Wikipedia リンク構造の関連度による用語抽出及び用語問題の自動生成

菅　亮太[†]　　徐海燕[†]

WWW の普及に伴い、Wikipedia に代表される Web 百科事典が公開され、文化、歴史、科学、社会、学問などの幅広い分野をカバーしている Wikipedia データを用いた様々な研究が行われている。Wikipedia の豊富なデータを活用するために、われわれは、用語知識の抽出およびアプリケーションへの適用を研究目的とした。Wikipedia の特徴であるリンク構造とカテゴリ構造のデータを利用した上で各構造の分析・改良を行い、概念（記事）同士の関連度を計算することで用語の抽出を行う。さらに、関連度の利用として用語問題の自動生成による e-Learning システムの開発を行っている。本発表はそれらについて報告する。

Extracting Related Terms and Automatically Constructing Exercises on Terminology by Wikipedia Link Structure

Ryota Suga[†]　and　Haiyan Xu[†]

With the spread of WWW, Wikipedia, a collaborative Web-based encyclopedia, has been published. Wikipedia data covers a wide-range of human knowledge such as Culture, History, Science, Social, and Academic. A considerable number of researches on Wikipedia mining have been carried out. In order to make use of the huge data of Wikipedia, this study aims to extract related terms and automatically constructing exercises on Terminology. Using category structure and link structure which is the characteristic of Wikipedia, we extract related terms and calculate the relatedness between terms. Furthermore, we are developing an e-Learning system which automatically constructs exercises with the related terms.

1. はじめに

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、科学、社会、学問、自然、技術、地理などの幅広い分野の概念をカバーしている。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編集することができることが大きな特徴である。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 Nature 誌の調査で報告されている[6]。

Wikipedia などの Web 事典と通常の電子事典の最大の違いは、記事どうしがハイパーリンクでお互いに参照していることである。Wikipedia の持つリンク構造は、近年知識抽出のために研究者から注目を集めており、Wikipedia を用いた様々な研究が行われている[2, 3, 4, 5, 7, 8, 9, 10]。概念間関連度の測定において、概念の網羅性を向上させることができ、自然言語処理における未知語の対応、同義語や多義語の判別など、これまでの自然言語処理手法における課題を意識せずに解析を行うことができる。Wikipedia の特徴を利用し、Wikipedia から用語の知識抽出を行う。

本研究では、Wikipedia のカテゴリリンク構造の特徴分析を行い、先行研究[15]を用いてカテゴリリンクの活用をはかる。さらに、先行研究によるカテゴリリンク利用の欠点を補う手法を提案し、先行研究と提案手法の比較を行う。さらに、カテゴリリンクの利用における問題点を挙げ、改善方法を提案する。

本稿は、次のように構成される。２章では、Wikipedia の特徴について述べる。３章では、Wikipedia のリンク構造の分析を行い、概念間関連度の測定を行う。3 章の概念間関連度計算に対する評価実験について、４章で報告する。５章はアプリケーションへの適用について報告し、６章は全体のまとめである。

2. Wikipedia の特徴と関連研究

Wikipedia（ウィキペディア）は、ウィキメディア財団が運営するインターネット百科事典である。Wikipedia は、閲覧によって情報を得るという活用以外に、研究者にとっては機械処理によって知識抽出を行う対象として注目されている。Wikipedia から知識抽出する際に有効な特徴を以下に示す。

1. 質の高いリアンカーテキスト
2. コンテンツの網羅性
3. 密なリンク構造
4. 多言語間のリンク

[†] 福岡工業大学大学院工学研究科情報工学専攻
Fukuoka Institute of Technology Graduate School/Master's Course/Computer Science and Engineering

