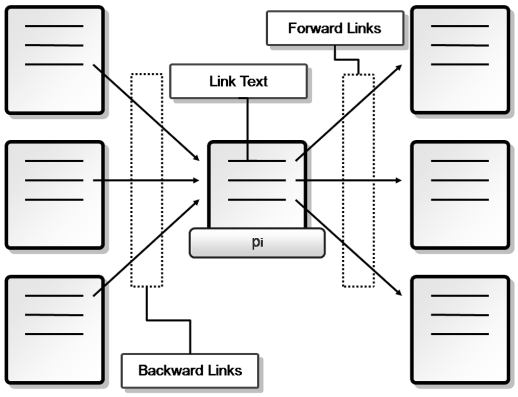


**図 1** リンクの概念
Fig. 1 Concept of links.

した解析による精度の問題などが残されている．また，膨大な Web 空間をコーパスとして用いた場合，探索空間が広すぎ解析内容が収束しない場合がある一方，ドメインを限定した場合には内容が偏るといった問題がある．

### 2.3 Wikipedia からのシソーラス辞書構築

Wikipedia では，Wiki によるコンテンツ管理を導入することにより，通常の Web コーパスや電子辞書とは異なる特徴を持つ．1 つ目は，ハイパーリンクによる記事どうしの参照である．各記事は，説明のテキスト，図表，そして別の記事に対する多数のリンクで構成される．従来の辞書や電子辞書では，機械可読なフォーマットで概念どうしの関係が表現されているものは少なく，概念どうしの関連を抽出するためには，説明文の中からさらに 1 度自然言語処理を行う必要があり，精度の低下を招く要因となっていた．しかし，Wikipedia の場合は Wiki をベースにしており，簡単に他の概念へのリンクを定義できることから，良質な概念どうしのリンクが多いという特徴を持つ．

2 つ目は，Wikipedia が高密度なリンク構造を持っていることである．筆者らは，予備実験として Wikipedia 内におけるリンク数をカウントしたところ，2006 年 9 月の段階で約 380 万ページ（Redirect リンクを含む）に約 8,000 万の内部リンク（Wikipedia 内へのリンク）を抽出し，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いるということを確認している．

3 つ目に，コンテンツの網羅性がある．従来，WWW を自然言語処理のコーパスとして利用する場合，その探索空間が膨大になりすぎることから，解析内容が収束しないもしくは偏ってしまうという問題があった．これに対し，Wikipedia は最新の幅広い分野の記事が網羅されており膨大な量のコンテンツが存在するものの，WWW の探索空間に比較するとそのリンク構造はサイト内で閉じられているため，現実的な時間での解析が可能である．

4 つ目に，URL によって概念を一意に特定できるということである．自然言語処理においては，様々な局面で未知語の問題に突き当たる．たとえば，形態素解析において未知語が存在すると，解析の精度が大きく下がることは周知である．また，「Apple」のような果物や社名など様々な意味が存在する多義語において，自然言語処理でそれぞれの意味を判別するには，前後の文脈で判断するなど高度な解析技術が必要であり，困難である．しかし，Wikipedia では URL によって一意に示される 1 つの記事（ページ）が 1 つの単語（概念）を表しており，多義を持つ単語には，意味に応じて別々の記事が用意されている．そのため，形態素解析における未知語の弊害もなく，多義語の判別も不要であり，高精度なシソーラス辞書の構築が可能である．

以上のような理由から，Wikipedia をシソーラス辞書構築のコーパスとすることは，多くのメリットがある．Wikipedia を解析してシソーラス辞書を構築する先行研究として，tfidf を使った手法と lfibf を使った手法がある．以下にその手法を解説する．

#### 2.3.1 tfidf

tfidf[11] は，Salton らによる文書中の重要なキーワードを抽出するための手法である．tfidf は tf（Term Frequency）と idf（Inverse Document Frequency）の 2 つの指標を利用し，それらの積によって文書中の各語の重要度を計算する．tf は文書中における特定の語の出現頻度であり，文書中に多く含まれる語が特徴語とされる．idf は全文書中に，特定の語が出現する文書数の逆数であり，出現する文書数が多い語は idf の値が小さくなる．つまり，広く使われている一般的な語ほど特徴語としての重要度が低くなる．

この tfidf を Wikipedia に適用した例として，Gabrilovich らによる研究[7] がある．Wikipedia においては，一ページが一概念（語）に対応し，リンクは他の概念に対する意味的かつ明示的な関係を示す．そのため，tfidf でページ内の重要なリンクを抽出することで語どうしの関係性を抽出することができる．tfidf によって記事中の各リンクの重要度を以下の式によって与える．

$$tfidf(l,d) = tf(l,d) \cdot idf(l) \quad (1)$$

$$idf(l) = \log \frac{N}{df(l)} \quad (2)$$

ここで，$tf(l,d)$ は記事 $d$ におけるリンク $l$ の出現回数であり，$df(l)$ はリンク $l$ を含む記事数，$N$ は全記事数である．

そして各概念をベクトル空間モデル[10] によって，リンクを次元，その各リンクの重要度（重み）を要素としたベクトルを生成する．各概念の関連度の算出は，それらのベクトル間のコサイン相関によって求められる．

