大規模 Web 事典からのシソーラス辞書構築

Thesaurus Construction from Large Scale Web Dictionaries

中山 浩太郎♥ 原 隆浩♦
西尾 章治郎♦

Kotaro NAKAYAMA, Takahiro HARA
Shojiro NISHIO

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 事典が数多く公開されてきた．筆者らはこれまでの研究で Web 事典に対して Web マイニング手法を適用することで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，効率的かつ精度の高いシソーラス辞書の構築手法が必要とされている．そこで，本研究では *n* ホップ先までのリンク構造を効率的に解析し，語同士の関連度を算出する手法 *lfibf* および 3 つの応用手法「単純法」「対数近似法」「Forward/Backward リンク重みづけ手法」を提案し，実験によりその有効性を示す．

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 efficient methods to analyze the huge amount of Web pages and Web links among articles in encyclopedias. In this paper, we propose "lfibf," an efficient method to construct a Web thesaurus from Web encyclopedias like Wikipedia by using three sub approaches: the "Simple method," the "Log method" and the "Forward/Backward link weight optimization method."

1. はじめに

近年，WWWの爆発的な普及に伴い，Wikipediaに代表される多数の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程度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．英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されている[5]．

一方，情報検索におけるクエリ拡張などの有用な応用から，シソーラス辞書の重要性が認識されている．クエリ拡張とは，（Web）文書の検索システムにおいて，ユーザがクエリとして入力したキーワードを拡張し，意味的に関連する語を抽出することである．この結果，キーワードを直接含まない文書であっても関連度を計算することが可能となる．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や part-of など）を明確に定義した「関連シソーラス」（Relation Thesaurus）と，与えられたクエリ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．本研究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．

筆者らは，これまでの研究において，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，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[7]．しかし，実験を進めていく中で，これまでの提案手法における問題点が明らかになった．それは，特定の条件下でシソーラス辞書構築の精度が低下することである．これは，詳細な調査の結果，一般的な語の解析の際に精度が低下していたことに起因することが分かった．この問題を回避するためには，Web事典のリンク特性を考慮してより最適化された手法が必要となる．

そこで，本研究では*n* ホップ先までのリンク構造を効率的に解析し，語同士の関連度を算出する手法*lfibf* と，3つの応用手法「単純法」「対数近似法」「Forward/Backward リンク重みづけ手法（以下FB 法）」を提案し，実験によりその有効性を示す．文献[7]の手法は，計算方法は異なるがその結果は単純法の結果に相当する．

本論文では，提案手法について詳述し，提案手法により生成されたシソーラス辞書を実験によって評価し，その有用性を示す．最後にまとめと今後の展開を記述する．

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

シソーラス辞書を構築する最も単純な方法は，人間の手によるものである．しかし，シソーラス辞書の構築においては，概念を追加・更新するためには膨大な手間がかかるため，最新の概念や一般的でない語彙などへの対応が難しいのが現状である．そのため，精度の高いシソーラス辞書を低コストで（半）自動的に構築する手法が必要とされている．

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[8]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[1]などがある．しかし，自然言語処理において，語義やかか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未だ残っており，シソーラス辞書構築の精度低下の主要因となっている．

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

♥ 学生会員 大阪大学大学院情報科学研究科マルチメディア工学専 nakayama.kotaro@ist.osaka-u.ac.jp
♦ 正会員 大阪大学大学院情報科学研究科マルチメディア工学専 {hara, nishio}@ist.osaka-u.ac.jp

いる[3]．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[4]．

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[2]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．Chen らの研究ではドメインを限定して Web サイトを選定した後にリンク構造の解析を行い，リンクテキスト上に出現する語の共起性を利用して語同士の関連度を算出している．しかし，Chen らの手法では，解析対象の Web コーパスの特性を考慮していない点や同義語や多義語に関する考察がない点など，語の関連性を解析する際に精度低下を招く要因が多いという問題がある．