5. URL による概念の一意性
6. カテゴリリンク構造
7. リダイレクトリンク

2011 年 6 月 28 日の段階で Wikipedia（日本語版）の記事数は約 121 万記事である。この記事の記事間リンク数は、約 5454 万であることが分かっている。これは、1 つの記事あたり平均 44.9 のリンク数を持っている。これらのリンクはサイト内に対するリンクのみをカウントしたものであり、サイト外へのリンクは含まれてない。これは、Wikipedia では閉じられた空間の中で密なリンク構造を持っており、リンク構造を解析することで有用な情報を抽出できる可能性が高いことを示している。

Wikipedia の特徴の 1 つとして言語間のリンクがある。Wikipedia は 2011 年 6 月現在、日本語、英語、中国語、ドイツ語など 283 言語で展開されている。次に、カテゴリリンクは、ある記事（概念）がどのようなカテゴリに属するかを指定するためのリンクである。カテゴリには専用のページ（カテゴリページ）があり、カテゴリページはさらに別のカテゴリページに属することが可能である。Wikipedia のカテゴリ構造は、実際にはネットワーク構造となっている。

Wikipedia に関する研究は、大きく 2 つの分野に分類できる。1 つは、Wikipedia を社会現象として解明する研究である。たとえば、Wikipedia に参加する人の目的や行動を調査し、社会現象として Wikipedia を解析するといった研究である。もう一方の研究分野は、Wikipedia を言語リソースとして利用や分析をする研究である。たとえば、記事（概念）間の関係性などの有用な情報を抽出し、アプリケーションに適用する研究がこの分野に分類される。Wikipedia を解析して概念間関連度を測定する先行研究として、大きく分けると記事間リンクに基づく手法、記事内テキストに基づく手法、カテゴリリンクに基づく手法がある。

3. Wikipedia リンク構造に基づく概念間関連度

3.1 Wikipedia カテゴリリンク構造の分析

Wikipedia は、記事（概念）をカテゴリに所属させることで管理している。Wikipedia のカテゴリのリンクおよび所属記事との関係の構図を図 1 に表す。Wikipedia の構造は木構造ではなく、循環を含むなど複雑な構造となっているため、タイトル語の分類のためにカテゴリのリンク構造を用いる際に、あるタイトル語に付与されたカテゴリのリンクをどの範囲まで取得すべきか明確ではない。そして、無作為なカテゴリのリンク構造抽出は、語の分類を行う際にノイズとなる恐れがある。具体的にノイズとは、図 1 の双方向リンクにあたる。これは、Wikipedia のカテゴリ構造は一部お互いのリンク先を参照し合うためループ構造が発生してしまうため、カテゴリリンクの経路をたどる際に問題となってしまう。

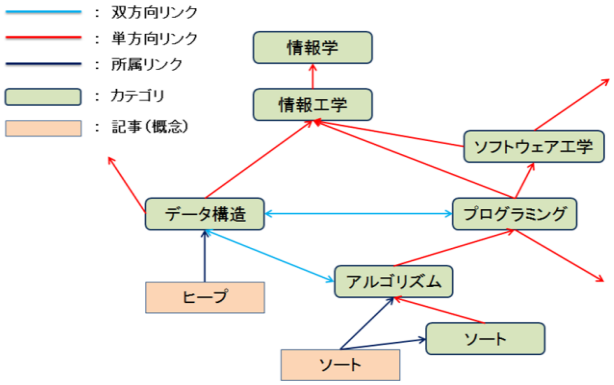


図 1 Wikipedia カテゴリ構造の例

3.2 Wikipedia カテゴリ構造の再構築

Wikipedia のカテゴリ構造はネットワーク構造であり、分類学のように厳密には整っておらず、3.1 節で述べたようにカテゴリ間をループするパターンが存在する。このカテゴリ構造をデータ利用するためにツリー構造へ変換し活用する新井らの研究がある[12]。新井らは、Wikipedia のカテゴリは複数の親カテゴリを持つことができるため、各カテゴリについて、トップカテゴリから最短経路による下位カテゴリ情報取得によるカテゴリの階層化によってツリー構造を構築した。