この手法では，1 つの概念の特徴ベクトルを抽出するには 1 つの記事に存在するリンク情報だけを解析すればよいため，スケーラビリティは高い．しかし，それゆえに記事の内容に信頼性がない場合やリンク数が少ない場合に，精度が低下する．

#### 2.3.2 lfibf

lfibf[14),15)] は，ある記事 $v_i$ から $v_j$ の関連度を算出する手法である．lfibf は lf（Link Frequency）と ibf（Inverse Backward link Frequency）の 2 つの指標を利用し，それらの積によって関連度を算出する．lf は記事 $v_i$ から $v_j$ へのパスの多さと，各パスの短さによって決定され，全経路 $T = \{t_1, t_2, \cdots, t_n\}$ によって以下の式で表される．

$$lf(v_i, v_j) = \sum_{k=1}^{n} \frac{1}{d(|t_k|)} \tag{3}$$

ここで，$d$ 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ことができる．

ibf は全記事中の記事 $v_j$ が参照された数，つまり記事 $v_j$ が持つ Backward リンク数の逆数である．この指標は，記事 $v_j$ に対するリンクが多いほど小さい値になる．したがって，記事 $v_i$ から記事 $v_j$ への関連度は lfibf によって以下の式で与えられる．

$$lfibf(v_i, v_j) = lf(v_i, v_j) \cdot ibf(v_j) \tag{4}$$

$$ibf(v_j) = \log \frac{N}{df(v_j)} \tag{5}$$

$N$ は全記事数，$df(v_j)$ 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，lfibf に基づく記事 $v_i$ から $v_j$ への関連度は，$v_i$ から $v_j$ へ多くの短いパスを持ち，$v_j$ の Backward リンク数が少ない場合に高い値を示す．

この手法では，$n$ ホップ先までのリンク構造を解析し，語彙どうしの関連度を算出している．そのため，1 つの概念に対する計算量が多く，全体として多量の計算が必要になる．

## 3. リンクの共起性解析

前章の 2.3 節にあげたように，Wikipedia からシソーラス辞書を構築する際，従来手法では概念の特徴を，局所的情報をもとにしたり，$n$ ホップ先までのリンク構造を解析することによって抽出しているため，精度が低下したり，多量の計算が必要であるという問題が存在した．そこで著者らは，リンクの共起性に着目した．Wikipedia 全体を通したリンクの共起性は，tfidf のような局所的情報ではなく大域的統計情報であり，ある特定の記事の質に大きく左右されることはない．その理由は，tfidf の場合は特定の記事に関する特徴を求めるとき，その記事内の情報を用いるため，特定の記事の質に大きく左右されるが，Wikipedia 全体における特定の記事へのリンクの共起性を解析した場合，その特定の記事の特徴は Wikipedia 全体の統計的情報に基づくからである．また，その計算時間はデータ量に対して線形であるため，lfibf のように多量の計算が必要になることはない．

ここで，提案手法において参照先 URL が同じリンクは，たとえリンクテキストや出現する記事が違っても，同じリンクであると見なす．たとえば**図 2** の例では，リンクテキスト “MS” と “Microsoft” は同じ記事 “Microsoft” を参照しており，同一のリンク（記事 “Microsoft” へのリンク）であると見なす．つまり，ここでのリンクは Wikipedia の各記事と 1 対 1 に対応しており，提案手法では特定の記事への参照に関する統計的解析を行うものである．たとえば，リンク A の Wikipedia 全体での出現回数は，リンク A が指し示す記事 A の被参照回数であり，記事 A の Backward リンク数と等しくなる．

本章では，Wikipedia におけるリンクの共起性に基

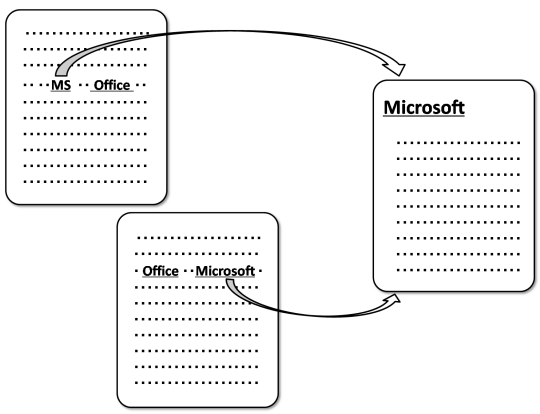


**図 2** 同一リンクの定義
Fig. 2 Definition of identical links.

づいて，2 つの概念間の関連度を求める手法を提案する．以下では，まず従来研究における単語の共起性によって関連度を求める手法を解説した後，提案手法におけるリンク間の関連度の算出方法を述べる．

### 3.1 単語の共起性解析による関連度の算出

単語の共起とは，特定の範囲において，ある組の単語が同時に出現することであり，単語の共起性解析は，頻繁に共起する単語ペアは関連度が高いという考えに基づいている．単語の共起性解析は，従来研究においては連想シソーラス辞書構築に利用されてきた．ここでは，単語の共起性を解析することによって単語ペアの関連性を求める手法についての 2 つの先行研究を紹介する．

