F-022

# Wikipedia の記事構造とリンク構造を用いた複数語からの連想手法
# A Method for Association from Multiword Based on Hierarchical Layouts and Link Structure in Wikipedia

長尾 和明† 吉村 枝里子‡ 土屋 誠司‡ 渡部 広一‡
Kazuaki Nagao Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe

## 1. はじめに

Web は人々が情報を取得するための手段として急速に普及し，今や我々の生活に欠かせない存在となっている．しかし，Web の情報の多くは単なるテキストとして表現されているため，意味を考慮した情報検索の実現が大きな課題とされている．意味を考慮した情報検索を実現するための基礎技術の一つとして，単語間の連想的な関係性（連想関係）を取得する技術がある[3].

連想関係を複数の語から取得する場合，それらの語はコンテキストを構成するため，連想される語もそのコンテキストに応じて変化すると考えられる[1]．たとえば，「富士山」と「阿蘇山」という語の組み合わせは，山というコンテキストを構成し，「富士山」と「熱海」という語の組み合わせは，静岡県というコンテキストを構成する．

本研究では，Wikipedia が膨大な記事を登録している Web 百科事典である点に着目し，さまざまな分野における単語間の連想関係を取得するためにデータベースを構築した．そのデータベースを用いて，複数語から連想される語を生成する手法を提案する．

## 2. Wikipedia

Wikipedia は，高い網羅性・即時性や記事間リンク，質の高いアンカーテキスト，URL による語義の一意性の確立など，知識を抽出するためのリソースとして有用な性質を多く持っている．また Wikipedia では，記事（概念）同士がハイパーリンクで互いに参照されている．ハイパーリンクは，ある 1 つのページに着目した場合，フォワードリンク（対象のページから別のページに移動するためのリンク）とバックワードリンク（別のページから対象のページへと移動するためのリンク）に分類できる．本稿では，記事から他の記事へのハイパーリンクを「記事間リンク」とする．

### 2.1 アンカーテキスト

アンカーテキストとは，HTML 文書におけるリンクが設定されたテキストである．Wikipedia において，他の記事へのリンクのアンカーテキストは，リンク先の記事の内容を端的に表す語が利用される．たとえば，企業である Apple に関する記事へのリンクのアンカーテキストは「アップル」「アップルインコーポレイテッド」などの表記が用いられており，これらは同義語であると判断できる．

### 2.2 URL

Wikipedia では 1 つの URL に 1 つの記事が割り当てられているため，多義性が URL によって解決されている．たとえば，「アップル」は，文脈によって意味が変化する多義語であり，企業の「Apple」を指す場合もフルーツの「リンゴ」を指す場合もある．Wikipedia では，これら 2 つの概念はそれぞれ別の記事として管理されており，別々の URL が割り当てられている．

### 2.3 リダイレクトリンク

リダイレクトリンクとは，ある記事が参照されたときに別の記事へとリダイレクト（転送）する機能を持つリンクである．たとえば「邦画」を参照した場合，記事「日本映画」へと自動的にリダイレクトされる．リダイレクトリンクは，主に同義語や表記ゆれを表現し，同一内容の記事が散在することを防ぐために利用される．

## 3. 関連研究

Wikipedia のデータを用いた単語間の関係性における研究として，隅田らによる，Wikipedia を用いた上位下位関係の獲得手法[4] が挙げられる．この手法では，Wikipedia の記事中の節や箇条書き表現（以下，階層構造）から大量の上位下位関係を獲得する．

図 1 に Wikipedia 記事の例として「紅茶」の記事を挙げる．たとえば，この記事には「代表的な産地」や「ブレンド」という節があり，「代表的な産地」の下位には「インド」や「インドネシア」といった小節がある．さらに「インド」の下位には，「アッサム」や「ダージリン」，「ニルギリ」といった項目が存在する．以後，これらの節見出し，小節タイトル，項目名を term とする．

隅田らが提案した上位下位関係の獲得手法は 2 ステップからなる．Step1 では，階層構造上の上下関係を守りながら，2 つの term から 1 つの上位下位関係候補を獲得する．Step2 では，教師あり機械学習を用いて上位下位関係候補から不適切な関係を取り除く．上位下位関係候補が適切か


紅茶

紅茶（こうちゃ、black tea）とは、摘み取った茶の葉と芽を萎凋（乾燥）させ、もみ込んで完全発酵させ、乾燥させた茶葉。

代表的な産地 [編集]

インド [編集]
アッサム (Assam)
ダージリン (Darjeeling)
ニルギリ (Nilgiri)

インドネシア [編集]
スマトラ
ジャワ (Java)

ブレンド [編集]
アフタヌーン
アールグレイ

カテゴリ: 紅茶 | 中国茶 | 発酵食品 | 喫茶文化

図 1 「紅茶」に関する Wikipedia の記事

† 同志社大学大学院理工学研究科
Graduate School of Science and Engineering, Doshisha University
‡ 同志社大学理工学部
Faculty of Science and Engineering, Doshisha University

否かを判定するため，SVM(Support Vector Machine)[2] で学習された分類器を用いて上位下位関係候補を選別する．

また隅田らは，Wikipedia 記事の定義文（記事の第一文に該当）を用いた手法と，記事下部にあるカテゴリを用いた手法も提案している[4]．これらの手法では記事タイトルが下位語として使われる．

