

手順の対応関係に基づくハウツー情報の階層化

湯本 高行[†]

[†] 兵庫県立大学大学院工学研究科 〒 671-2201 兵庫県姫路市書写 2167

E-mail: tyumoto@eng.u-hyogo.ac.jp

あらまし 料理のレシピなどのハウツー情報を検索する需要は高いが，ハウツー情報に特化した検索・閲覧手法は実現されていない．そこで本研究では，ハウツー情報の検索結果を，手順の対応関係に基づいて階層化する手法を提案する．ここで手順とはハウツー情報を構成する単位であり，動作とその対象のペアとして表現する．2つのハウツー情報の間で手順の対応関係を求め，これに基づいて関連度を計算する．関連度は共通度，被覆度，ばらつき度を用いて定義される．この関連度の大小関係に基づき，ハウツー情報を階層化する．この階層化では関連するハウツー情報の集合が1つのツリーを構成し，深い階層にあるハウツー情報ほど詳しい情報を含む．

キーワード ハウツー情報，手順，情報検索，情報抽出

Takayuki YUMOTO[†]

[†] Graduate School of Engineering, University of Hyogo

2167 Shosha, Himeji, Hyogo, 671-2201 Japan

E-mail: tyumoto@eng.u-hyogo.ac.jp

1. はじめに

近年，Web 情報検索は一般的になり，その用途も多様化している．中村らの 2006 年の調査によると，検索した時に知ろうとしているものの中で「方法・手段」は「検索キーワードの意味・定義」や「もの」に次いで多く，1000 人中 575 人がよく検索すると回答している [1]．すなわち，方法や手段などのハウツー情報に対する需要は高いと言える．

これに対して，ハウツー情報のうち，レシピなど一部のドメインに特化した検索エンジンやハウツー情報のみを登録できるサイトなどは存在するが，ハウツー情報を自動的に収集したり，閲覧を効果的に行う手法については不十分な状況にある．野中らはハウツー情報を効率的に発見するために，Web ページからハウツー情報を抽出し，見やすさによってランキングする手法を提案している [2]．しかし，この手法では，同じキーワードを含むハウツー情報でも内容が異なるものが含まれる可能性があり，1次元のランキングでは不十分であると考えられる．

そこで本研究では，ハウツー情報の検索結果に対して，関連のあるハウツー情報を階層的にまとめてツリーのリストとして提示する手法を提案する．作成するツリーでは，深い階層にあるハウツー情報ほど説明が詳しい．ハウツー情報間の関連は，ハウツー情報の構成要素である手順の対応関係によって判定する．なお，手順は動作とその対象のペアとして表現する．

2. 関連研究

我々の研究では，ハウツー情報検索・提示は以下のような段階に分けて実行する．

- (1) ハウツー情報の収集・抽出
- (2) ハウツー情報のモデル化
- (3) ハウツー情報の提示

それぞれに対して関連研究をまとめる．

2.1 ハウツー情報の収集・抽出

検索対象を特定のサイトに限定しない場合には，まずハウツー情報の収集が必要になる．多くの場合，Web 検索を行うことになるが，検索クエリによっては，ハウツー情報以外に本の目次やショッピングサイトのページが含まれたり，むしろハウツー情報以外のページが占める割合の方が多くなってしまうことがある．これに対して，あらかじめドメインを限定することができれば，小久保らの検索隠し味 [3] のように機械学習を用いることによってハウツー情報を多く含むようにクエリ拡張を行い，ハウツー情報の収集効率を向上させることができる．これに対して，野中らはハウツー情報の抽出を同時に行うことでハウツー情報の判定も行う手法を提案している [2]．

また，ハウツー情報の抽出についてであるが，汎用的な HTML 文書からの本文抽出 [4], [5] を利用する方法が考えられる．しかし，Web ページによってはブログなどのように本文中にハウツー情報以外を含む場合も存在するため，ハウツー情報に特化

した抽出手法が必要になる。これについては、野中らのように手がかり表現を用いた手法 [2] や小澤らのようにこれに加えて機械学習を併用した手法 [6] などが提案されている。本研究では、ハウツー情報の収集と抽出の両方を行うため、野中らの手法を利用する。

