社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

# 概念構造に基づく周辺の話題のコンテンツホール検索

渡邉　康平† 　灘本 明代‡

†甲南大学理工学部 〒658-8501 兵庫県神戸市東灘区岡本 8-9-1
‡甲南大学知能情報学部 〒658-8501 兵庫県神戸市東灘区岡本 8-9-1
E-mail: †heavenlystar613@yahoo.co.jp, ‡nadamoto@konan-u.ac.jp

**あらまし** 現在，Web 検索サービスではユーザがクエリとして与えた単語に対して，WWW 上でそれと一致した Web ページ群が返される．そのため，ユーザが目的の単語がわからないときユーザは目的の情報を得ることができないことが課題とされている．我々は，それらを「思いつかない情報」として捉え，ユーザに与えられた単語に概念構造を持たせ，それに基づく周辺の話題のコンテンツホール検索を提案する．
**キーワード** Web 検索サービス　単語　クエリ　WWW　HTML　情報　概念構造　コンテンツホール検索

# Surrounding Topic Content Hole Search based on Content Hierarchy

Kohei WATANABE† and Akiyo NADAMOTO‡

†Faculty of Science and Engineering 8-9-1 Okamoto, Higashinada-ku, Kobe, Hyogo, 658-8501 Japan
‡Dept. of Intelligence and Informatics 8-9-1 Okamoto, Higashinada-ku, Kobe, Hyogo, 658-8501 Japan
E-mail: †heavenlystar613@yahoo.co.jp, ‡nadamoto@konan-u.ac.jp

## 1. まえがき

近年我々は日常的に様々な情報を，インターネットを用いて調べることが多くなると共に，一般の人々が情報を発信することも，SNS やブログなどを介して容易になった．その膨大な情報量の中でいかにユーザの目的に即した有益な情報を取得できるかどうかが課題とされている．現在，インターネット検索システムの進歩により上記の課題を解決するために様々なシステムがある．例えば，Yahoo!や Google ではある単語のスペルの過ちに関して，似たものを提示することにより対応している．しかしながら，仮にユーザが「目的のキーワードを忘れてしまったがそのことについて調べたい.」時や「関連するキーワードは思いつくが，それを調べることが目的ではない.」時に，「目的のキーワード」について検索することは難しい．現在，既存の検索システムでは，ユーザがシステムに入力した単語やフレーズがそのままクエリとして用いられて検索さるため，以下の問題が考えられる．

- 検索キーワードがわからない場合，ユーザが欲しい情報が存在するページを web 上から探し当てることが困難である．
- ユーザは，自分自身が認知している情報しか検索できず，システムではそれらの単語についてのみの情報しか知ることができない．

インターネットにおける検索行為そのものが能動的で，その使い方によっては人々の得る情報が偏ったものになる恐れがある．例えば，ユーザが「にんにく」と「唐辛子」「オリーブオイル」の入ったパスタが食べたいが，「ペペロンチーノ」しか思いつかない．しかしながら「ペペロンチーノ」ではないパスタを食べたいがその名前がわからないといったような検索ができないのが実状である．これらの問題を解決するために，我々はユーザの入力した単語の周辺の情報を検索してユーザに提示する事が必要であると考えられる．これまで我々は７つのコンテンツホール検索を提案してきた[1]．本研究では，この７つのコンテンツホールのうち，周辺の話題のコンテンツホール検索を提案する．

本研究ではユーザが入力する単語に構造化を持たせ，その構造をもつ単語群を Web 検索エンジンや Wikipedia を用いて，上位語と同位語を発見することにより概念構造をリアルタイムで構築し，ユーザが欲しい情報として提供を試みる．つまりは，クエリとしてまず，１つの単語 X を入力する．そして単語 X と part-of 関係にある，もしくは X を構成する要素 l，m，n を入力する．その X の上位概念である単語 Z を Wikipedia を用いて抽出し，それらの単語 X，Z を用いて，概念的に同列でかつで，できるだけ多くの同じ要素を持つ単語 A’$_i$(i=1，…n)を結果として返す．例えば，ユーザは入力キーワード X:「ペペロンチーノ」，l，m，n として「ニンニク」，「唐辛

Copyright ©2009 by IEICE

子」,「オリーブオイル」を入力する．システムは「ペペロンチーノ」の上位概念を示す「イタリア料理」を Wikipedia から抽出する．その「イタリア料理」と「ニンニク」,「唐辛子」,「オリーブオイル」を用いて,「ペペロンチーノ」と同位語である「アラビアータ」を抽出し，ユーザに提示することを行う．