本節では、新井らが用いたカテゴリ構造の最短経路によるツリー構造への変換で構築したデータを検証し、新たに提案手法として最短経路によって欠損したデータを取得するためにカテゴリ階層構造の再構築を行う。

3.2.1 最短経路によるツリー構造への変換

Wikipedia のカテゴリリンクの概念図を図 2(a)に示す。Top カテゴリの直下にカテゴリ A とカテゴリ B があり、カテゴリ A の子カテゴリとして、カテゴリ C がある。また、カテゴリ C とカテゴリ B の子カテゴリとしてカテゴリ D がある。この場合、Top カテゴリからカテゴリ D に至る経路は、A－C－D、B－D の 2 経路あり、それぞれ Top カテゴリからの深さが、3 と 2 で異なってしまう。各カテゴリについて、図 2(b)のように Top カテゴリまでの経路が最短となる親カテゴリを選択することで、ツリー構造を構築している。

3.2.2 提案手法によるカテゴリツリー構造の改良

新井らによるカテゴリ構造の階層化は、図 2(b)で示したように Top カテゴリからの最短経路によってルートを選択している。しかし、最短経路を取ることによって本来得られる情報が欠損する問題がある。つまり、最短経路を取る場合、カテゴリ D はカテゴリ B の間との関係のみになってしまい、カテゴリ D とカテゴリ C 間の関連性が失われてしまう。

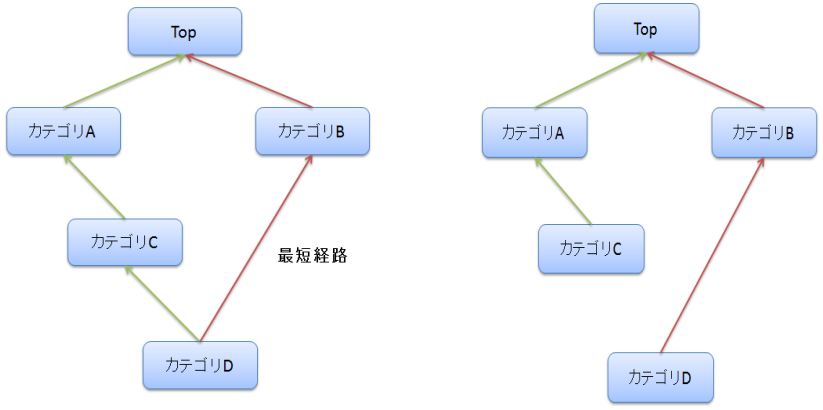


図 2　Wikipedia のカテゴリリンクとツリー構造への変換

提案手法では、最短経路による手法で欠損した関連性を保つために、カテゴリ間を親子関係で表わすことで問題の解決を図る。図 3 に示すようにカテゴリ C とカテゴリ D の関係を保持することで欠損したデータを補う。つまり、カテゴリ D には、カテゴリ C との親子関係、カテゴリ B との親子関係をそれぞれ持たせる。

3.2.3 提案手法でのカテゴリの再構築

新井らの手法を Wikipedia のカテゴリデータに適用し Category_Hierarchy_Short テーブル(最短経路テーブル)を構築した。このテーブルのスキーマを表 1 に示す。最短経路テーブルには、カテゴリを階層化し得られた、カテゴリ ID、カテゴリ名、親カテゴリ名、所属階層位置が格納されている。さらに、最短経路テーブルのデータ妥当性を判断するために、新井らの構築したカテゴリ階層データとの比較を行った。

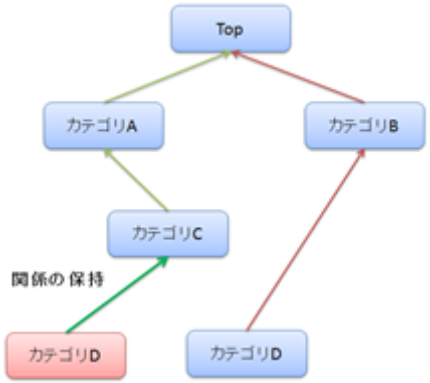


図 3　提案手法

表 1　Category_Hierarchy_Short テーブルのスキーマ

フィールド	データ型	概要
h_id	int unsgined	pageテーブルのpage_idと紐付く。主キー
h_title	varchar(255)	カテゴリ名
h_belong	varchar(255)	所属カテゴリ名。(親カテゴリ)
h_layers	int signed	所属階層位置