3. lfibf

先述のとおり筆者らは，Wikipedia のリンク構造を解析することで高精度にシソーラス辞書を構築できることを示してきた．しかし，文献[7]の手法では，ドメイン特有の語彙や固有名詞に関しては精度が高くシソーラス辞書の構築が出来ているものの，一般名詞や国名など特定の状況下ではシソーラス精度が低下することが判明した．そのため，本研究では大規模 Web 事典に対して効率的かつ高精度にシソーラス辞書を構築する手法の実現を目指し，*lfibf* と 3 つの応用手法「単純法」「対数近似法」「Forward/Backward リンク重みづけ手法」を提案する．本章では，これらについて詳述する．

3.1 lfibf の基本方針

Web 事典は，記事とその関連記事が互いにリンクで参照されているネットワークであるため，記事をノード集合 V，参照（リンク）をエッジ集合 E とする有向グラフ $G = \{V, E\}$ で表現できる．このとき，2 記事間（v_i, v_j）の関連の強さを計測する問題を考えた場合，関連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 への各パスの短さ

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，記事間の関連性は強く，またそのパスの長さが短ければ短いほど強く関連すると考えられる．ここで，パスとはリンクを伝って記事 v_i から記事 v_j へと移動可能な経路を示す．ただし，パスを抽出する際には，リンクの順方向だけでなく，逆方向のパスも抽出するものとする．これは，記事 v_i から記事 v_j の関連の強さを計測するとき，記事 v_i のリンク先の記事が重要であることと同様に，記事 v_i 対するリンク元の記事関係も関連度を算出するための指標として重要であると考えられるためである．上記の考察に基づき，記事 v_i から v_j への全経路 $T = \{t_1, t_2, ..., t_n\}$ が与えられたとき，記事 v_i から v_j の関連性 *lf*（Link Frequency）を以下の式により表現する．

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

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．一方，個々の記事が持つリンクの数も記事間の関連性に影響すると考えられる．例えば無数のリンクを持つ記事は，他のどの記事に対しても多数の短いパスを持つことが考えられる．例えば，記事「United States」は非常に多くの記事から参照されている一種の一般語である．このような記事は，上記の *lf* だけでは他のどの記事に対しても強い関連度を持つことになる．そのため，上記の二つの要素に加え，被参照数も考慮して関連度を計算する必要がある．そのため，上述の *lf* に加え，*ibf*（Inversed Backward link Frequency）を導入し，*lfibf* を以下の通りに定義する．

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

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

N は全記事数，df (v_j) は記事 v_j が持つ他の記事へのリンク数とする．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．

このとき，*ibf* では記事 v_j の持つリンク数だけ考慮し，中継ノードのリンクの数を考慮していないのは，リンクを多く含む記事（通常は多くのユーザによって精査されている記事）内のリンクが軽視されることを防ぐためである．

3.2 単純法

前述の通り，*lfibf* はリンク構造を解析し，経路の多さと距離に応じたスコアリング（*lf*）と被リンク頻度（*ibf*）を利用した，語彙同士の関連性の数値化アルゴリズムである．しかし，グラフ内の全経路を算出することは，ノード数とリンク数に応じて莫大な計算量が必要となる NP 困難問題であることが知られている．そのため，*lfibf* では探索距離を限定し，近似解を求めるアルゴリズムを実現している．以下にアルゴリズムの詳細を解説する．

有向グラフ G は，隣接行列（A）で表現することが可能であり，隣接行列のべき乗 A^n を計算することで距離 n の全てのノード（記事）への経路数を算出できることは周知である．ところで，記事 v_i は，Forward リンクと Backward リンクの 2 種類のリンクを持つ．v_i の Forward リンクとは，記事 v_i から別の記事へジャンプするリンクの集合であり，$F_{vi} = \{f_{i1}, f_{i2}, f_{i3}, ..., f_{im}\}$ と定義する．また，Backward リンクは別の記事から記事 v_i へジャンプするリンクの集合であり，$B_{vi} = \{b_{i1}, b_{i2}, b_{i3}, ..., b_{il}\}$ と定義する．このとき，*lfibf* では Forward リンクだけでなく Backward リンクの解析も行うために，隣接行列 A と転置行列 A^T を加算した行列 A' を解析に利用する．

