Vol.2009-MPS-76 No.26
Vol.2009-BIO-19 No.26
2009/12/17

関連語ネットワークへの未知語挿入アルゴリズムの検討

芝野　功一郎[†1]　廣　安　知　之[†3]
三　木　光　範[†2]　横　内　久　猛[†3]

シソーラスは，情報検索や機械学習を行う際に広く利用されている．通常は，シソーラスは木構造であり，新しい単語は既存の木構造に追加される．一方で，単語は複数のカテゴリの意味を有することが多々あり，このような場合には，シソーラスをネットワーク構造として取り扱うことが適切である．本研究では，単語間の関連を示した関連語ネットワークにおいて，ネットワーク上に存在しない語（未知語）を挿入するアルゴリズムを提案する．これまでの研究では，単語の上位/下位関係を表したツリー構造における未知語挿入が可能であったが，本研究ではそれをネットワーク構造においても可能としている．提案アルゴリズムでは，tf-idf 法を用いて未知語に対する関連語を取得し，それらの関連語の隣接ノードの数や，各関連語間の最短距離などを用いて，未知語をネットワークへ登録する際にエッジをつなぐノードを決定し，挿入する．

A study of algorithm inserting unknown word to related term network

KOICHIRO SHIBANO,[†1] TOMOYUKI HIROYASU,[†3]
MITSUNORI MIKI[†2] and HISATAKE YOKOUCHI[†3]

Thesaurus is very useful for information search, machine learning and so on. Usually thesaurus has a tree topology and new words should be inserted into this tree properly. On the other hand, words have several types of categories and to express this relationship between the words network topology is suitable. In this paper, we propose an algorithm where unknown words are inserted into thesaurus which has network topology. In this algorithm, related words of unknown word are derived by tf-idf method, and nodes which unknown word is inserted to network are determined. This inserted node is determined using the information of a number of their related word's adjoining node and shortest distance between each related words. Using this algorithm, unknown words can be inserted into thesaurus whose topology is not only tree structure but also network structure.

1. はじめに

近年，人工知能の自然言語処理の分野において数多くのシソーラスが構築され，情報検索や機械翻訳などに利用されている[1)2)]．シソーラスとは，単語の上位/下位関係，部分/全体関係，同義関係，類義関係などによって単語を分類し，体系づけた辞書のことであり，主にツリー構造で表される．シソーラスに完成形はなく，日々新しい単語が出現するごとに，新しいシソーラスへと更新してゆく必要がある．シソーラス上に存在しない語（未知語）のシソーラスへの登録は非常に重要な役割を担っている．その一方で，単語は複数のカテゴリの意味を有することが多々あり，このような場合には，シソーラスをネットワーク構造として取り扱うことがある．これまでの研究では，未知語の新たな登録は主にツリー構造を対象に行われてきた[3)4)]．

そこで，本発表では未知語の登録をネットワーク構造においても可能にするアルゴリズムを提案する．ネットワーク構造へ未知語を新しく登録することにより，未知語と隣接した単語とのつながりもわかり，ネットワーク全体で見た時の未知語の位置付けがわかりやすくなる．

本研究では，Yahoo! API を用いて構築した単語間の関連を示した関連語ネットワークを既存ネットワークとし，未知語を新たに登録する．登録方法は，tf-idf 法を用いて未知語に対する関連語を取得し，それらと既存ネットワーク上での共通のノードを，未知語と関連のある候補ノードとし，各候補ノード間のネットワークにおける最短距離や，候補ノードの隣接ノードの数などを用いて未知語とつなぐノードを決定し，未知語を新たに既存ネットワークへ登録する．

2 章では関連技術，3 章ではネットワークの構築方法，4 章では未知語挿入アルゴリズム，最後に 5 章でまとめと今後の課題を述べる．

†1 同志社大学大学院生命医科学研究科
Graduate School of Life and Medical Science, Doshisha University
†2 同志社大学理工学部
Department of Science and Engineering, Doshisha University
†3 同志社大学生命医科学部
Department of Life and Medical Sciences, Doshisha University