表 2　カテゴリ構造データ

階層	カテゴリ数
1	9
2	259
3	2,089
4	6,813
5	12,399
6	11,275
7	5,787
8	3,448
9	230
10	50

(a)

階層	カテゴリ数
1	8
2	241
3	2,499
4	9,084
5	20,011
6	21,219
7	9,774
8	4,623
9	406
10	39
11	4

(b)

階層	カテゴリ数
1	8
2	241
3	2,623
4	10,883
5	26,668
6	39,181
7	34,016
8	22,674
9	7,066
10	993
11	92
12	24

(c)

新井らが構築したカテゴリ階層データは、日本語版 2007 年 10 月 13 日のダンプファイルであり、今回使用したデータは、日本語版 2011 年 6 月 28 日のダンプファイルである。Wikipedia 全体のデータとして 2007 年に 430,344 件、2011 年に 756,699 件と推移している。表 2(a) 新井らによるカテゴリ階層データ と表 2(b) 構築したカテゴリ階層データ からデータの分布を表 3.4 に示す、非常に酷似した分布が見られる。階層数が増加した点は、データ増大によるカテゴリの細分化が行われた結果である。以上のことから、構築したカテゴリ階層データは妥当だといえる。表 2(c)は提案手法での結果である。

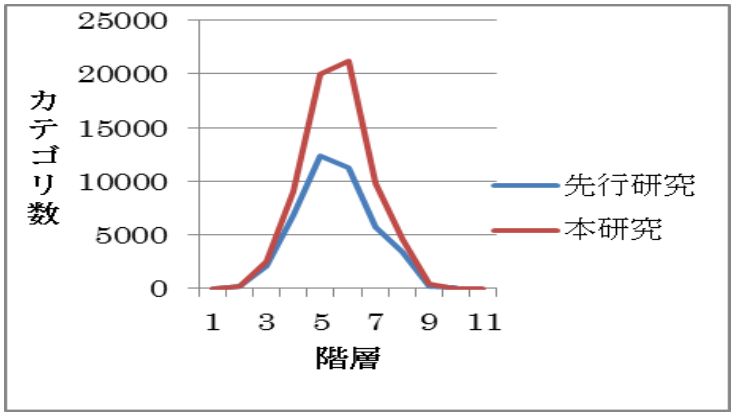


図 4　先行研究と本研究のカテゴリ階層データ比較

3.2.4 評価実験

本実験の目的は、提案手法によって作成されたカテゴリツリーと先行研究手法によるカテゴリツリーの Lch[8]計算による値の評価をすることである。Wikipedia から概念間関連度を測定し、解析時間の比較を行う。また、提案手法によるカテゴリツリーが先行研究手法より多くの情報が得られるか比較する。

解析対象の Wikipedia のデータとしては、2011 年 6 月時点の日本語版 Wikipedia のデータからノイズ記事を除去した。ノイズ記事の定義は、杉原ら[14]による Wikipedia の管理用カテゴリを除外するフィルター条件（表）にしたものである。

表 3　ノイズ記事の定義

カテゴリの文字列に関する条件	先頭が「User_」
	末尾が「の画像」
	末尾が「ユーザー」
	末尾が「スタブ項目」
	末尾が「のスタブ」
	末尾が「のテンプレート」
	末尾が「ウィキペディアン」

対象データの準備

実験で使用するデータは、日本語版(2011 年 6 月 28 日)Wikipedia のダンプファイルである。今回用いたファイルは、pages-articles.xml(記事本文データ)と categorylinks.sql (カテゴリのリンク情報)で、MySQL にデータベースを構築している。