$$A' = A + A^T.$$

次に，行列 A' 内の各要素を各列における要素の和（被リンク数）で除算した推移確率行列 P およびその積 P^n を利用することで，n ホップ先のノード（記事）への推移確率を算出し，記事同士の関連性を算出する．ここで，各行における要素の和ではなく，各列の要素の和で除算するのは，隣接行列 A' の各列の和は，記事 v_j の被リンク数（被リンク数）であり，前述の *ibf* を近似できるためである．

ここで，v_i から v_j への n ホップ以内の全推移確率行列 P^1, P^2, ..., P^n が与えられた時，*lfibf* を以下の通り定義する（$p^l_{i,j}$ は P^l の i 行 j 列要素）．

$$lfibf(i, j) = \sum_{l=1}^{n} \frac{1}{d(n)} \cdot p^l_{i,j}.$$

$d(n)$ は，ホップ数 n に応じて増加する単調増加関数もし

くは指数関数である．また，P^{n+1} の算出には P^n の解析結果を利用するため，解析対象の記事を起点として再帰的に解析を繰り返す方式に比べ，効率よく解析することが可能である．

3.3 対数近似法

上述の単純法のように，行列 A' の各要素を各列の和で単純に除算した場合，被リンク数に反比例して値が線形で減少するため，*ibf* の近似としては不十分であることが考えられる．そのため，行列内の各要素を被リンク数で除算する代わりに，対数関数を利用して除算することで，*ibf* を近似する手法（対数近似法）を提案する．対数近似法では，A' から推移行列 P を計算する前に，A' の各要素を以下の数式に従って更新する．

$$a'_{i,j} = a'_{i,j} \cdot \log \frac{N}{|B_{v_j}|}.$$

3.4 Forward/Backward リンク重みづけ手法

ここで，リンク解析における Forward リンクと Backward リンクの重要性について考察する．Forward リンクは，通常記事の著者の主観に依存し，重要な単語へのリンクがあるか，その数が妥当かなど，情報の信頼性にはばらつきがみられるのが一般的である．しかし，その一方で多数のユーザによって長期間洗練された記事は，関連の深い記事へのリンクが豊富であり，内容に間違いも少ない．このような記事（ページ）の情報の信頼性は，Backward リンクの数によって判断できる場合が多い．これは，Backward リンクがページに対する「投票」と見做すことができるためである．HITS アルゴリズムや PageRank アルゴリズム[6]など，最近の Web 構造解析のアルゴリズムでも，Backward リンク解析が客観的な情報を得るために有用であることが示されている．実際に，ドメイン特有の単語（専門用語）の場合には特に Backward リンクが重要な役割を果たすことが予備実験によって判明している．これは，ドメイン特有の単語の場合，ドメイン内で密なリンク構造が形成されており，Forward リンク解析では発見できなかった語彙同士の関連情報を Backward リンク解析から抽出できたためだと考えられる．しかし，その逆に一般的な語の場合は，様々な分野の記事から参照されるため，Backward リンク解析の結果が分散してしまい，関連語の抽出精度が下がってしまうという現象がみられた．これは，Backward リンク数の多い語（一般的な語）は，記事の内容が信頼できるため，Forward リンクを重視して解析することが望ましい一方で，Backward リンク解析の結果は分散してしまうため，比重を下げる必要があることを示唆している．そのため，記事の Backward リンク数に応じた Forward/Backward リンクの重み付けを適用した FB 法を提案する．FB 法では，A' の各要素を以下の通りに更新する．

$$a'_{i,j} = W(|B_{v_j}|) \cdot a^T_{i,j} + (1 - W(|B_{v_j}|)) \cdot a_{i,j},$$

$$W(|B_{v_j}|) = 0.5/(|B_{v_j}|^{\alpha}).$$

$W()$ は，記事 v_j の持つ Backward リンク数に応じて Backward リンクの重みを変更する関数である．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．