Vol.2009-MPS-76 No.26
Vol.2009-BIO-19 No.26
2009/12/17

2. 関 連 技 術

2.1 tf-idf 法

tf-idf 法とは，文章中の特徴的な単語（重要とみなされる単語）を抽出するためのアルゴリズムであり，主に情報検索や文章要約などの分野で利用されている[5)6)]．tf とは，Term frequency(単語出現頻度) の略であり，同じ文書に何回も現れる単語ほど重要であるという考え方である．ある文書で出現する単語の頻度を f , 文章で出現する総単語数を $max(f)$ とすると，

$$tf = \frac{f}{max(f)} \tag{1}$$

と表すことができる．また，idf とは Inverse document frequency(逆文書出現頻度) の略であり，多くの文書に出現する単語はあまり重要ではないため，重要度を低くするという考え方である．検索エンジンの総ドキュメント数を N，ある単語を検索した時のヒットページ数を df とすると，

$$idf = \log_2(\frac{N}{df}) \tag{2}$$

と表すことができる．よって，文章中のある単語の重要度 w は，

$$w = tf \times idf = \frac{f}{max(f)} \times \log_2(\frac{N}{df}) \tag{3}$$

と表すことができる．

2.2 ツリー構造における未知語登録

ツリー構造における未知語登録が関連研究として行われている[7)]．ツリー構造における未知語登録を行う第一段階として，未知語の関連語の抽出を行う．未知語はシソーラスに登録されていないため，従来のシソーラスの階層距離計算[8)] によって関連語を得ることはできない．そこでウェブから関連語を取得する方法を用いる．ウェブからの未知語の関連語を取得する方法として，まず，ある未知語の tf（単語出現頻度）を検索結果より計算する．tf の値が最も高いページに出現する全単語の重要度を tf-idf 法を用いて計算する．そして，tf-idf の値が高い単語を抽出する．これにより，ある未知語に対する関連語を取得することが可能となる．続いて，取得した関連語を用いて未知語をツリー構造のシソーラスに登録する処理を行う．はじめに，シソーラスのルートに注目し，取得した関連語の数が一番多い枝を選択し，下位の階層に進む．次の分岐点においても関連語の数が一番多い側を選択しさらに下位の階層へ進む．このようにして最も下位の階層にある関連語へ行き着くまで，この処理を繰り返す．最後に，最も下位の階層の関連語の親ノードの下に未知語を登録することで，ツリー構造のシソーラスへ未知語を登録することが可能となる．

3. ネットワークの構築方法

本章では，Yahoo! API を用いた関連語ネットワークの構築方法について述べる．

3.1 Yahoo! API

Yahoo! API とは，Yahoo!検索で使用されたキーワード情報をもとに，指定されたキーワードと組み合わせて検索される関連語を提供している API である．例として,「京都」に対する関連語は図 1 のような XML 形式で取得することができる．これらの内，今回は上位 10 件を京都の関連語としてネットワーク構築に用いる．

```
<Result>京都大学</Result>
<Result>京都 観光</Result>
<Result>京都御所</Result>
<Result>京都 ホテル</Result>
<Result>京都市</Result>
<Result>京都 市バス</Result>
<Result>京都新聞</Result>
<Result>京都府</Result>
<Result>京都 紅葉</Result>
<Result>京都駅</Result>
```

図 1 京都に対して得られた関連語

3.2 関連語を用いたネットワークの構築

図 2 の XML 形式のデータを正規表現を用いて，関連語のみを抜き出す．例として，図 2 の 1 行目の「京都大学」を正規表現により抜き出し，ノードにする．そして，そのノードとクエリワードである「京都」は関連のあるノードであるので，エッジをつなぐ．その他の単語もクエリワードとし，関連語を取得し，それぞれのクエリワードを繋げることでネットワークを構築する．図 2 に構築したネットワークの一部を例として示す．

図 2 の例は，クエリワード「京都」,「奈良」,「ハワイ」,「イタリア」,「中国」を用いて，構築したネットワークの一部である．

Vol.2009-MPS-76 No.26
Vol.2009-BIO-19 No.26
2009/12/17

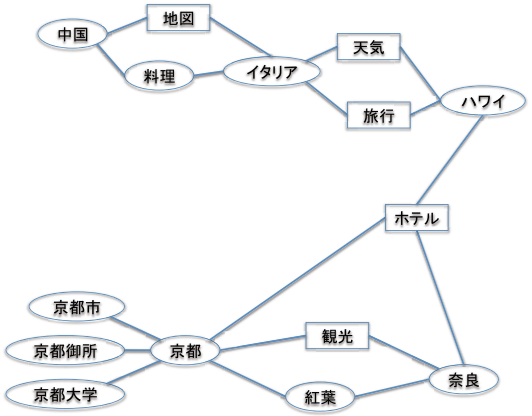

図 2 構築したネットワークの一部

4. 未知語挿入アルゴリズム

本章では，未知語挿入アルゴリズムについて説明する．3 章の方法で構築したネットワークには，存在しない語が存在する．本研究では，その語を未知語といい，未知語を登録するアルゴリズムを提案する．