#### 3.1.1 共起回数による単語間の関連度

共起回数から関連度を求める代表的な手法として，Cosine，相互情報量，Dice 係数がある[9),16)]．以下では，それぞれにおける関連度の計算式を示す．$P(x)$，$P(y)$ は単語 $x$ と $y$ がそれぞれ独立に出現する確率，$P(x,y)$ は $x$ と $y$ が同時に出現する確率，$f_x$，$f_y$ は $x$ と $y$ がそれぞれ独立に出現する回数，$f_{xy}$ は $x$ と $y$ が同時に出現する回数とする．

- Cosine

$$Cosine(x,y) = \frac{f_{xy}}{\sqrt{f_x f_y}} \tag{6}$$

- 相互情報量

$$MI(x,y) = \log \frac{P(x,y)}{P(x)P(y)} \tag{7}$$

- Dice 係数

$$Dice(x,y) = \frac{2 \cdot f_{xy}}{f_x + f_y}, \quad (0 \leq Dice(x,y) \leq 1) \tag{8}$$

北村らは，Dice 係数の欠点は単語ペア出現回数の大小にかかわらず，独立出現回数と同時出現回数の相対比により関連度が決まるという点であり，出現回数が少ない場合の信頼性の違いを考慮していないと指摘し，Dice 係数に共起回数による重み付けを行った改良版の Dice 係数を提案している[16)]．以下にその式を示す．

$$IDice(x,y) = w(f_{xy}) \frac{2 \cdot f_{xy}}{f_x + f_y}, \tag{9}$$

$$w(f_{xy}) = \begin{cases} f_{xy} \\ \log f_{xy} \end{cases}$$

#### 3.1.2 二次共起

Schütze らは文書コーパスから単語の共起に基づくシソーラス辞書を構築し，情報検索に応用する手法を提案している[12)]．具体的には，3.1.1 項に示すような共起回数だけで，ある組の関連度を算出する一次共起（first-order co-occurrence）に異論を唱え，ある組の語がどれくらい同じ語と共起しているかで関連度を算出する二次共起（second-order co-occurrence）を提案している．

この手法では，まずすべての単語を行と列においた正方行列 $C$ を作り，その各要素 $c_{ij}$ を単語 $i$ と $j$ の共起回数としている．ここで任意の単語 $i$ における行ベクトルをシソーラスベクトル（thesaurus vector）とし，単語 $i$ と $j$ の関連度はそれぞれのシソーラスベクトルのコサイン相関によって求められる．

### 3.2 リンク間の関連度の算出

先に述べたように，筆者らの提案手法では，リンクの共起性を解析することによってリンク間（記事間）の関連度を算出する．リンクの共起とは，単語をリンクとして扱うということ以外，基本的な概念は単語の共起と同様である．つまり，リンクが共起するということは，特定の範囲においてある異なる 2 つのリンクが同時に出現するということである．リンクの共起性解析では，リンクは参照先 URL が同じなら同じリンクと見なされ，Wikipedia 全体でのリンクの共起性を解析する．ここで，先に述べたように，Wikipedia におけるリンクは，参照先の記事を 1 対 1 で表している．そのため，2 つのリンクの関連度を求めることは，Wikipedia の記事が表す 2 つの語（概念）の関連度を求めることと等価である．

ところで，Wikipedia を解析するとき，同じ記事内での共起をカウントすると，リンク数の多い記事の場合，非常に膨大な共起の組合せが存在する．そこで解析範囲を近傍のリンクに限定するウインドウを設定して，ウインドウ内のリンクのみにおいてだけ共起していると見なす[12)]．たとえば **図 3** の，ある “○○○” という語に関する記事（図上部）から，解析対象データであるリンクを出現順で並べたデータ（図下部）を生成した例を題材に説明する．図中のアルファベットは，その記事のリンクにおけるリンク先記事を表している．この例では，ウインドウサイズが 3 の場合の解析例を示しており，解析対象データの先頭からウインドウが図に示すように 1 つずつ移動する．各ウインドウの位置で，図に示すような 2 つのリンクペアが共起していると見なされる．このような指定されたサイズのウインドウに基づいて計算された各記事のリンクペアの共起回数を，全記事で合算することによって，Wikipedia 全体におけるそれぞれのリンクペアの共起回数を算出することができる．