測定方法

データベースに作成している最短経路によるカテゴリツリーテーブル（Category_Hierarchy_Short テーブル）と経路網羅によるカテゴリツリーテーブル（Category_Hierarchy テーブル）を用いて各テーブルに対して、page テーブルからランダムに抽出した 2 つの概念（記事）の関連度と解析時間、共通の親が現れる階層の差を測定する。関連度の計算は、Strube らのカテゴリ間関連度を測定する Lch 手法を用いる。

測定するデータの数は、カテゴリの全体数が、Category_Hierarchy_Short テーブルで 67,909 件、Category_Hierarchy テーブルで 144,470 件あるためサンプル数を以下の式(1.1)によって求める。また、2 つの手法による共通カテゴリが出現する階層の比較を行う。

$$n = \frac{N}{\left[\left(\frac{\varepsilon}{\mu(\alpha)}\right)^2 \times \left\{\frac{N-1}{\rho(1-\rho)}\right\} + 1\right]} \qquad (1.1)$$

n	必要サンプル数
μ(α)	信頼度100－αのときの正規分布の値。今回αは信頼度95%の1.96。
N	調査したい母集団の大きさ。
ε	精度。今回は上下3%とした。
ρ	母比率。今回は0.5とした。

図 5　式(1.1)の詳細

実験結果

本項では、先行研究による手法と提案手法の処理時間の比較と Lch 計算における数値の差異についてそれぞれ解説し考察する。解析には、上述の手順で構築したデータを利用して行った。

解析処理時間

ここで評価する解析処理時間は、2 つの概念ペアが与えられ、即時にその概念ペアの関連度を算出できる状態までカテゴリ階層テーブル全体を解析するまでの時間である。表 4 に、解析にかかった処理時間の平均を示す。

表 4　解析に要する時間

手法	平均処理時間(sec)
先行研究手法	0.766165191
提案手法	3.598486347

両手法の比較

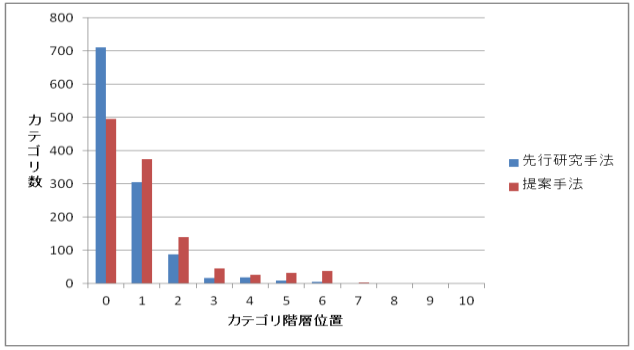
ここで評価する比較は、サンプルデータの結果から得られた各手法の Lch の値であり、提案手法からみた先行研究手法との値の差をみる。両手法による Lch 計算の結果は、サンプル数 1050 件中 214 件で値の差があらわれた。Lch の値は、平均で 0.116765 の差であった。このことから、提案手法は全体の 2 割程のデータに対してより概念間の関係を表すことができたと考えられる。

表 5　両手法の差

概要	結果数
両手法による差が表れたカテゴリ数	214
両手法によるLchの平均差	0.116765

また、両手法の共通カテゴリの位置が現れる階層の比較を行った。これは、階層数が小さいほど早い段階で共通カテゴリを見つけていることが分かる。図から提案手法は、先行研究手法より早い段階で共通カテゴリを発見し、より関係性が強い共通カテゴリを選択していることが分かった。

表 6 各手法における共通カテゴリ位置



4. カテゴリリンク間関連度計算による問題とその改良法

記事（概念）が同一カテゴリに所属する場合、親カテゴリの位置が同じであるためLch による関連度計算の数値に差がないという問題がある。

一方、LSP 法とは、記事のリード部分（冒頭文）を重要文と見做して解析する手法である[13]。これは、Wikipedia の各記事において、リード部分が多くの場合に他の概念との明確な意味関係を定義した文であることを利用した手法である。特に、Wikipedia におけるリード部分は、ほかの概念に対する is - a 関係が豊富に定義されていることが中山らの調査によって判明している。

