Vol.2011-DBS-153 No.3
2011/11/3

Web 上のハウツー情報の構成要素の抽出

湯　本　高　行[†1]

Web 情報検索の目的は多様化しているが，レシピやソフトウェアのインストール方法などのハウツー情報の需要は高い．そこで，本研究では，ハウツー情報の構成要素を “手順” と考え，“手順” を “動作” と “対象” のペアでモデル化する．さらに，Web ページから手順を抽出する手法を提案する．手順の抽出の際は，文末表現に注目して手順を説明する文のみを発見し，それらの文に対して，動詞と目的語を発見して，動作と対象とみなす．また，対象が省略されている文に対しては，前の文から対象を補う．提案手法の評価を行ったところ，平均で適合率が 0.49, 再現率が 0.57, F 値が 0.51 という結果であった．結果を分析したところ，対象の補完，慣用句による動作の表現などの問題があることがわかった．

Extracting Components of How-to Information on the Web

Takayuki Yumoto[†1]

The purposes of Web search become various. Many users want to obtain how-to information such as recipes and a way to install the software. We regard a “procedure” as a component of how-to information, and model it as a pair of an “action” and a “target”. We propose a method to extract the components from the Web pages. In our method, we find sentences explaining procedures by using the expression at the end of the sentences. Then, we find verbs and objects from them, and regard them as actions and targets respectively. When targets are omitted in the sentences, we complement them from the previous sentence. We evaluated our algorithm. Its precision was 0.49, and its recall was 0.57, and its F-measure was 0.51. From the analysis of the results, we found that there are problems in complement of target and in actions expressed by idioms.

†1 兵庫県立大学工学研究科
Graduate School of Engineering, University of Hyogo

1. はじめに

現在，検索エンジンを利用して情報を探すことは一般的になっているが，その一方でユーザの欲する情報は多様化している．中村らは情報検索に対する信頼性に関する調査[1)] で 1000 人に対してアンケート調査を行った．その中の「Q4. 検索した時に知ろうとしているもので多い物は何ですか」という質問に対して，各回答者が 1〜5 位として挙げたものの総数をまとめたものを**図 1** に示す．図 1 で挙げられている目的の中でハウツー情報は 3 位を占めており，重要な目的の一つである．

これに対して野中らはハウツー情報を検索し，見やすい順にランキングする手法を提案している[2)]．この手法では，既存の検索エンジンを用いて Web ページを取得し，その後，ハウツー情報に含まれる表層的な特徴に注目して，ハウツー情報が掲載されている部分を抽出，最後にハウツー情報を見やすさに基づき，ランキングを行う．特にハウツー情報の抽出の際には,「1.」,「(2)」,「c)」などの箇条書きに注目し，その各項目を手順の最小単位として扱い，抽出している．さらに，ランキングの際にはこの単位に基づいてスコアを計算している．

しかし，箇条書きの項目の粒度をどのようにするかは Web ページの作成者による．たとえば，料理のレシピでは，あるページでは複数の食材を切る部分を 1 つの手順として扱い，別のページでは食材ごとに切り方が違うので，複数の手順として扱う場合などが考えられる．一方，ハウツー情報の間の詳しさや内容の一致について分析が必要になった場合にはより細かい粒度の手順を扱う必要がある．また，ハウツー情報の中には箇条書きを用いずに説明しているものもあり，これらへの対応を考えた場合もページの作成者によって粒度が変わらないような単位で手順を扱うことが妥当であると考える．

そこで，本研究では，ハウツー情報の手順を表す最小単位を動作と対象のペアで表現し，これをハウツー情報の構成要素とするモデルを提案する．これにより，ハウツー情報を動作と対象のペアの集合として表現する．また，ハウツー情報の構成要素を抽出する手法を提案する．この手法では，文末表現によって手順を説明している文のみを発見し，そこから動詞と目的語を発見して，動作と対象とみなす．また，対象が省略されている文に対しては，前の文から対象を補う．