本システムの特徴は以下の通りである．

- 本システムでは上位単語の抽出，概念関係の判定を Wikipedia 用いて抽出する.Wikipedia は一般的な単語を十分に網羅しており，またリアルタイムで編集を行っているので新しい単語にも対応している．したがって，不特定多数の WWW 上の単語に適応していることが考えられる．
- この単語 X と，単語 l，m，n とには part-of 関係にあり，構造を持たせる．これは，ユーザの欲しい情報の概念を絞り込むことで有益性を高める為である．
- 単語 X の同位単語には様々な単語群がある為，同じ概念を持たせるために，上位単語を用いる手法も提案する．

以下，第 2 章に関連研究を，第 3 章に本システムの手法について，第 4 章にプロトタイプシステムについて，第 5 章で実験について述べ，第 6 章でまとめと今後の課題について述べる．

## 2. 関連研究

本研究では，ユーザが入力したキーワードに対して，ユーザが思いつかない周辺の情報を Web 上から連想，抽出する．また最終的に提供する単語はユーザが始めに入力した単語 X の概念構造的に一致もしくは同じであることを目指す．関連語に関する研究は多く提案されている．しかしながら本研究と関連語検索に関する研究とは少し異なる．芳鐘ら[2]は Web 上から，複合語を対象として，ある用語と関連する概念を持つ関連語をテキストから自動的に収集する手法の提案を行っている．これらの最終結果には異なる概念も含まれているので，本研究の目的と異なる．大塚ら[3]は大域 Web アクセスログを用いた関連語の発見により，ユーザに検索語を想起させるシステムの提案を行っている．これらの研究に対して，我々が提案するのは，Web 検索エンジンのスニペットと Wikipedia の記事から概念構造を構築する点で異なっている．我々の研究は即時性に優れていると考える．

概念構造を見出すために我々は，上位，同位単語を求める．そのためには，EDR 電子化辞書[4]やシソーラスのような電子化された辞書を用いることも考えられる．しかし，時が経つにつれて，更新が必要になり，あらゆる Web 上の単語を網羅することが困難である．新里ら[5]は Web から df/idf に基づく上位語候補の獲得を行っている．また，それらと下位単語との関係を意味的類似度の計算により精度をあげることに成功している．しかしながら，この場合多くのデータ収集と計算量を要する．

大島ら[6]は Web 検索エンジンスニペットのみを用いて同位単語の発見を提案した．これは，同位単語が並列助詞「や」によって接続されることに着目している．精度は平均で 70%～80%程度の適合度でクエリ 1 語あたり 6 語程度同位単語が求められたと報告している．本研究では大島らの研究に基づき同位単語を抽出する．

上位単語の抽出において，本研究では Wikipedia を用いる．Wikipedia を用いた概念抽出に関する研究は数多く行われている．灘本ら[1]は Wikipedia の記事の目次（TOC）構造を 1 つの属性情報と見なして，概念構造取得のために用いている．それに対して，本研究では記事のリード部分（冒頭文）1 文を用いる．このリード部分の重要性は中山ら[7]の研究により，その有用であることが報告されている．

## 3. 周辺の話題のコンテンツホール検索の手法

本論文では単語 X から以下の 3 つの発見手法により得られた単語群をコンテンツホールと考え，提案を行う．

### 3. 1 Wikipedia による上位単語の発見

本節では単語 X の上位単語の発見手法について述べる．その手法は下記の仮定に基づいている．

（1） 単語 X が,「X は～である.」もしくは,「X とは～である.」という文章に存在する場合,「である.」の直前の単語は X の上位単語 Z である．

（2） Wikipedia に記載されている記事は世間一般的な単語を網羅している．またそれらの記事間には概念の上下関係が存在する．

（3） リンクタグ（<a>～</a>）に囲まれる文字列は 1 つの単語もしくは，複合語としての意味を持つ．

（4） Wikipedia 内で，ある単語 X に関して説明文の始めの 1 文は（1）の形式をとる可能性が高い．

1 つめの仮定について，単語 X と Z とは is-a 関係にあることは中山ら[7]の研究よって証明されている．この場合，”単語 X は Z である”と言い換えることができる．2 つ目の仮定は，Wikipedia が従来の研究において単語間の概念構造を構築する手法に数多く利用されている．また数々の研究によりそれらの有用性が示め

されている．3 つ目の仮定では，リンクタグのそのアドレス先には必ず情報が存在することが見てとれる．4 つ目の仮定においても，同じように仮定（１）の文脈が存在することが最も多い．