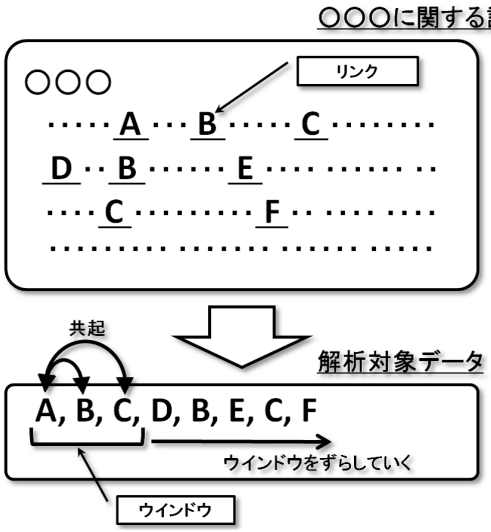


**図 3** 記事の解析例
Fig. 3 An example of co-occurrence analysis.

ここで提案手法では，リンク間の関連度を算出するためにリンクの二次共起による関連度を用いる．これは，リンクの一次共起のみを用いた場合，直接共起しないリンクペアは関連性がないと判断されるが，本来関連性はあるが直接共起しないようなリンクペアも存在するためである．一方，リンクの二次共起はこの問題を解決するために，たとえ直接共起しなくても，同じ共起特性を持つかどうかで関連性を測る．実際，予備実験により，高い関連度であると考えられるリンクペアのうち，リンクの一次共起では関連性がないとされる場合でも，リンクの二次共起では高い関連度を示すものがいくつも存在することを確認している．さらに，数百のリンクペアの関連度を求めたところ，リンクの一次共起だけでは約 50%ものリンクペアにおいて関連性がないと算出された．これは，そのリンクペアが 1 度も直接共起していないということを表している．一方，リンクの二次共起ではすべてのリンクペアの関連度を求めることができていた．たとえば，2 つの語「OPEC」と「Oil」は，直感的にも高い関連性を持つことが分かる．しかし，一次共起による解析の場合，関連性はないと算出されている．一方，二次共起の場合はこの 2 つの語に対して比較的高い関連性を示していた．以上の理由により，リンクの共起性解析をする際，一次共起よりも二次共起の方が，より関連度を求めるために適していると判断した．

二次共起では，まずリンクの一次共起による関連度を求めた後，その関連度を使ってリンクの二次共起による関連度を算出する．以下では，それぞれについて解説する．

#### 3.2.1 リンクの一次共起による関連度の算出

一次共起による関連度の算出方法として，最も単純なものが共起回数を共起性として利用する方法である．しかし共起回数だけを利用した場合，多く出現しているリンクは出現回数の低いリンクより，どのリンクとも共起する可能性が高くなる．つまり，出現回数の高いリンクほど関連度が高くなる可能性がある．たとえば，あるリンク A と B がそれぞれ 1,000 回出現していて，A と B の共起回数が 100 回であるのと，あるリンク C と D がそれぞれ 100 回出現していて，C と D の共起回数が 100 回であるのとでは同じ関連度であると見なされる．しかしこの場合，リンク C と D はすべての出現において共起しているので，C と D の関連度の方が高いことは明白である．この問題を解決するために考慮しなければならないことは，共起ペアのそれぞれの出現回数に対して共起回数が何回であるかということである．そこで，リンクの出現回数を考慮した計算方法として 3.1.1 項にあげた 4 つの式をリンク間の一次共起による関連度として定義する．ここで，本研究ではリンクの共起による解析を行うため，語の共起性解析で示した 4 つの式の中に出現する $x$, $y$ を単語ではなくリンクとして扱う．

#### 3.2.2 リンクの二次共起による関連度の算出

二次共起による関連度を求める際，各リンクにおいてどのようなリンクと共起するかという，各リンクの共起特性を表すリンクベクトルを生成する．リンクベクトルは，ベクトル空間モデルに基づく，リンクを次元，各リンクに対する重み（一次共起による関連度）を要素とする多次元ベクトルであり，リンク $i$ の共起特性を表すベクトル $v_i$ は以下のように表される．

$$v_i = \{l_{i1}, l_{i2}, l_{i3}, \cdots, l_{in}\} \tag{10}$$

ここで，$l_{ij}$ はリンク $i$，$j$ 間の重みである．

このように作成されたリンクベクトルを利用し，以下の式 (11) で 2 ベクトル間のコサイン相関によって，それぞれのリンクの共起性パターンがどれだけ同じかという関連度を求めることができる．

$$\begin{aligned} cos(v_i, v_j) &= \frac{v_i \cdot v_j}{|v_i||v_j|} \\ &= \frac{\sum_{k=1}^{n} l_{ik} l_{jk}}{\sqrt{\sum_{k=1}^{n} l_{ik}^2} \sqrt{\sum_{k=1}^{n} l_{jk}^2}} \end{aligned} \tag{11}$$

#### 3.2.3 tfidf との融合

前項における共起性解析によるリンクベクトルは，Wikipedia 全体に散在するリンクの共起性によって生成されているが，各記事内の記述は直接利用していない．共起性解析から得られるような大域的統計情報は，

特定の記事の質に大きく左右されることがない反面，全体的なリンクの使われ方の傾向を用いているため，記事に書かれている重要なキーワードを見落としている可能性がある．

そこで，各記事内における記述（リンク）の特性も考慮するため，2.3.1 項で述べた tfidf によるベクトルを合成することによって，精度向上を図る．共起性解析と tfidf によるベクトルは，両方ともベクトル空間モデルによって，リンクを次元，その各リンクに対する重みを要素とする多次元ベクトルで表されているため，一般的なベクトルに対する演算を適用できる．それぞれのベクトルは，合成（加算）する前に各要素をすべての要素の合計値で除算することによって正規化する．

