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

2007－DBS－143 (91)
2007／7／4

Wikipediaのリンク共起性解析による シソーラス辞書構築のスケーラビリティ向上

伊藤 雅弘† 中山浩太郎† 原 隆浩† 西尾章治郎†

† 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
〒565-0871 大阪府吹田市山田丘 1-5
E-mail: †{ito.masahiro,nakayama.kotaro,hara,nishio}@ist.osaka-u.ac.jp

あらまし 近年，Wikipediaに代表されるような，記事同士がハイパーリンク（以降リンク）で結び付けられたWebベースの事典が公開されている．筆者らはこれまでの研究において，Web事典から精度の良いシソーラス辞書を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Web事典を解析するためには，スケーラビリティの更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．

キーワード Webマイニング，Wikipediaマイニング，共起，シソーラス

A Scalable Thesaurus Construction Method based on Link Co-occurrence Analysis for Wikipedia

MASAHIRO ITO†, KOTARO NAKAYAMA†, TAKAHIRO HARA†, and SHOJIRO NISHIO†

† Dept. of Multimedia Engineering, Grad. School of Information Science and Technology, Osaka Univ.
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
E-mail: †{ito.masahiro,nakayama.kotaro,hara,nishio}@ist.osaka-u.ac.jp

Abstract Web based encyclopedias, such as Wikipedia, have become dramatically popular among internet users. We have already proved how effective they are 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.

Key words Web Mining, Wikipedia Mining, Co-occurrence, Thesaurus

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挙げられる．シソーラス辞書は，語彙同士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の爆発的な普及に伴い，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，市販の百科事典の記事数が数万～10万であることと比較してもその規模が膨大であることがわかる．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）同士がハイパーリンク（以下リンク）で互いに参照されていることである．