2.2 ハウツー情報のモデル化

ハウツー情報をモデル化する方法であるが、これに関しては、レシピについて多くの研究がなされている。Miyamori はレシピなどのアイテムの妥当性を評価するために、それを構成する材料とその分量によってアイテムをモデル化し、それを基準と比較する手法を提案している [7]。この手法では、基準はユーザが事前に与えるとしており、ユーザに知識が要求される。また、浜田らはレシピから、素材 (本研究で「対象」と呼ぶもの) と動作をノードとする、データフローグラフを生成する手法を提案している [8]。この手法では、レシピに特化した辞書などが必要となるため、ハウツー情報全般に用いることが難しい。これに対して我々は動作と対象のペアのリストで表現する方法を提案しており [9]、本研究でもこれを用いる。

2.3 ハウツー情報の提示

ハウツー情報の提示方法としては、まずランキングが考えられる。ハウツー情報に対して有効なランキングの尺度としては、わかりやすさや見やすさ [2], [10], [11] を利用する方法が考えられる。しかし、1つの目的に対して複数の方法が存在する場合は、ランキング中に方法の異なるハウツー情報が混在してしまう。これに対して、検索結果のクラスタリング [12] によって内容の近さに基づき、検索結果を整理して提示する方法がある。本研究では内容の近さだけではなく、内容の対応関係と詳しさに応じて階層化を行う。

3. ハウツー情報の階層化

3.1 ハウツー情報の抽出

Web 検索を行い、得られた Web ページに対して、野中らの手法 [2] によってハウツー情報の抽出を行う。我々はハウツー情報を手順のリストとして表現する。また、手順を動作とその対象のペアとして表現する。ここで、ハウツー情報を h 、手順を $s_i (i = 1, \dots, n)$ 、手順 s_i の動作と対象をそれぞれ a_i, t_i とすると、以下のような関係になる。

$$h = \langle s_1, s_2, \dots, s_n \rangle \quad (1)$$

$$s_i = (a_i, t_i) \quad (2)$$

手順は、文献 [9] で提案したアルゴリズムを改良し、以下ののように抽出する。

(1) 文への分割：句点を基準に文に分割する。この段階では重文を残したままにする。

(2) 説明文かどうかの判定：文末の表現に注目して、過去の経験について述べた文、状態について述べている文、感想などを述べている文、疑問文などを排除することで消去法で説明文のみを抽出する。

(3) 動作と対象の抽出：説明文を形態素解析した結果を文の末尾から分析し、品詞のパターンから動作と対象を決定する。

自立動詞または「する」の直前の名詞を動作とする。また、格助詞の「は」または「を」が続く名詞またはそれらの名詞と並列関係にある名詞を対象とする。さらに、「を」が続く名詞よりも「は」が続く名詞を優先することとする。なお、この段階では、対象は集合として扱う。

(4) 省略された対象の補完：(3) の処理で対象が取得できない場合は、以下の処理をする。

- 重文を構成する文のうち、後ろに部分から抽出された手順の場合は、前の部分から抽出された手順の対象によって補完する。たとえば、「野菜を切って、炒める」という文は、「野菜を切って」と「炒める」からなる重文である。この場合、「炒める」の対象は「野菜を切って」の対象である、「野菜」になる。

- 上記以外の場合は、この段階では対象を確定させない。

(5) 手順の分割：対象の部分が集まっている手順を、対象が集合でないような手順に分割する。たとえば、(炒める, {肉, 野菜}) という手順を、(炒める, 肉), (炒める, 野菜) の2つの手順に分割する。

3.2 ハウツー情報間の関係の発見

3.2.1 手順間の対応関係

動作と対象がそれぞれ一致していれば、2つの手順は対応しているとみなす。動作間または対象間には以下のいずれかの条件を満たしていれば一致しているとみなす。

- 形態素解析した結果の読みが同じである。
- 日本語 WordNet [13] で同じ概念に分類される。
- Wikipedia で対象とする語と一致するリンクアンカーから同一のページにリンクしている。

これらの条件を同義語判定の条件と考えた場合、再現率は高いが適合率は低いことが予想される。たとえば、白川ら [14] は Wikipedia データ上でリンクされる頻度を考慮することによって異義語の判定を行っており、これを応用することで同義語の判定を行うことも考えられる。しかし、本研究では、判定はハウツー情報の手順を構成する語に対して限定的に行われるため、同音異義語などによる問題が起りにくいと考えている。一方、ハウツー情報では表記ゆれが問題となることが非常に多いため、再現率を重視して、上記の条件を採用することとする。