## 4. 提案手法

本章では，上位下位関係データベースおよび記事間リンクデータベースの構築方法と，構築したデータベースを用いて複数語から連想を行う手法について述べる．

### 4.1 上位下位関係データベース

本研究では，隅田らの手法[4] により，2016 年 6 月 1 日の Wikipedia ダンプデータから約 930 万の上位下位関係ペアを獲得し，データベース（以下，上位下位関係データベース）を構築した．訓練データは隅田らが実験で用いたデータと同じものを使用した．

### 4.2 記事間リンクデータベース

Wikipedia では，ノード（記事）は概念，リンクは意味的な関係を表し，概念どうしがリンク構造を形成している．そこで本研究では，Wikipedia のリンク構造を解析して大量の概念（記事タイトル）とその属性（アンカーテキスト，およびフォワードリンク先記事タイトル）を抽出し，データベース（以下，記事間リンクデータベース）を構築した．

本研究におけるリンク構造の解析では，リダイレクトリンクを用いて同義語や表記ゆれも考慮しており，記事間リンクデータベースには約 170 万の概念が格納されている．

### 4.3 連想手法

まず入力した複数語に対し，上位下位関係データベースから上位語と下位語，記事間リンクデータベースからバックワードリンク先の記事タイトルとフォワードリンク先の記事タイトル，およびアンカーテキストを各入力語の関連語として獲得する．それぞれが獲得した関連語の和集合を各入力語の関連語集合とする．その後，全ての関連語集合の共通部分を計算し，その全要素を連想語として出力する．表 1 に連想の具体例を示す．

表 1　複数語連想の具体例

| 入力語 | 獲得源 | 関連語 | 出力語 |
|---|---|---|---|
| 出前 | 上位下位 | 輸送，外食産業 | ピザ 洋食 |
| | リンク | ピザ，洋食，電話 | |
| オーブン | 上位下位 | ガス機器，調理器具 | |
| | リンク | ピザ，洋食，グラタン | |
| チーズ | 上位下位 | 具材，乳製品 | |
| | リンク | ピザ，洋食，グラタン | |

## 5. 評価実験

### 5.1 実験方法

提案手法（以下，手法①）の有効性を検証するため，比較実験を行った．比較対象としては，概念ベース[5]（国語辞書や新聞記事などから構築された知識ベース）とシソーラス[6]を用いた連想手法（以下，手法②）を用いた．本実験では，あらかじめ用意した複数語（2，3 語）とその出力結果（100×2 セット）を評価セットとした．

被験者 3 名が，各複数語により生成された連想語群が正しいかどうか，連想語 1 語ごとに三段階評価（○，△，×）を行った．その後，連想語として必須な語が出力されているかどうかなどを参考に，各複数語とその出力に対し，正しい連想ができているといえるか，適不適で評価を行った．

### 5.2 実験結果

被験者による実験結果を表 2 に示す．結果として，手法②に比べ，手法①の方が○のみを正解とした適合率の平均は 23.4%高い．また適切な連想の割合の平均も手法①の方が 21.4%高い．

表 2　連想の実験結果

| | 総出力語数 | ○のみを正解とした各連想語の適合率の平均 | 適切な連想の割合の平均 |
|---|---|---|---|
| 手法① | 1,720 | 39.2% | 54.4% |
| 手法② | 2,885 | 15.8% | 23.0% |

## 6. 考察

評価実験より，手法②と比較すると，手法①の方が出力語数が少なく適切な連想の割合の平均が高いことから，手法①による連想の方が優れていると考えられる．

例として「出前」「オーブン」「チーズ」と入力としたとき，手法①では，ピザに関連する語を獲得するのに対し，手法②では，調理に関連する語を獲得するため，出力の多くが雑音と見なされるといった傾向があった．

## 7. おわりに

本稿では，大規模な Web 辞典である Wikipedia を解析することによって二つのデータベースを構築し，構築したデータベースを用いて複数語から連想される語を出力する手法を提案した．

今後の課題として，定量化された概念間の関連の強さを利用した連想方法の検討が挙げられる．本研究における連想では，各入力語の関連語集合の共通部分の要素を全て出力していた．しかし，データベース内の各概念間の関連の強さを定量化することによって，関連性の低い語の除去を行うことができ，より優れた連想ができると考えられる．

### 謝辞

本研究の一部は，JSPS 科研費 16K00311 の助成を受けたものです．

### 参考文献

[1] Kraft, R., Maghoul, F. and Chang, C.C., “Y!Q:Contextual Search at the Point of Inspiration”, in *Proceedings of International Conference on Information and Knowledge Management* (*CIKM*) , pp.816-823(2005).
[2] Vapnik, V. N., *Statistical Learning Theory*. Wiley-Interscience, (1998).
[3] 岡本潤，石崎俊，“概念間距離の定式化と既存電子化辞書との比較”，自然言語処理，Vol.8，No.4，pp.37-54(2001).
[4] 隅田飛鳥，吉永直樹，鳥澤健太郎，“Wikipedia の記事関係からの上位・下位関係抽出”，自然言語処理，Vol.16，No.3，pp.3-24(2009).
[5] 奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14，No.5, pp.41-64(2007).
[6] NTT コミュニケーション科学研究所監修，“日本語語彙体系“，岩波書店(1997).