## 4. 実　　験

本章では，提案手法の有効性を示すために行った実験について述べる．

### 4.1 実験概要

本実験の目的は，提案手法によって構築されたシソーラス辞書の精度と構築時間を評価することである．提案手法は，共起性解析だけによる手法，共起性解析から生成したベクトルと tfidf のベクトルを融合する手法の 2 つである．比較対象としては，tfidf と lfibf を用い，それぞれの手法によって Wikipedia からシソーラス辞書を構築し，構築時間と精度を比較した．なお，lfibf においてはホップ数 $n$ を 2 に設定した．

解析対象の Wikipedia のデータとしては，2006 年 9 月時点の英語版 Wikipedia のデータからノイズ記事を除去した，記事数約 82 万，総リンク数約 4,000 万のデータを用いた．ノイズ記事の定義は，トップページやカテゴリページなどの通常の記事ではないもの，記事内のリンク数が 5 つ以下のものである．また精度については次に述べるデータセットに基づいて計測した．

#### 4.1.1 精度評価用の実験データセット

本実験では，データセットとしてシソーラス辞書の精度を計測するためによく利用[7),13)]されている「WordSimilarity-353 Test Collection」[6)]によって評価を行った．このデータセットは，353 組の単語ペアを 13 人から 16 人の被験者によって関連性を主観で 10 段階評価してもらい，その平均を関連度としている．このデータセット内のすべての単語ペアに対する関連度をシソーラス辞書から抽出し，データセットでの関連度と，実験によって抽出した関連度をそれぞれ順位付けする．そして，2 つの順位の相関性を「スピアマンの順位相関係数（Spearman rank-order correlation coefficient）」によって求め，シソーラス辞書の精度とした．スピアマンの順位相関係数とは，2 つの数値の系列における数値の大きさの順位が，どの程度近いかという順位の相関性を計算する手法であり，順位がまったく同じならスピアマンの順位相関係数は最大の相関である 1 となり，順位がまったく逆なら負の相関である $-1$ となる．

ここで，「WordSimilarity-353 Test Collection」では単語のペアが与えられているが，この単語を Wikipedia のページにマッピングしなければならない．文献 7)，13) では，そのことに対して言及していないが，本研究では以下の手順でマッピングを行った．

(1) データセットに存在する各単語に対して，各単語がリンクテキストとして利用されている Wikipeda の記事を割り当てる．しかし一般に，あるリンクテキストを用いたリンクによって参照される記事は複数ある．そこで，あるリンクテキストによって参照された記事の中で，最も被参照数の多い記事をそのリンクテキストが表す記事とする．

(2) しかし，(1) の処理だけでは，データセットで想定されていた単語と違う意味の記事にリンクが割り当てられており，比較には適さない単語ペアが存在する．たとえば，“Aluminium” と “Metal” の比較で，鉄に関する比較をしているにもかかわらず，“Metal” が音楽のジャンルの記事になっている場合である．そこで，多義性の高い単語を含む単語ペアから優先的に，一般的に比較する際に適切だと思われるリンクに手動で置き換える．

(3) 最後に，Backward リンク数 500 以下の単語を含む組を除外する．

この処理の結果，テストデータに残ったのは 100 組であった．ここで，Backward リンク数 500 以下を除外する理由は，十分な情報がないリンクは正確に関連度を測定できないためである．予備実験によって，Backward リンク数 100 以下では精度がすべての手法において急激に低下することが分かっている．また，Backward リンク数 100 から 500 にかけて徐々に提案手法の優位性が顕著になっていた．そのため本実験では，高い精度を示し，なおかつある程度の実験データ数を確保できる，Backward リンク数 500 を閾値とした．ここで，Wikipedia は現在もそのリンク数を増加させ続けており，将来的には非常に膨大なリンク数を持つようになると予想される．つまり，Backward リ

**表 1** 計算機環境
Table 1 Computer environment.

| 項目 | 仕様 |
|---|---|
| CPU | Intel Xeon 5160 3.0 GHz × 4 |
| メモリ | 16 GB |
| OS | SUSE Linux Enterprise Server 10 |
| 開発言語 | C++ |
| コンパイラ | Intel C++ Compiler 9.1 |

ンク数の多い記事数が増加し，共起性解析によってより精度の高い概念間の関係性を抽出できるようになると考えられる．

### 4.2 実験結果と考察

本節では，本実験の結果における計算時間と精度のそれぞれについて解説し考察する．シソーラス辞書の構築には，**表 1** に示す計算機環境を用いた．

#### 4.2.1 シソーラス辞書構築に要する時間

**表 2** に，提案手法におけるウインドウサイズ 2 から 5 の場合と，比較手法におけるシソーラス辞書構築に要する時間を示す．なお，提案手法における一次共起の各計算手法によって計算時間に変化がなかったため，ここでは Cosine を用いたシソーラス辞書の構築時間を示す．

提案手法の共起性解析ではウインドウサイズが 2 から 5 に増えるにともなって計算時間が増加しているが，その増加率はウインドウサイズの増加に対して 160 秒程度と線形となっている．

共起性解析と tfidf の計算時間を比較すると，共起性解析のウインドウサイズ 2 においては tfidf と比べて約 0.8 倍の時間を，ウインドウサイズ 5 においては約 2.5 倍の時間を要している．