4.1 概　　要

本アルゴリズムは，既存ネットワークに対し，ネットワークに存在しない語（未知語）を新たに登録するアルゴリズムである．既存ネットワークは図 2 のネットワークを用いる．未知語を新たに既存ネットワークへ登録する際に関連のあるノード，つまり，エッジをつなぐノードを決める必要がある．そのノードを決めるため，tf-idf 法を用いて，未知語の関連語をウェブから取得する．既存ネットワークのノードの中で，tf-idf 法により得られた重要度が上位 10 件の関連語と同じ語のノードを本研究では候補ノードと呼ぶ．そして，後に説明する「tf-idf 法による順位」,「候補ノード間の最短距離」,「候補ノードの隣接ノードの数」により重み付けを行い，閾値を定め，最終的に候補ノードの中から未知語とエッジをつなぐノードを決定する．

4.2 候補ノードへの重み付け

4.2.1 tf-idf 法による順位

本項では，tf-idf 法による順位に注目した重み付けについて説明する．tf-idf 法については，2.1 節で説明した通りである．例として,「グアム」という単語を検索した際に上位にあるページを対象として，形態素解析を行い，tf-idf 法を用いて，上位 10 件を得る．今回の場合は,「旅行」,「観光」,「天気」,「ホテル」,「ショッピング」,「地図」,「レオパレス」,「アウトリガー」,「お土産」,「グアムプラザホテル」を取得した．この中で図 2 との共通のノードは,「旅行」,「観光」,「天気」,「ホテル」,「地図」である．よってこれらを未知語とエッジをつなぐ候補ノードとし，四角で表している．これらのノードに，tf-idf 法により得られた順位通りに重みをつけてゆく．重み付けの方法は，表 1 に示すように，順位の一番低いノードの重みを 1 とし，順位が 1 あがるごとに重みを 1 増やす．

表 1 tf-idf 法の順位による重み付け

候補ノード名	tf-idf 法の順位	重み
旅行	1	5
観光	2	4
天気	3	3
ホテル	4	2
地図	5	1

4.2.2 候補ノード間の最短距離

本項では，候補ノード間の最短距離による重み付けについて説明する．ある候補ノードから別の候補ノードまでの最短距離の和が小さいほど，候補ノードの中でも中心的な位置にあると考えられる．例として図 3 のようなネットワークがあった場合，1〜4 までの番号がふってあるノードを候補ノードとする．

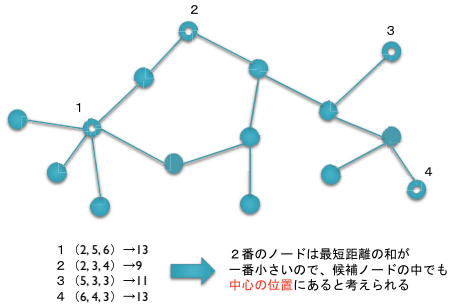

図 3 候補ノードの最短距離

Vol.2009-MPS-76 No.26
Vol.2009-BIO-19 No.26
2009/12/17

番号 1 のノードを例に挙げて説明すると，ノード 1 から 2 までの最短距離は 2，3 までは 5，4 までは 6 となり，最短距離の和は 2 + 5 + 6 で 13 となる．他のノードも同様に計算を行い，各ノードへの最短距離の和を求める．その結果，和が一番小さいのは 2 番のノードとなる．よって，2 番のノードは候補ノードの中でも中心的な位置にあるノードと考えられる．中心的な位置にあるノード，つまり最短距離の和が小さなノードほど高い重みを付け，和が大きなノードほど低い重みを付ける．重み付けの方法は，まず，和の一番大きいノードの重みを 1 とおく．そこから，和が小さい順に重みをつけてゆく．和の値が同じ場合は，同じ重みとし，和の値に同じものが続いた後のノードの重みは同じ数が続いた回数分，重みを大きくする．このアルゴリズムのフローチャートを図 4 に示す．和の値が同じものより，同じでないものの方が重要であり，差をつけなければならないと考え，この方法をとった．図 2 のネットワークの例では表 2 の結果が得られる．