LSP 法を利用した LSPLch による関連度計算

我々は、LSP 法を用いて同一カテゴリの Lch 計算結果の問題点を解消することを試みる。提案手法 LSPLch では、LSP 法の考え方を基に Wikipedia の重要文を冒頭の概要部分と定め、重要文に含まれるハイパーリンクを計算の対象概念とした。2 つの概念(記事)A,B から以下に式(1.2)を示す。

概念 A のハイパーリンク $=\{a_1, a_2, \cdots a_{n-1}, a_n\}$

概念 B のハイパーリンク $=\{b_1, b_2, \cdots b_{m-1}, b_m\}$

$$\mathrm{LSPLch} = \frac{1}{2}\left(\frac{1}{n}\sum_{k=1}^{n}\mathrm{lch}(a_n, B) + \frac{1}{m}\sum_{l=1}^{m}\mathrm{lch}(b_m, A)\right) \quad (1.2)$$

概念 A のハイパーリンクは、概念 A に関する説明用語であると考え、対象の概念 B と Lch 計算することで概念 A との関係性の指標を増やす目的がある。同じように、概念 B のハイパーリンクも概念 A に対して Lch 計算を行う。各リンク数の平均をとり足し合わせ、最後に全体で割ることで LSPLch の値とする。

適用例

カテゴリ「アルゴリズム」に属している概念（記事）に対して、LSPLch 計算を行う。対象ページは、モンテカルロ法を軸に 9 個の記事、1. 数値積分、2. 粒子フィルタ、3. ページ置換アルゴリズム、4. 高速フーリエ変換、5. トポジカルソート、6. ベクトル空間モデル、7. 最良優先探索、8. ファジィ集合、9. K 平均法を対象にした。

表 7　LSPLch による結果

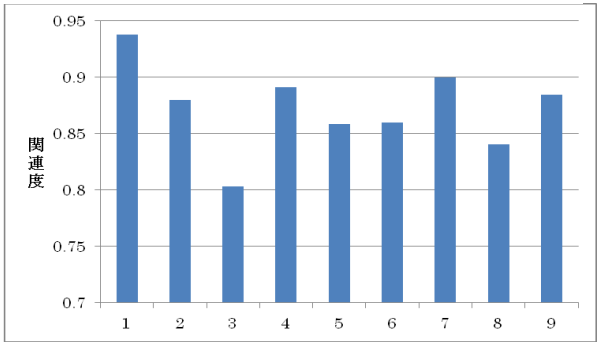


表 7　LSPLch による結果表 7 から、同一カテゴリ内における各概念間の差別化が見て取れる。この表から特に注目したいのが、3 番目のページ置換アルゴリズムである。対象ページはモンテカルロ法で、関係する概念は数値解析や乱数列であり、ページ置換アルゴリズムが関係する概念は、メモリ管理やページング方式といったややモンテカルロ法とは関連性が薄い印象を受ける。結果から見ても、数値として表れているのが分かる。

しかし、記事によってハイパーリンクの使い方に差があり、影響を受けやすい点が問題である。

5. アプリケーションへの適用

アプリケーションへの適用として､用語問題に関する e-learning システムを構築する。出題問題は、Wikipedia のデータを用いて行う。問題文は、概念（記事）の概要を説明している冒頭の部分を用いて、出題用語名を空欄としている。解答方式は、四択である。

関連度計算を e-learning システムへ適用する点は、ダミー解答の部分である。正答の用語を対象として、その用語と意味的に近い用語の候補選びに用いる。アプリケーションへの利用としては、関連度処理時間の関係からまだ問題が多く残る。