我々は上記の仮説に基づいて，上位単語発見に以下の提案を行う．まず，単語 X を Wikipedia にクエリとして渡す．大多数の Wikipedia は，単語 X の記事の冒頭部の 1 文は「X は～.」という文が存在する．我々はこの Wikipeda の構造に注目し,この冒頭の 1 文を取り出す．そして，この文の文末に最も近いリンクタグで囲まれた文字列は，単語 X の上位単語である場合が多いため，そのリンクタグで囲まれた単語が上位単語であると考える．例えば，単語 X を「ペペロンチーノ」とした場合,「ペペロンチーノはイタリア料理の一種である.」という冒頭の 1 文から,「イタリア料理」は「ペペロンチーノ」の上位単語であることがわかる．

### 3．2　検索エンジンを用いた同位単語の発見

本節では，3．1 の上位単語に加えて単語 X の同位単語の発見手法について述べる．同位単語とはここでは単語 X と同じ属性を持ち，かつ同じ上位単語とより関係が近い単語のことを示す．得られた単語 X の同位単語が単語 X の上位単語とより関係が近いかどうかの判定は次節 3．3 で述べることにする．

我々は大島ら[6]の研究に基づき，下記の 4 つのアプローチで同位単語を Web 検索エンジンから抽出する．

（１）　ユーザが始めに入力した,単語 X の前に１つ，後ろに１つ並列助詞「や」を付随し，クエリを２つ用意する．例えば，単語 X が「ペペロンチーノ」であれば，クエリは「”ペペロンチーノや“」と「“やペペロンチーノ”」となる．ここで，引用符で括られているフレーズがそのまま出現するようなページを検索するためにフレーズ検索を用いる．

（２）　さらに，その２つのクエリの各々に，単語 X の構成要素である単語 l，m，n を加えて，これらを，検索クエリとして用いる．また，それらを Web 検索エンジンで AND 検索し，結果を得る．これは，単語 X と同じ構成要素を含む同位単語をより多く取得する為である．

（３）　検索結果のタイトルとスニペットを解析対象とする．そして，形態素解析 Mecab を用いて解析し，「や」に続く，もしくは，「や」の直前の名詞のみを抽出する．

（４）　（１）～（３）を 2 つのクエリ両方に対して行い，両方に得られた単語のみを単語 X の同位単語候補とする．

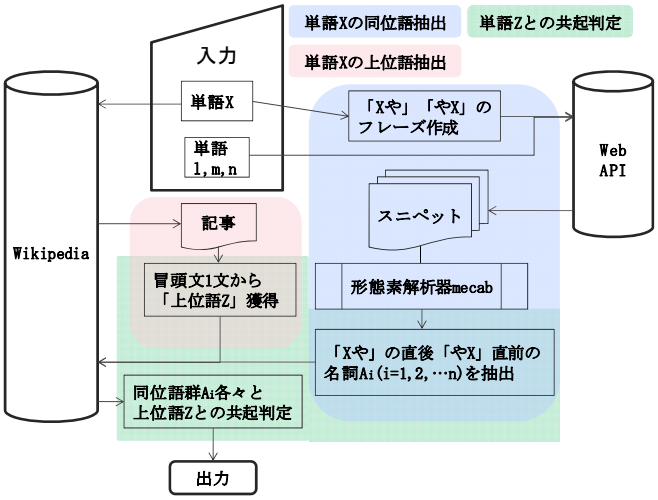


図 1．プロトタイプシステムの概要

図 2．プロトタイプシステム

### 3．3　同位単語候補の判定

得られた同位単語候補にユーザの目的とは違った単語が含まれている可能性がある．ここでいう目的とは **3．1** で抽出した単語 X の上位単語である単語 Z に包含されていることである．すなわち，単語 X が「ペペロンチーノ」である場合，単語 Z が「イタリア料理」であるとすると，同位単語候補は「イタリア料理」に包含される．その単語 Z と関係のない単語は同位単語候補から省く手法をこの節で提案する．Wikipedia では 1 つの記事に様々な記事へのリンクが存在する．もし，3．1 で得られた単語候補が単語 X と同じ概念を持つと考えると，単語 X の上位単語と関係があると考える．そこで，我々は以下の手法により同意単語を取得する．