筆者らは，Wikipedia のこれらの特性に着目し，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，語彙同士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 [13] [14]．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更なる向上が技術的な課題であった．文献 [13] [14] の手法では，n ホップ先までのリンク構造を解析し，語彙同士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，計算量が爆発的に増大する可能性がある．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，第 2 章で関連研究について述べ，第 3 章でリンクの共起性解析について解説する．第 4 章では実験により本手法を評価する．最後に，第 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．例えば，語の共起関係に基づいて構築するもの [11] 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，語義の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り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．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 [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．例えば，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">
Apple
</a>
```

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，探索空間が広すぎ内容が発散する一方，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同士の参照である．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同士の関係が表現されているものは少なく，概念同士の関連を抽出するためには，説明文の中からさらに一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，良質な概念同士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（リダイレクトリンクを含む）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，解析内容が発散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，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．形態素解析において未知語が存在すると，解析の精度が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，その判別は困難である．しかし，Wikipediaでは一つの記事（ページ）が一つの概念を表しており，多義語においても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．筆者らは，Wikipediaに対してWebマイニングを行い，有益な情報を抽出する手法の総称をWikipediaマイニングと呼んでいる．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[10]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）の二つの指標を利用し，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同士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によって，リンクを次元，その各リンクの重要度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，一つの概念の特徴ベクトルを抽出するには一つの記事に存在するリンク情報だけを解析すれば良いため，スケーラビリティは高い．しかし，それ故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，精度が低下する可能性がある．

2.3.2 lfibf

lfibf[13][14]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）の二つの指標を利用し，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わされる．

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

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が持つ他の記事からのリンク（Backward Link）の数の逆数である．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．

$$lfibf(v_i, v_j) = lf(v_i, v_j) \cdot ibf(v_j) \quad (4)$$

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

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対してより高い値を示す．

この手法では，nホップ先までのリンク構造を解析し，語彙同士の関連度を算出している．そのため，一つの概念に対する計算量が多く，全体として膨大な計算が必要になるという問題がある．

3. リンクの共起性解析

前章の2.3項に挙げたように，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，局所的情報を元にしたり，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，精度に影響を与えたり膨大な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．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のように膨大な計算が必要になることはない．

本章では，Wikipediaにおけるリンクの共起性に基づいて，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，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た解析手法である．単語の共起解析は，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][15]．以下にそれぞれの式を示す．$P(x)$,$P(y)$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

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

北村らは，Dice 係数の欠点は単語ペア出現回数の大小に関わらず，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 [15]．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

Hinrich らは文書コーパスから単語の共起に基づくシソーラスを構築し，情報検索に応用する手法を提案している [11]．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（ページ間）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．

ところで，Wikipedia を解析する時，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み合わせが存在する．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みなす [11]．そして，各記事において各共起毎の数をカウントし，全記事の計算結果を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の共起回数を算出することができる．

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．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，もっと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．例えば，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みなされる．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 3.1.2 項に挙げた 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，リンクを次元，その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 の共起特性を表わすベクトル v_i のベクトル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．

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

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

本節では，実験内容について述べる．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．比較としては，tfidf，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，精度と構築時間を比較する．解析対象の Wikipedia のデータとしては，2006 年 9 月時点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 4000 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する．

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] [12] 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 組の単語を 13 人〜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から抽出し，正解データとの順位相関を「スピアマンの順位相関係数 (Spearman rank-order correlation coefficient)」で求め，シソーラスの精度とした．

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い．文献 [7] [12] 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．例えば，“Aluminium” と “Metal” の比較で，鉄に関する比較をしているにも関わらず，“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，一般的に比較する時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．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．将来的には Wikipedia の成長に伴い，Backward リンク数が少ないページも減少していくと思われる．

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．

表 1 計算機環境

項目	仕様
CPU	Intel Xeon 5160 3.0GHz × 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

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 2〜5 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．

提案手法ではウインドウサイズが 2 から 5 に増えるに伴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が増えるに従ってほぼ 300 秒程度と線形に増加している．

提案手法と tfidf の計算時間を比較すると，提案手法のウインドウサイズ 2 においては tfidf と比べて約 1.5 倍の時間を，ウインドウサイズ 5 においては，約 4.9 倍の時間を要している．

次に，提案手法と lfibf の計算時間を比較すると，lfibf は提案手法のウインドウサイズ 2 に対して約 204 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 63 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 [14] に述べられている近似手法を用いても，膨大な計算が必要である．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．

表 2 シソーラス辞書構築に要する時間

手法	計算時間 (秒)
提案手法 (Win2)	419
提案手法 (Win3)	741
提案手法 (Win4)	1,063
提案手法 (Win5)	1,365
tfidf	278
lfibf	85,472

4.2.2 シソーラス辞書の精度

提案手法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 2〜5 のそれぞれにおいて，3.1.1 節で示した 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．

表 3 シソーラス辞書の精度：提案手法

ウインドウサイズ	一次共起計算手法	スピアマン
2	Cosine	0.65
	MI	0.56
	Dice	0.59
	IDice	0.60
3	Cosine	0.62
	MI	0.62
	Dice	0.59
	IDice	0.58
4	Cosine	0.62
	MI	0.61
	Dice	0.58
	IDice	0.59
5	Cosine	0.62
	MI	0.59
	Dice	0.58
	IDice	0.60

表 4 シソーラス辞書の精度：比較手法

手法	スピアマン
tfidf	0.57
lfibf	0.68

まず表3より，共起性解析による精度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隣り合うリンクを共起とみ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みなすことは，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みられなかった．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．特にウインドウサイズ2のCosineによる一次共起性の計算手法が最も精度が高い．この理由は，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ためである．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．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と提案手法を比較すると，提案手法はlfibfに比べて低い精度となっている．しかし，提案手法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，僅かな差にとどまっている．前項で述べた提案手法とlfibfの計算時間が約204倍も違うことを考えると，提案手法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．

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，本手法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関わらず，高い精度を保っていることがわかった．特に，一次共起性の計算手法の一つであるCosineは，最も高い精度でシソーラス辞書を構築できることがわかった．

今後の展開としては，単に共起回数による計算だけではなく，共起する距離，記事内の出現位置や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．

さらに，自然言語処理技術の適用も課題の一つである．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．

文　献

[1] Brill, E.: A Simple Rule-Based Part of Speech Tagger., *Proceedings of 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., *Proceedings of 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., *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 309–320 (1988).

[5] Davison, B. D.: Topical locality in the Web., *Proceedings of 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., *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 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 Cooccurrence 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. and McGill, M.: *Introduction to Modern Information Retrieval.*, McGraw-Hill Book Company (1984).

[11] 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).

[12] Strube, M. and Ponzetto, S. P.: WikiRelate! Computing Semantic Relatedness Using Wikipedia., *Proceedings of National Conference on Artificial Intelligence and Innovative Applications of Artificial Intelligence Conference* (2006).

[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, pp. 2917–2928 (2006).

[14] 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 19 (2007).

[15] 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).