表 2 候補ノードの最短距離の和による重み付け

候補ノード名	最短距離の和	重み
ホテル	9	5
天気	10	3
旅行	10	3
地図	14	2
観光	16	1

4.2.3 候補ノードの隣接ノードの数

本項では，候補ノードの隣接ノードの数による重み付けについて説明する．隣接ノードとは，対象ノードにつながっている全てのノードを指す．隣接ノードの数が多いほど，重要な語である可能性が高いと考えられる．図 2 を例に挙げると，各候補ノードの隣接ノードの数は表 3 の通りである．隣接ノードの数が多い順に重み付けを行う．候補ノードの数を x とした時，最大を x とし，隣接ノードの数が同じ場合は，同じ重みとする．重み付けの方法は，図 4 中の「和の値が前の和の値と等しい」が「隣接ノードの数が前の隣接ノードの数と等しい」に置き換わったアルゴリズムを用いる．この手法による，重み付けの結果を表 3 に示す．

4.3 閾値の設定

本節では，表 1〜3 で求めた各候補ノードの重みの和における閾値を設定する．重みの和が閾値以上のノードを，未知語と関連のあるノード，つまり未知語をネットワークへ登録する際のエッジをつなぐノードとする．

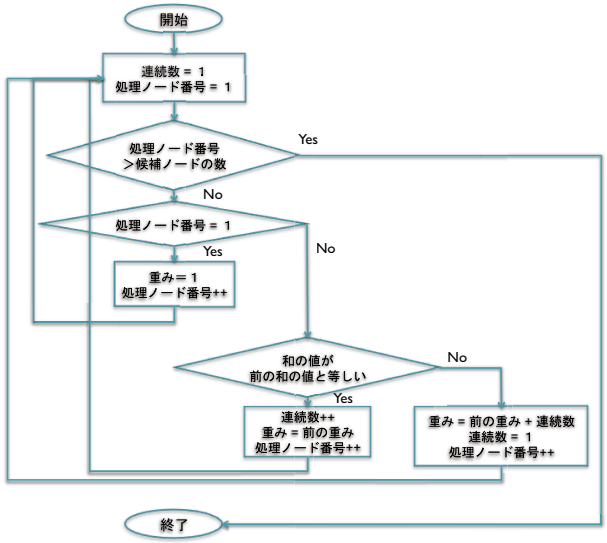

図 4 重み付けのフローチャート

表 3 候補ノードの隣接ノード数による重み付け

候補ノード名	隣接ノードの数	重み
ホテル	3	5
天気	2	1
旅行	2	1
地図	2	1
観光	2	1

閾値の設定方法は,「tf-idf 法による順位」,「候補ノード間の最短距離の和」,「候補ノードの隣接ノード数」によって重み付けを行った全ての候補ノードの重みの和の平均値を求める．例えば,「tf-idf 法による順位」の場合は各候補ノードの重みは，5, 4, 3, 2, 1 なのでこれらの重みの和は 15 となり，平均は 3 である.「候補ノード間の最短距離の和」と「候補ノードの隣接ノード数」の場合も同様にして求めると，2.8 と 1.8 という値になる．これらの各要素の平均値の和を閾値に設定する．よって今回の場合の閾値は，3 + 2.8 + 1.8 で 7.6 となる．これにより，未知語を登録する際にエッジをつなぐノードは，重みの和が閾値 7.6 以上の候補ノード，つまり以下の表 4 に示すように「ホテル」と「旅行」となる．

Vol.2009-MPS-76 No.26
Vol.2009-BIO-19 No.26
2009/12/17

表 4 各候補ノードの重みの和

候補ノード名	tf-idf 法の順位による重み	最短距離の和による重み	隣接ノードの数による重み	重みの和
ホテル	2	5	5	12
旅行	5	3	1	9
天気	3	3	1	7
観光	4	1	1	6
地図	1	2	1	4