（１）　得られた単語候補各々に対して Wikipedia の記事を検索する．

（２）　その記事中に単語 Z が出現するか否かを判定する．

例えば,「ペペロンチーノ」の同位単語候補の１つと

表 1. 上位単語抽出の実験結果のまとめ

| | 全体71 | 全体の割合% |
|---|---|---|
| 正解数 | 45 | 63.3 |
| 不正解(誤り) | 18 | 25.3 |
| 不正解(記事なし) | 7 | 9.8 |
| 不正解(リンクなし) | 2 | 2.8 |
| 適合率 | | 63.3 |

| 単語X | x1 x2 x3 | 上位単語Z | Web取得件数 50件 |
|---|---|---|---|
| ペペロンチーノ | にんにく<br>オリーブオイル<br>唐辛子 | イタリア料理 | **アラビアータ**<br>唐辛子 **ボンゴレ**<br>パスタ |
| ゴーヤーチャンプルー | ゴーヤ<br>卵<br>豚肉 | 沖縄料理 | **フーチャンプルー**<br>**トーフチャンプル** |

図 3. 同位単語抽出判定された単語の具体例

して,「アラビアータ」があった場合，Wikipedia 内の「アラビアータ」の記事中に「イタリア料理」のリンクがあると考えられる.

## 4. プロトタイプシステム

我々は提案手法を用いて，コンテンツホール検索のプロトタイプシステムを PHP にて開発した．ここで Web 検索には Yahoo!API を使用した． 図 1，2 に示すようにプロトタイプシステムでは，ユーザが 4 つのクエリを入力できるフォームがある．内 1 つは単語 X で，もう 3 つは X と part-of 関係にある単語を入力する．ユーザが「Submit」のボタンをクリックすると，単語 X の記事の冒頭部分と X との同位単語候補をリスト表示される．それらのリストから，単語 X と同じ要素をもつ同位語が提示される.

## 5. 評価実験

プロトタイプシステムを用いて，評価実験を行った．我々は 71 件の構造化されたキーワードについての検証を行い，それらの検索結果の有用性について以下のようにまとめた.

- 上位単語に関して，全体の 71 語に対して，正解数が 45 で，適合率は 63.3%であった．表 1 にその結果の詳細を示す.
- 単語 X と概念構造が同じでかつ同じ要素を含む単語に関しては，単語取得数 94 に対して，正解数 50，適合率は 53.1%であった.

具体的な抽出例を図 3 に示す．このように，コンテンツホールを抽出できたが，問題点として以下のことが挙げられる.

- 上位単語の取得ができなければ，最終結果を得られない.
- 上位単語抽出の際，複数個の上位単語が存在した場合についての対応できていない.
- 同位単語抽出の際，複合語に対応できていない.

## 6. まとめと今後の課題

本論文ではユーザが「思いつかない情報」を概念構造を用いて上位単語，同位単語を抽出し，周辺話題のコンテンツホール検索の発見，提案を行った．我々は Yahoo!JAPAN の Web 検索エンジンのスニペット結果と Wikipedia の記事を解析した．利点として，不特定多数の WWW 上の単語に対応し，かつ，リアルタイムに構築を行うので即時性に優れている．また，クエリに概念構造を持たせたことで，ユーザの目的に即した結果を抽出することができた.

今後は，これら Web 検索エンジンや Wikipedia を用いた，概念抽出手法やそれに基づく周辺話題の抽出の精度向上の検証と改良が必要である．またこれらを用いた他のコンテンツホール検索についても，検討していく予定である.

## 謝辞

本研究の一部は，平成 20 年度科研費特定領研究域「Web2．0 時代のコミュニティ型コンテンツのコンテンツホール検索に関する研究」(課題番号：19024072，代表:灘本明代)による．ここに記して謝意を表します.

## 文 献

[1] 灘本明代，荒牧英治，阿辺川武，村上陽平,"Wikipedia を用いたコンテンツホール検索の提案"，社団法人 情報処理学会 研究報告，pp.146-44,2008

[2] 芳鐘冬樹,野澤孝之辻慶太,影浦峡," ウェブからの関連語・下位語の収集手法の検討と検索システムへの応用"

[3] 大塚真吾,豊田正史,喜連川優,"大域ウェブアクセスログを用いた関連語の発見法に関する一考察", 情報処理学会論文誌，Vol.46 No.SIG8，pp.82-92，2005

[4] EDR 電子辞書 http://www2.nict.go.jp/r/r312/EDR/J_index.html

[5] 新里圭司，鳥澤健太郎,"HTML 文書からの単語間の上位下位関係の自動獲得"，情報処理学会研究報告，pp. 95-102, 2003.

[6] 大島裕明，山口雅史,小山聡,田中克己,"Web 検索とクエリログを併用した同位語発見手法",日本データベース学会 Letters,Vol.5,No.4, pp. 37-040(2007).

[7] 中山浩太郎,西尾章治郎,原隆浩,"自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築",日本データベース学会論文誌， Vol.7, No.1, pp.67-7