次に，共起性解析と lfibf の計算時間を比較すると，lfibf は共起性解析のウインドウサイズ 2 に対して約 390 倍もの時間を要している．ウインドウサイズ 5 の処理と比較しても約 120 倍の計算時間を要している．これは，明らかに提案手法の方が計算時間において大幅に有利であることを示している．lfibf は手法の特性上，$n$ ホップ先のリンクを再帰的に計算する．lfibf に関する論文[15] に述べられている近似手法を用いても，多量の計算が必要である．一方，共起性解析や tfidf はリンク先を再帰的に処理することはしないため，少ない計算量に抑えられている．

また，共起性解析と tfidf を融合した手法の計算時間は，共起性解析と tfidf の計算時間を足した時間とほぼ一致した．

ここで，本実験により生成されたリンクベクトルの数と次元数は，実験に用いた Wikipedia の記事数 816,463 と等しくなる．しかし，実際には長さがゼロのベクトルや，ゼロ要素を含むベクトルは圧縮しており，現実的な計算機リソース上での計算を可能にしている．圧縮後のデータ量としては，ウインドウサイズ 2 の場合，ベクトル長がゼロ以外のリンクベクトルの数は 782,417 で，ベクトルの非ゼロ要素の最大数は 62,830 次元，平均長は約 33 次元であった．また，ウインドウサイズ 5 の場合，ベクトル長がゼロ以外のリンクベクトルの数は 782,417 で，ベクトルの非ゼロ要素の最大数は 162,431 次元，平均長は約 114 次元であった．

**表 2** シソーラス辞書構築に要する時間
Table 2 Total analysis time.

| 手法 | 計算時間（秒） |
|---|---|
| 共起性解析のみ（ウインドウサイズ 2） | 220 |
| 共起性解析のみ（ウインドウサイズ 3） | 385 |
| 共起性解析のみ（ウインドウサイズ 4） | 545 |
| 共起性解析のみ（ウインドウサイズ 5） | 701 |
| tfidf と融合（ウインドウサイズ 2） | 500 |
| tfidf と融合（ウインドウサイズ 3） | 664 |
| tfidf と融合（ウインドウサイズ 4） | 825 |
| tfidf と融合（ウインドウサイズ 5） | 980 |
| tfidf | 278 |
| lfibf | 85,472 |

**表 3** シソーラス辞書の精度：提案手法
Table 3 Accuracy of thesaurus: Proposed method.

| ウインドウサイズ | 手法 | スピアマンの順位相関係数 | |
|---|---|---|---|
| | | 共起性解析のみ | tfidf と融合 |
| 2 | Cosine | 0.65 | 0.68 |
| | MI | 0.56 | 0.60 |
| | Dice | 0.59 | 0.69 |
| | IDice | 0.60 | 0.66 |
| 3 | Cosine | 0.62 | 0.66 |
| | MI | 0.62 | 0.60 |
| | Dice | 0.59 | 0.68 |
| | IDice | 0.58 | 0.65 |
| 4 | Cosine | 0.62 | 0.65 |
| | MI | 0.61 | 0.59 |
| | Dice | 0.58 | 0.67 |
| | IDice | 0.59 | 0.66 |
| 5 | Cosine | 0.62 | 0.64 |
| | MI | 0.59 | 0.59 |
| | Dice | 0.58 | 0.66 |
| | IDice | 0.60 | 0.67 |

#### 4.2.2 シソーラス辞書の精度

2 つの提案手法（共起性解析のみと共起性解析と tfidf の融合）によって構築したシソーラス辞書の精度として，ウインドウサイズ 2 から 5 のそれぞれにおいて，3.1.1 項で示した 4 つの一次共起の計算手法を用いた場合の結果を**表 3** に示す．また，**表 4** に比較手法におけるシソーラス辞書の精度を示す．

まず表 3 の共起性解析のみの場合の結果より，精度

**表 4** シソーラス辞書の精度：比較手法
Table 4 Accuracy of thesaurus: Other methods.

| 手法 | スピアマンの順位相関係数 |
|---|---|
| tfidf | 0.57 |
| lfibf | 0.68 |

はウインドウサイズの違いで変化が見られた．総じてウインドウサイズが小さい方が高精度となっており，ウインドウサイズが 2 の結果が最も良い．これは，リンクの共起性解析においては隣り合うリンクを共起と見なすだけで十分であり，隣り合っていない離れたリンクを共起と見なすことは，精度の低下を招くということを示唆している．また，各一次共起の計算手法による精度の違いを比較すると，どのウインドウサイズにおいても Cosine が最も高い精度を示し，他の手法による精度の違いはほとんど見られなかった．結果的に，ウインドウサイズが 2 における Cosine による共起性解析が最も精度が高い結果となった．