また、「入れる」、「足す」、「加える」、「追加する」についてはハウツー情報においては同じ意味で使われることが多いため、動作がこれらのいずれかの語どうしであれば、一致しているとみなす。

さらに、ハウツー情報の抽出の際に対象が確定していない手順についてはその手順より前に出現した対象のいずれかであれば一致するとみなす。たとえば、 i 番目の手順の対象 t_i が確定していない場合は、対象 $t_1, \dots, t_{i-1}$ のいずれかと一致するかどうかで対応関係を判定する。

3.2.2 対象の類似性

与えられた2つのハウツー情報が同一のものであることを保証するため、使用している対象がどの程度類似しているかを事前に検証する。ここでハウツー情報 h を構成する対象の集合を T_h で表す。ハウツー情報 h_1, h_2 の対象の類似度を以下のよう

$$sim(h_1, h_2) = \frac{|T_{h_1} \cap T_{h_2}|}{\max(|T_{h_1}|, |T_{h_2}|)} \quad (3)$$

$|T|$ は T に含まれる対象の数である．ここでは対象が大きく異なるものハウツー情報は別の目的であると考え、 $sim(h_1, h_2) < \theta_{sim}$ のときは h_1 と h_2 は別の目的であるとみなす．なお、 θ_{sim} は 0 以上 1 以下のしきい値である．

3.2.3 ハウツー情報間の関係

ハウツー情報間の関係を定義するにあたって、まず、2 つのハウツー情報の関連度を定義する．この関連度は方向を持ち、ハウツー情報 h_1 から見た h_2 との関連度の強さを表現する．また、以下の条件を満たす場合に値が大きくなると考える．

- (1) h_2 が h_1 と共通の手順を多く含む．
- (2) h_1 と共通の手順が h_2 に広く分布している．

これらのそれぞれを表現するための指標を定義する．

まず、前者を表現するために、 h_1 と h_2 の共通度 $com(h_1, h_2)$ を以下のように定義する．

$$com(h_1, h_2) = \frac{|h_1 \cap h_2|}{|h_1|} \quad (4)$$

ただし、 $|h_1|$ は h_1 に含まれる手順の数、 $|h_1 \cap h_2|$ は h_1 と h_2 に共通して含まれる手順の数である．たとえば、 h_1 が 4 つの手順からなり、そのうちの 3 つが h_2 に含まれていた場合、 $com(h_1, h_2) = 3/4 = 0.75$ である．共通度は 0 以上 1 以下の値をとり、この値が大きいほど 2 つのハウツー情報は共通していると言える．

続いて、後者に対応する指標であるが、被覆度とばらつき度の 2 種類を定義し、これらを組み合わせて用いる．被覆度 $cover(h_1, h_2)$ は h_1 に含まれる手順が h_2 にどの程度広く分布しているかを表す指標であり、0 以上 1 以下の値をとる．被覆度を定義するにあたって、ハウツー情報 h における手順 s_i, s_j の距離 $d_h(s_i, s_j)$ を以下のように定義する．

$$d_h(s_i, s_j) = |i - j| \quad (5)$$

なお、 i, j はハウツー情報 h での出現順を表す．これを用いて、被覆度 $cover(h_1, h_2)$ は以下のように定義する．

$$cover(h_1, h_2) = \frac{\max_{s, s' \in h_1 \cap h_2} d_{h_2}(s, s')}{\max_{s, s' \in h_2} d_{h_2}(s, s')} \quad (6)$$

なお、 $s \in h$ は手順 s がハウツー情報 h に含まれることを示し、 $h_1 \cap h_2$ は h_1, h_2 をそれぞれを手順の集合と見なしたときの積集合である．つまり、式 (6) の分子は h_1 と共通して含まれる手順が h_2 のどのくらいの範囲に分布しているかを表し、分母はその最大値によって正規化している．

ここで例を示す．図 1 はそれぞれの四角が h_2 を構成する手順を表し、そのうち色がついたものが h_1 にも含まれる手順であるとする．つまり、 h_2 が 6 つの手順からなり、そのうち、1 番目、4 番目、5 番目の手順が h_1 にも含まれていることを示している．この場合、 $cover(h_1, h_2) = 4/5 = 0.80$ である．