4. 実験と考察

本章では，*lfibf* の有効性を示すために行った実験について述べる．

4.1 実験の概要

本研究では，提案手法の有効性を示すために，*lfibf* の 3 応用手法と Chen らの手法を比較した．

本実験では，クエリとして入力されたキーワードから関連語を 30 件抽出する簡易の検索エンジンを構築し，評価に利用した．この検索エンジンでは，まず与えられたクエリに対して各手法で構築された関連語のリストを関連度の高い順に 30 件抽出する．次に，それぞれのシソーラス辞書で構築した関連語を順不同で被験者に提示し，被験者が各関連語とクエリとの関連度を 5 段階（1: 関係しない← 3: どちらともいえない→ 5: 関係する）で評価した．

ただし，関係があるか否かの判断が被験者の偏った主観に依存することを防ぐために，is-a 関係や is-apart-of 関係など，語から連想できる語のことを「関係ある語」と定義していることを被験者に明確に示した上で実験を行った．さらに，実験結果をより公正なものとするために，被験者には「関係のある語も関係のない語も含まれている可能性がある」と伝えた．評価値としては，シソーラス辞書の精度評価でよく利用される CP 値(Concept precision)[1]を以下の式により算出した．ここで，「関係が深いと評価された関連語の数」は，被験者の評価で 5 もしくは 4 と評価された関連語の数であり，「システムが抽出した関連語の数」は，クエリから抽出された関連語の数を示す．

$$CP = \frac{\text{関係が深いと評価された関連語の数}}{\text{システムが抽出した関連語の数}}.$$

本実験では，合計 15 名の被験者に対して検索エンジンを利用させ，それぞれ別々の任意の一単語をクエリとして，各シソーラス構築手法で抽出された関連語 30 件に対して CP 値による評価を行った．

4.2 実験の結果

シソーラス辞書の精度に関する実験の結果を図 1 に示す．横軸の Rank は，関連語 30 件を関連度の高い順にソートした時の順位（ランク）を示し，縦軸の CP は該当ランクでの CP 値の平均値である．Simple, Log, FB はそれぞれ *lfibf* の 3 手法を示す．まず，単純法と対数近似法を比較した場合，対数近似法のほうが精度よくシソーラス辞書を構築できていることがわかった．これは，推移確率行列 P を作成する際には，単に被リンク数で除算するより，対数関数で *ibf* を近似するほうが精度良くシソーラス辞書構築ができることを示している．次に対数近似法と FB 法を比較した場合，「Microsoft」や「iPod」といったドメインに特化した専門的な用語等の場合は，ランキング下位では FB 法のほうが約 10%高い精度を実現していたが，ランキング上位において精度に大差は生じなかった．しかし，「Book」や「Music」，「Sport」といった，一般的な語（Backward リンクの数が非常に多い語）の関連語に関しては，FB 法の方が精度よくシソーラス辞書を構築できており，全体の精度に約 4.74%の差が生じた．Chen らの手法は，形態素解析による精度低下が発生した．これは，語の共起性解析において，リンクテキストを自然言語処理ツールにより空白，ハイフン，カンマ，ピリオドなどの区切り文字で単語・フレーズに分割する際に，適切ではない箇所で形態素に分割されたことに起因する．

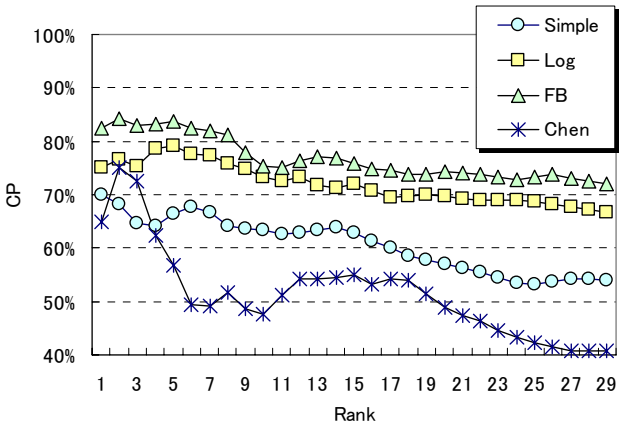

図 1　最終的なランクと CP 値の関係

Fig.1 CP vs. Rank.