次に表 4 の tfidf と比較すると，すべてのウインドウサイズとすべての手法において，tfidf より高い精度を実現している．これは，tfidf では記事内に含まれるリンクのみを利用し，記事（概念）に対する特徴ベクトルを抽出しているのに対して，共起性解析では Wikipedia に存在するすべての記事を通して共起しているリンク組を抽出していることに起因する．各記事は限られたユーザによって編集されているので，各記事のリンク数や信頼性は均質でない．そのため，各記事のリンクによって得られる情報は必ずしも一般的というわけではなく，偏った内容となっている可能性がある．しかし，Wikipedia のすべての記事を通して出現する各記事へのリンクから得られる情報は，一部のユーザによる偏った情報ではなく，各記事に対する一般的な認識による情報となっている．つまり，ある記事へリンク付けを行うかどうかなどの，書き手によるリンク付けの偏りが存在した場合においても，共起性解析は様々な書き手を通しての統計的情報を得られるため，書き手によるリンク付けの偏りは平均化され，客観的情報が得られる．この理由により，tfidf より共起性解析の方が高い精度を実現したものと考えられる．

また表 4 の lfibf と共起性解析を比較すると，共起性解析は lfibf に比べて低い精度となっている．しかし，共起性解析における最大の精度である 0.65 は，lfibf の精度である 0.68 に迫っており，わずかな差にとどまっている．前項で述べた共起性解析と lfibf の計算時間が約 390 倍も違うことを考えると，共起性解析は大幅に少ない計算量で lfibf と同等の高い精度を実現しているといえる．

さらに表 3 の共起性解析と tfidf を融合した手法の結果は，共起性解析のみの手法と比較して，ほとんどの場合で精度が向上している．特に，Dice 係数を用いた手法の精度向上が著しい．tfidf と比較して精度が良いのはもちろんであるが，lfibf と比較しても同等かそれ以上の精度を実現している．精度が向上した理由として，共起性解析の得意とする性質と，tfidf が得意とする性質の違いが影響していると考えられる．すでに述べたが，共起性解析では Wikipedia 全体を通しての統計的情報を用いるため，特定の記事の質によらない一般的な語の使われ方による情報から関連性を導いている．一方，tfidf は記事の記述内容からその記事（語）の特徴を求め，その特徴を比較することによって関連性を導いている．tfidf の場合，精度は記事の質に左右されるが，その記事の特徴を端的に表すものとして重要な情報となりうる．この情報が，共起性解析での統計的情報では特徴を表現しきれない場合に，特徴情報を補間する役割をしていると考えられる．仮に，tfidf による記事の特徴ベクトル中に，実際にその記事の特徴を表すデータのほかに，ノイズデータが含まれているとしても，共起性解析によるベクトルと融合した場合，ベクトルの次元数は非常に大きいため，ノイズデータの次元が共起性解析と tfidf のベクトルで偶然共通し，合成ベクトルにおいて大きな値になる確率はきわめて低いと考えられる．一方で，本来重要な特徴であるはずの次元のデータは双方で共通する可能性が高く，ベクトルの合成によってより強調されると考えられる．また，共起性解析において Backward リンク数が少なく統計的情報が十分に取得できなかった場合でも，tfidf の特徴情報によって補完された合成ベクトルを生成するため，精度が向上するものと考えられる．

ここで，本実験によって得られた実験結果の精度が，構築したシソーラス辞書全体においても有効であるかについて考察する．本実験で用いた 100 組の実験データセット中からランダムに数十組を抽出し，そのデータセットを用いて実験を 50 回行った．その結果，精度は多少変化するものの，提案手法と従来手法の精度の順位は 3 回しか変化せず，ほぼ同じ傾向を示した．このことより，構築したシソラース辞書全体においてもこのような傾向が見られるものと考えられる．

## 5. おわりに

本論文では，大規模な Web 事典である Wikipedia を解析し，シソーラス辞書を構築するスケーラビリティの高い手法として，リンクの共起性解析に基づく

手法を提案した．実験の結果から，共起性解析によって構築されたシソーラス辞書は，従来研究である tfidf よりも高い精度を実現し，また lfibf よりも計算時間が大幅に短いにもかかわらず，lfibf に迫る高い精度を保っていることが分かった．特に，一次共起性の計算手法としては，Cosine が最も高い精度でシソーラス辞書を構築できることが分かった．さらに，共起性解析と tfidf を融合した手法ではさらに精度が向上し，lfibf と同等かそれ以上の高い精度を実現できることが分かった．

今後の展開としては，リンクの記事内における出現位置や，共起するリンクが同一センテンス内であるかなどの位置関係も考慮した重み付けを行い，精度向上を目指す．また，記事内のリンクの近さによって共起を定義付けるだけではなく，Backward リンクのリンク元などとの関係も共起と定義するなど，共起情報の量や網羅性の向上などを図り，さらなる精度向上を目指す．共起性解析と tfidf の融合手法においては，本論文のような単に正規化したベクトルの和をとるだけではなく，それぞれのベクトルに対して何らかの特徴に応じた重み付けを行うことによって，より精度の高い合成ベクトルの生成を目指す．

また，提案手法は Wikipedia だけでなく，多数の内部リンクを持つ他の Web サイトにおいても適用可能だと考えられる．たとえば，Wikipedia の姉妹プロジェクトである Wikinews などがその適用対象である．ただし，これらのプロジェクトは，現時点では提案手法を適用できるほどの十分な量のデータがない状態である．