一方、被覆度が高くても、対応する手順が一箇所に集中して存在している場合は、2 つのハウツー情報は関連が弱いと考え

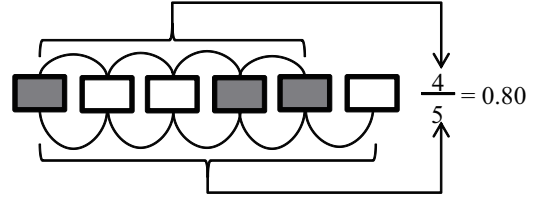


図 1 被覆度の例

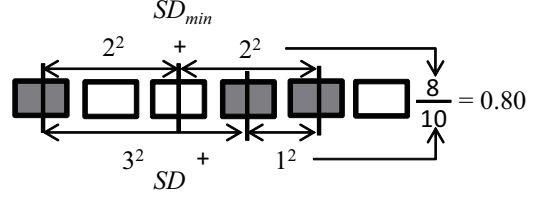


図 2 ばらつき度の例

られる．そこで、手順間のばらつきを表現するため、ばらつき度を定義する．ばらつき度は $disp(h_1, h_2)$ は h_1 に含まれる手順が h_2 でどの程度均等に広く分布しているかを表す指標である．0 以上 1 以下の値をとり、1 の時に最も均等に出現している（ばらついている）ことを表す．ばらつき度を定義するにあたり、2 乗距離 $SD(h_1, h_2)$ を以下のように定義する．

$$SD(h_1, h_2) = \sum_{i,j} \{d_{h_2}(s_i, s_j)\}^2 \quad (7)$$

ただし、 i, j は以下の 2 条件を満たすような任意の組合せである．

- (1) s_i, s_j は $i < j$ の順で h_2 に含まれると共に h_1 にも含まれる．
- (2) $i < k < j$ かつ $s_k \in h_1 \cap h_2$ を満たすような k は存在しない．

すなわち、2 乗距離は h_1, h_2 の両方に含まれ、かつ隣り合う手順の間の距離の 2 乗の総和である．この 2 乗距離は h_2 内で等間隔に出現する場合に最小となり、その値 SD_{min} は以下のように計算できる．

$$\begin{aligned} SD_{min}(h_1, h_2) &= \left(\frac{\max_{s, s' \in h_1 \cap h_2} d_{h_2}(s, s')}{|h_1 \cap h_2| - 1} \right)^2 \times (|h_1 \cap h_2| - 1) \\ &= \frac{\left(\max_{s, s' \in h_1 \cap h_2} d_{h_2}(s, s') \right)^2}{|h_1 \cap h_2| - 1} \end{aligned} \quad (8)$$

2 乗距離の逆数を $1/SD_{min}$ で正規化すると、0 以上 1 以下の値となる．この値をばらつき度とし、以下のように定義する．

$$disp(h_1, h_2) = \frac{SD_{min}(h_1, h_2)}{SD(h_1, h_2)} \quad (9)$$

図 2 を用いて説明する．この場合、 $SD(h_1, h_2) = 3^2 + 1^2 = 10$ 、 $SD_{min}(h_1, h_2) = 4^2/(3 - 1) = 8$ であるため、 $disp(h_1, h_2) = 8/10 = 0.80$ である．

関連度 $rel(h_1, h_2)$ はこれらの値の相乗平均として以下のように定義する．

$$rel(h_1, h_2) = \sqrt[3]{com(h_1, h_2) \times cover(h_1, h_2) \times disp(h_1, h_2)} \quad (10)$$

関連度には方向性があり、以下のような関係が考えられる。

- $rel(h_1, h_2) \geq \theta_{rel} \wedge rel(h_2, h_1) \geq \theta_{rel}$: h_1 と h_2 は同じ内容について説明している。
- $rel(h_1, h_2) \geq \theta_{rel} \wedge rel(h_2, h_1) < \theta_{rel}$: h_2 は h_1 より詳しい内容を説明している。