この方法により，未知語をネットワークへ登録する際につなぐノードとして，不必要なものを除去するとともに，適切なノードを選ぶ精度の向上が見込める．

4.4 未知語を既存ネットワークへ登録した例

未知語を新たに登録したネットワークを以下の図 6 に示す．手法としては，未知語「グアム」に対する関連語を取得し，それらの内から 4.2 節で説明したように，未知語をネットワークへ登録する際に関連のあるノード，つまりエッジをつなぐノードの候補を決定する．今回の場合は，図 5 の四角マークのノードである「旅行」,「観光」,「天気」,「ホテル」,「地図」を候補ノードとした．これらを 4.2.1〜4.2.3 項で説明した手法により重み付けを行い，それらの重みの和が 4.3 節で設定した閾値以上の候補ノードと未知語をつなぎ，未知語をネットワークへ登録した．

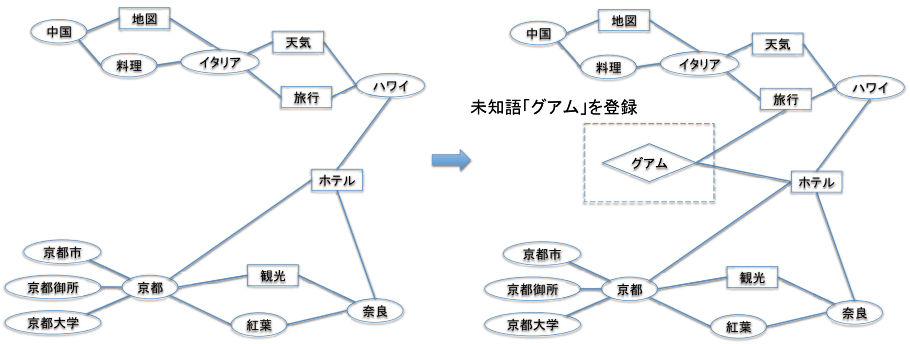

図 5 未知語「グアム」を登録したネットワーク

5. まとめと今後の課題

本アルゴリズムは，既存ネットワークへ未知語を登録する際に「tf-idf 法による順位」,「各候補ノード間の最短距離の和」,「候補ノードの隣接ノードの数」の 3 つの点から重み付けを行い，エッジをつなぐノードを決定した．これにより，関連語ネットワークに未知語を挿入することが可能となった．

今後の課題として，各重みに対するパラメータの設定が挙げられる．パラメータを設定し変化させることで，未知語をネットワークに登録する上で，より適したノードとリンクさせることができると考えられる．tf-idf 法の順位による重みを a，最短距離の和による重みを b，隣接ノードの数による重みを c，各パラメータを α，β，γ とした時，新しい重みの和は，下記のように求められる．新しい重みの和を s とすると，

$$s = a\alpha + b\beta + c\gamma \quad (\alpha + \beta + \gamma = 1) \tag{4}$$

今後，パラメータを変化させた際の新しい重みの和を比較し，より最適なノードとリンクさせることができるパラメータを検証してゆく．

参 考 文 献

1) 野美山 浩．事例の一般化による機械翻訳．情報処理学会論文誌 Vol. 34 No.5，1993
2) 甲田 彰，森田 歌子．科学技術文献検索システムにおける大規模用語辞書の活用について．情報知識学会誌 Vol. 16 (2006) , No. 2 pp.2-45-2-48
3) 浦本直彦．コーパスに基づくシソーラス: 統計情報を用いた既存のシソーラスへの未知語の配置．情報処理学会論文誌, 1996
4) 前田 康成，統計的決定理論に基づく既存名詞シソーラスへの未知語登録方法に関する考察．電子情報通信学会論文誌 A Vol.J83-A No.6 pp.702-710
5) 森 純一郎，松尾 豊，石塚 満．Web からの人物に関するキーワード抽出．人工知能学会論文誌 C 20 巻 5 号，2005
6) Kiyoshi Naganuma，Satoru Hayamizu．Extraction of Topics from Web Documents in the Medical Domain．The 19th Annual Conference of the Japanese Society for Artificial Intelligence，2005
7) 芝野 功一郎，渡部広一，河岡 司．ニュース記事見出しを用いた会話処理システムの構築．同志社大学卒業論文．2009
8) 川島貴広 石川勉．言葉の意味の類似性判別に関するシソーラスと概念ベースの性能評価, 人工知能学会論文誌 20 巻 5 号, pp.326-336, 2005.