さらに，自然言語処理技術の適用も課題の 1 つである．リンクの前後の文章を構文解析することで，関連度だけでなく，関連の種類（is-a や part-of）の抽出も可能であると考えられる．

**謝辞** 本研究の一部は，文部科学省特定領域研究（18049050）およびマイクロソフト産学連携研究機構 CORE 連携研究プロジェクトの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) Brill, E.: A Simple Rule-Based Part of Speech Tagger, *Proc. Applied Natural Language Processing*, pp.152–155 (1992).
2) Chen, H., Yim, T., Fye, D. and Schatz, B.R.: 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) Chen, Z., Liu, S., Wenyin, L., Pu, G. and Ma, W.-Y.: Building a Web Thesaurus From Web Link Structure, *Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.48–55 (2003).
4) Crouch, C.J.: A Cluster-Based Approach to Thesaurus Construction, *Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.309–320 (1988).
5) Davison, B.D.: Topical locality in the Web, *Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.272–279 (2000).
6) Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E.: WordSimilarity-353 Test Collection (2002).
7) Gabrilovich, E. and Markovitch, S.: Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis, *Proc. International Joint Conference on Artificial Intelligence*, pp.1606–1611 (2007).
8) Giles, J.: Internet Encyclopaedias Go Head to Head, *Nature*, Vol.438, pp.900–901 (2005).
9) Peat, H.J. and Willett, P.: The Limitations of Term Co-occurrence Data for Query Expansion in Document Retrieval Systems, *Journal of the American Society for Information Science*, Vol.42, No.5, pp.378–383 (1991).
10) Salton, G., Wong, A. and Yang, C.S.: A vector space model for automatic indexing, *Comm. ACM*, Vol.18, No.11, pp.613–620 (1975).
11) Salton, G. and McGill, M.: *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company (1984).
12) Schütze, H. and Pedersen, J.O.: A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval, *Information Processing and Management*, Vol.33, No.3, pp.307–318 (1997).
13) Strube, M. and Ponzetto, S.P.: WikiRelate! Computing Semantic Relatedness Using Wikipedia, *Proc. National Conference on Artificial Intelligence and Innovative Applications of Artificial Intelligence Conference* (2006).
14) 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, pp.2917–2928 (2006).
15) 中山浩太郎，原 隆浩，西尾章治郎：Web 事典からのシソーラス辞書構築手法，情報処理学会論文誌：データベース，Vol.48, No.SIG 11 (TOD 34), pp.27–37 (2007).

16) 北村美穂子，松本祐治：対訳コーパスを利用した対訳表現の自動抽出，情報処理学会論文誌，Vol.38, No.4, pp.727–736 (1997).

(平成 19 年 6 月 18 日受付)
(平成 19 年 10 月 9 日採録)

(担当編集委員　今村 誠)

**伊藤　雅弘**（学生会員）

2007 年立命館大学理工学部情報学科卒業．現在，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博士前期課程在学中．人工知能，WWW からの知識獲得および情報検索に関する研究に興味を持つ．日本データベース学会の学生会員．

**中山浩太郎**（正会員）

2001 年関西大学総合情報学部卒業．2003 年同大学院総合情報学研究科修士課程修了．この間（株）関西総合情報研究所代表取締役社長，同志社女子大学非常勤講師に就任．2004 年関西大学大学院を中退後，2007 年大阪大学大学院情報科学研究科にて博士号を取得し，同年から大阪大学大学院情報科学研究科特任研究員となり，現在に至る．人工知能および WWW からの知識獲得に関する研究に興味を持つ．IEEE，ACM，電子情報通信学会，人工知能学会の各会員．

**原　　隆浩**（正会員）

1995 年大阪大学工学部情報システム工学科卒業．1997 年同大学院工学研究科博士前期課程修了．同年同大学院工学研究科博士後期課程中退後，同大学院工学研究科情報システム工学専攻助手，2002 年同大学院情報科学研究科マルチメディア工学専攻助手，2004 年より同大学院情報科学研究科マルチメディア工学専攻准教授となり，現在に至る．工学博士．1996 年本学会山下記念研究賞受賞．2000 年電気通信普及財団テレコムシステム技術賞受賞．2003 年本学会研究開発奨励賞受賞．データベースシステム，分散処理に興味を持つ．IEEE，ACM，電子情報通信学会，日本データベース学会の各会員．

**西尾章治郎**（フェロー）

1975 年京都大学工学部数理工学科卒業．1980 年同大学院工学研究科博士後期課程修了．工学博士．京都大学工学部助手，大阪大学基礎工学部および情報処理教育センター助教授，大阪大学大学院工学研究科情報システム工学専攻教授を経て，2002 年より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教授となり，現在に至る．2000 年より大阪大学サイバーメディアセンター長，2003 年より大阪大学大学院情報科学研究科長，その後 2007 年より大阪大学理事・副学長に就任．この間，カナダ・ウォータールー大学，ビクトリア大学客員．データベース，マルチメディアシステムの研究に従事．現在，Data & Knowledge Engineering 等の論文誌編集委員．本会理事を歴任．電子情報通信学会フェローを含め，ACM，IEEE 等 8 学会の各会員．