θ_{rel} はしきい値である。一方、どちらの方向からの関連度も低い場合、2つのハウツー情報の目的は異なっているとみなす。

3.3 階層化のアルゴリズム

ハウツー情報の詳しさに応じて階層化することで、検索結果をツリーのリストとして提示する。このツリーは、ノードを(複数の)ハウツー情報、エッジを詳細関係とみなし、同じ内容について説明しているハウツー情報を同一のノード、詳しい内容について説明しているハウツー情報に対応するノードを子ノードとするツリーである。このようなツリーのリストを $L = \langle R_1, R_2, \dots, R_m \rangle$ で表現する。ただし、 R_j はツリーとする。

野中ら[2]の手法によってハウツー情報を抽出し、3.1節の手法で手順を抽出したものに対して、ツリーのリスト L は以下のように作成される。

- (1) $L \leftarrow \langle \rangle$ とする。
- (2) 手順の数が少ないハウツー情報 h_k から順に (3) を実行する。
- (3) $L = \langle R_1, \dots, R_m \rangle$ の R_j の根 $root(R_j)$ に対して、式 (11) および式 (12) を満たすものを探し、(4) を実行する。

$$sim(h_k, root(R_j)) \geq \theta_{sim} \quad (11)$$

$$rel(h_k, root(R_j)) \geq \theta_{rel} \quad (12)$$

これらを満たすものが存在しない、もしくは $L = \langle \rangle$ の場合は、 h_k を根とするツリーを作成し、 L に追加する。

- (4) 以下の (a)~(c) を実行し、 R_j に対して h_k を追加する。
 - (a) 式 (13) を満たす場合は $root(R_j)$ に追加する。

$$rel(root(R_j), h_k) \geq \theta_{rel} \quad (13)$$

(b) 式 (13) を満たさない場合は、 $root(R_j)$ の子とす部分木 R'_j に対し、式 (14) を満たすかどうか調べ、満たすものについては R'_j に対して h_k を追加する処理として (4) と同様の処理を行う。

$$rel(h_k, root(R'_j)) \geq \theta_{rel} \quad (14)$$

(c) 式 (14) を満たす R'_j が存在しない場合は、 $root(R_j)$ の新たな子として h_k を追加する。

4. 実行例

実行例として、肉じゃがのレシピを提案手法によって階層化した結果について説明する。表 1 に URL、表 2 にタイトルを示す。手順の数の昇順に $h_1, h_2, h_3, h_4, h_5, h_6$ であった。なお、 $\theta_{sim} = 0.40$, $\theta_{rel} = 0.42$ とした。階層化の結果を図 3 に示

表 2 使用したページのタイトル

ID	タイトル
h_1	だしも水もいっさい使わない肉じゃが
h_2	洋風トマト肉じゃが
h_3	肉じゃがレシピ
h_4	だしいらすの肉じゃが
h_5	基本のおかず・肉じゃが
h_6	肉じゃがのレシピ・作り方

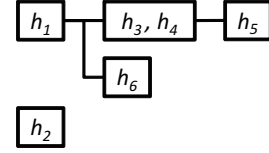


図 3 提案手法による階層化の結果

す。図 3 では、四角はノードを表し、1つ以上のハウツー情報に対応する。ノード間にはエッジが存在し、最も左側のノードが根となるツリーを構成する。 h_1 を根とするツリーと h_2 のみからなるツリーの2つからなるリストが作成された。 h_1 は簡略化されたつくり方で、 h_3, h_4, h_5, h_6 はいずれも h_1 よりも手のかかるつくり方であった。 h_3 と h_4 はどちらもオーソドックスなレシピであり、説明は類似していた。 h_5 にはこれらよりも詳しい説明があった。 h_6 は詳しい説明がされていたが、隠し味などが違っていた。また、 h_2 は洋風肉じゃがのレシピであり、他のものとは材料が異なっていた。

提案手法を用いなかった場合との比較結果について述べる。類似度による判定を行わず、関連度のみで階層化を行った場合は、 h_2 と h_3, h_4 の関連度が高くなり、目的が異なるハウツー情報が同一のツリーに混在してしまった。類似度を併用することによって、目的が異なるハウツー情報の混在は避けられる。また、関連度の代わりに共通度を用いて階層化を行った場合は、 h_3, h_4 間の共通度よりも h_5, h_6 間の共通度が高くなり、人の感覚と異なる結果になった。このように、共通度、被覆度、ばらつき度を併用することで人の感覚に近い階層化が期待できる。