5. まとめと今後の展開

本論文では，Wikipedia などの大規模 Web 事典をマイニングし，シソーラス辞書を構築する手法として *lfibf* を提案した．実験の結果から，生成されたシソーラス辞書は，関連度の高い語を抽出していることがわかった．特に，被リンク頻度を考慮した FB 法は高精度のシソーラス辞書を構築する上で有効であることがわかった．

本プロジェクトの成果は，Wikipedia を解析することで有用な情報を抽出するプロジェクトである「Wikipedia Mining プロジェクト」として「http://wikipedia-lab.org」で公開している．本サイトでは，*lfibf*（FB 法）によって構築されたシソーラス辞書を利用することが可能である．また，開発者用の API を XML Web サービスとして公開しており，WSDL（Web Services Description Language）によって仕様を定めている．

今後の展開としては，日本語を含めた多言語 Wikipedia での実験が非常に興味深いと思われる．例えば，言語間リンクの解析による翻訳用シソーラス辞書の構築などが応用例として考えられる．ただし，これら別プロジェクトとの連携するためには十分な量のコーパスが必要となるが，現在の段階では十分なデータが他プロジェクトに揃っていないのが現状である．また，自然言語処理技術との統合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 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．

[文献]

[1] Chen, H., Yim, T. and Fye, D.: Automatic Thesaurus Generation for an Electronic Community System, Journal of the American Society for Information Science, Vol. 46, No. 3 (1995).
[2] Chen, Z., Liu, S., Wenyin, L., Pu, G. and Ma, W.Y.: Building a Web Thesaurus from Web Link Structure, Proceedings of the ACM SIGIR, pp.48-55 (2003).
[3] Craswell, N., Hawking, D. and Robertson, S.: Effective Site Finding using Link Anchor Information, Proceedings of the ACM SIGIR, pp. 250-257 (2001).
[4] Davison, B.D.: Topical Locality in the Web, Proceedings of the ACM SIGIR, pp. 272-279 (2000).
[5] Giles, J.: Internet Encyclopaedias Go Head to Head, Nature, Vol.438, pp.900-901 (2005).
[6] Lawrence, P., Sergey, B., Rajeev, M. and Terry, W.: The PageRank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford Digital Library Technologies Project, pp.39.41 (1999).
[7] 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).
[8] Schutze, H. and Pedersen, J.O.: A Cooccurrencebased Thesaurus and Two Applications to Information Retrieval, International Journal of Information Processing and Management, Vol.33, No.3, pp.307-318 (1997).

中山 浩太郎　Kotaro NAKAYAMA
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，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，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後期課程在学中．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．情報処理学会の正会員および電子情報通信学会，日本データベース学会，ACM，IEEE の各学生会員．

原 隆浩　Takahiro HARA
1997 年大阪大学大学院工学研究科博士前期課程修了．同年，博士後期課程中退後，同大学大学院工学研究科情報システム工学専攻助手，同大学大学院情報科学研究科マルチメディア工学専攻助手を経て，2004 年より同大学大学院情報科学研究科マルチメディア工学専攻助教授となり，現在に至る．工学博士．データベースシステム，モバイルコンピューティングなどの研究に従事．IEEE，ACM，電子情報通信学会，情報処理学会，日本データベース学会の各会員．

西尾 章治郎　Shojiro NISHIO
1980 年京都大学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教授を経て，2002 年より同大学大学院情報科学研究科教授となり，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を併任．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，本学会理事，電子情報通信学会，情報処理学会の各フェローを含め，ACM，IEEE など 8 学会の会員．