6. まとめ

本研究では、Wikipedia のリンク構造のカテゴリリンクに着目をし、概念間関連度計算を行った。先行研究によるカテゴリ構造のツリー化の問題点を取り上げ、提案手法によってカテゴリ構造の再構築を行い、より概念間の関連性を得ることができた。また、同一カテゴリに所属する概念（記事）の差別化を図るため、LSP 法の考えを基にした LSPLch 手法を提案し、Wikipedia の重要文部分からハイパーリンクを取得し、サンプルを用いて考察を行った。その結果、同一カテゴリ内の記事を差別化することができたが、記事の質によって左右される問題が課題である。

参考文献

[1] Bray, T., Paoli, J., and Sperberg-McQueen, C.M. : Extensible Markup Language (XML), The World Wide Web Journal, Vol.2, No.4, pp.27-66 (1997).
[2] 伊藤 雅弘：Wikipedia を用いた概念間の関連度測定に関する研究, Osaka University Knowledge Archive(OUKA) (2011).
[3] 森竜也, 増田英孝, 清田陽司：Wikipedia を活用した言語間差異比較システムの提案, DEIM Forum 2010, A5-5 (2010).
[4] 玉川奨, 桜井慎弥, 手島拓也, 森田武史, 和泉憲明, 山口高平:日本語 Wikipedia インフォボックスからのプロパティ自動抽出, the 24th Annual Conference of the Japanese Society for Artificial Inelligence (JSAI 2010), 2I3-NFC4-3 (2010).
[5] 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 47(10), 2917-2928, 2006-10-15.
[6] Giles, J. : Internet Encyclopedias Go Head to Head, Nature, Vol.438, pp.900 - 901 (2005).
[7] Gabrilovich, E., and Markovitch, S. : Computing Semantic Relatedness Using Wikipedia^based Explicit Semantic Analysis., in Proceedings of International Joint Conference on Artificial Intelligence(IICAI 2007), pp. 1606-1611 (2007).
[8] Strube, M., and Ponzetto, S. P. : WikiRelate! Computing Semantic Relatedness Using Wikipedia, in Proceedings of the American Association for Artificial Intelligence (AAAI 2006), pp. 1419-1424 (2006).
[9] 山崎由佳, 井庭崇, 熊坂賢次：Wikipedia における編集者の活動分析, 第 21 回セマンティックウェブとオントロジー研究会（第 2 回 Wikipedia ワークショップ）(SIG-SWO), A901-01 (2009).
[10] 鈴木優, 吉川正俊：Wikipedia におけるキーパーソン抽出による信頼度算出精度および速度の改善, 第 21 回セマンティックウェブとオントロジー研究会（第 2 回 Wikipedia ワークショップ）(SIG-SWO), A901-01 (2009).
[11] Torsten Zesch, Iryna Gurevych. : Analysis of the Wikipedia category graph for nlp applications, in Proceedings of the Workshop TextGraphs-2:Graph-Based Algorithms for Natural Language Processing at HLT-NAACL 2007, pp. 1-8 (2007).
[12] 新井 嘉章, 福原 知宏, 増田 英孝, 中川 裕志：Wikipedia の言語間リンクに関する分析, 第 22 回人工知能学会 全国大会 (JSAI 2008), 2D3-02 (2008).
[13] 中山浩太郎, 原隆浩, 西尾章治郎：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に関する一手法, 電子情報通信学会 第19回データ工学ワークショップ, A3-2 (2008).
[14] 杉原 大悟, 増市 博, 梅基 宏, 鷹合 基行：Wikipedia カテゴリ階層構造の固有名詞分類実験における効果, 情報処理学会研究報告. 情報学基礎研究会報告 2009(2), 57-64, 2009-01-15
[15] 新井 嘉章, 福原 知宏, 増田 英孝, 中川 裕志：Wikipedia を用いた多言語情報アクセスに関する研究：言語間リンクの分析と応用, 第 20 回セマンティックウェブとオントロジー研究会, pp.SIG-SWO-A803-15(2009).

【 この位置に改ページを入れ，以降のページを印刷対象外とする 】