5. おわりに

本研究では、手順の対応関係に基づいてハウツー情報の検索結果を階層的にまとめ、ツリーのリストとして提示する手法を提案した。具体的には、手順を動作とその対象のペアとして表現し、日本語 WordNet および Wikipedia を用いて、手順間の対応関係を発見する手法を提案した。また、この対応関係を用いてハウツー情報間の関連度を定義し、これに基づいてハウツー情報を階層化する手法を提案した。関連度は共通度、被覆度、ばらつき度を用いて定義される。

今後は手順間の対応関係、対象の類似度、ハウツー情報間の関連度のそれぞれについて評価実験を行うと共に、実例の規模や数を増やして考察を行う。

謝 辞

本研究の一部は、平成 24 年度科研費若手研究 (B)「情報の詳

表 1 使用したページの URL

ID	URL
h_1	http://cookpad.com/recipe/5629
h_2	http://cookpad.com/recipe/956932
h_3	http://recipe.inoinox.com/gohan/nikujoyaga.html
h_4	http://www.kikkoman.co.jp/homecook/search/recipe.php?numb=00000596
h_5	http://erecipe.woman.excite.co.jp/features/okazu/o16.html
h_6	http://erecipe.woman.excite.co.jp/detail/246ff0f634c81a014920b3bc14f1ef9c.html

細関係に基づく Web ページの組織化」(課題番号: 24700097)
 によるものです。ここに記して謝意を表すものとします。

文 献

- [1] “情報検索に対する信憑性に関する調査”, <http://www.dl.kuis.kyoto-u.ac.jp/i-explosion/report/index.html>
- [2] 野中 諒志, 湯本 高行, 新居 学, 高橋 豊, “概要・詳細の見やすさに基づく手法情報のランキングと閲覧支援”, WebDB Forum 2010, 2A-1 (2010).
- [3] 小久保 卓, 小山 聡, 山田 晃弘, 北村 泰彦, 石田 亨, “検索隠し味を用いた専門検索エンジンの構築”, 情報処理学会論文誌, Vol.43, No.6, pp.1804-1813 (2002).
- [4] 鶴田 雅信, 増山 繁, “レイアウト情報を用いた Web ページの主要な DOM ノードの抽出法”, 人工知能学会論文誌, Vol.25, No.6, pp.742-756 (2010).
- [5] Deng Cai, Shipeng Yu, Ji-Rong Wen, Wei-Ying Ma, “VIPS: A Vision based Page Segmentation Algorithm”, Microsoft Technical Report (2003).
- [6] 小澤 俊介, 内元 清貴, 松原 茂樹, “モノの使われ方の情報がノウハウ獲得に与える影響”, 電子情報通信学会論文誌 D, Vol.J95-D, No.3, pp.506-517 (2012).
- [7] Hisashi Miyamori, “Assisting the Validity Assessment of Items based on Composition Similarity”, Proceedings of the ACM multimedia 2009 workshop on Multimedia for cooking and eating activities, pp.15-21(2009).
- [8] 浜田 玲子, 井手 一郎, 坂井 修一, 田中 英彦, “料理テキスト教材における調理手順の構造化”, 電子情報通信学会論文誌, J85-D-II(1), pp.79-89 (2002).
- [9] 湯本 高行, “Web 上のハウツー情報からの構成要素の抽出”, 情報処理学会第 153 回データベースシステム研究発表会, A-1-3 (2011).
- [10] Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh, “Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus”, Proceedings of the 6th International Conference on Language Resources and Evaluation, pp.654-660 (2008).
- [11] Makoto Nakatani, Adam Jatowt, and Katsumi Tanaka, “Easiest-First Search: Towards Comprehension-based Web Search”, Proceeding of the 18th ACM conference on Information and knowledge management, pp.2057-2060 (2009).
- [12] Douglass R. Cutting, David R. Karger, Jan O. Pedersen and John W. Tukey, “Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections”, Proc. of SIGIR '92, pp. 318-329 (1992).
- [13] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchi-moto, Takayuki Kuribayashi and Kyoko Kanzaki, “Enhancing the Japanese WordNet”, The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009 (2009).
- [14] 白川 真澄, 中山 浩太郎, 荒牧 英治, 原 隆浩, 西尾 章治郎, “Wikipedia と Web の情報を組み合わせたオントロジー構築の試み”, 電子情報通信学会論文誌 D, Vol.J94-D, No.3, pp.525-539(2011).