Vol.2011-DBS-153 No.3
2011/11/3

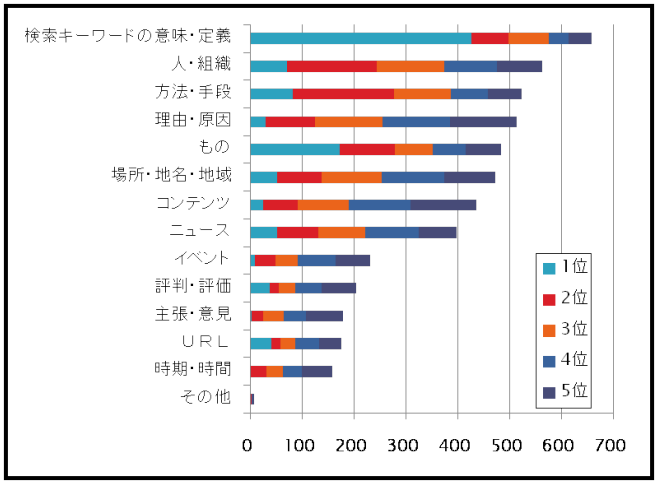

図 1 検索用途に関するアンケート結果

2. 本研究の位置付け

2.1 ハウツー情報検索

本節では，野中らの提案しているハウツー情報検索について概説し，本研究の位置付けについて述べる．野中らのハウツー情報検索は以下のように実行される[2]．

(1) 既存の検索エンジンを用いて Web ページを取得する
(2) ハウツー情報が掲載されている部分を抽出する
(3) ハウツー情報を見やすさに基づき，ランキングする

概略を**図 2** に示す．

まず，(1) では，既存の検索エンジンを用いて，候補を取得する．すなわち，野中らのハウツー情報検索のシステムはメタサーチエンジンとして機能する．検索のクエリは任意のものを許すとしており，この段階でハウツー情報を集中的に集める工夫などは行っていない．そのため，以下のような問題がある．

- ハウツー情報を効率的に収集できない
- 検索結果内に目的の異なるハウツー情報が混在する場合がある

前者の問題は，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である．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．一方，後者の問題はクエリの多義性の問題である．たとえば，「Linux インストール」というクエリで検索を行った場合に，Linux 自体のインストールについてのページと，Linux にあるソフトウェアをインストールする方法のページが混在する可能性がある．

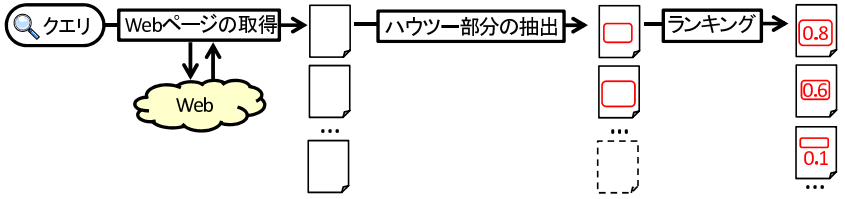

図 2 ハウツー情報検索の概略

続いて，(2) では，ハウツー情報を説明する箇所に出現する，表層的な特徴に基づいてハウツー情報の掲載部分 (ハウツー部分) を特定する．野中らの手法で注目している特徴は，箇条書き (「1.」，「(2)」，「c)」) や手順を表す語句 (「まず」,「次に」など) の存在，過去形の文が少ないということである．なお，ハウツー部分が存在しないページは検索対象から除外する．特に箇条書きの各項目を手順という単位であるとみなし，次のステップ (3) で用いている．このため，汎用的な HTML 文書からの本文抽出[4)–6)] とは異なり，より細かい粒度でのページの分割が必要になる．

最後に，(3) では，各ハウツー情報およびそれに含まれる手順を分析し，ランキングを行う．野中らは各ハウツー情報に出現する画像や語句に注目し，概要を把握する目的と詳細を把握する目的の 2 つの目的について見やすさのスコアを定義している．これは，語句の難易度[9] や専門度[10] などとは違い，ハウツー情報を理解するのに用いることが容易かどうかを表す尺度であるとしている．

2.2 ハウツー情報検索における本研究の位置付け

野中らの手法と提案手法との関係を説明する．提案手法は野中らの手法の (2), (3) で用いている手順のモデルより細かい単位で手順をモデル化するものである．これにより，(3) においてより詳細な分析が可能になると考えている．

また，(3) では，ランキングの対象となるハウツー情報集合の全体での頻度が高いものを

重視し，それを説明語としてランキングの計算に用いている．野中らはユーザに結果を提示する際にも，その語を強調することによってユーザのハウツー情報に対する理解を助けるとしている．しかし，(1) ではクエリの多義性の問題が存在するため複数の目的のハウツー情報が混在した状態で，説明語を抽出したり，ランキングを行ったりすることは，ランキングの精度やユーザにとってのわかりやすさを損なうと著者は考える．そこで，(2) の後にハウツー情報のクラスタリングを行い，クラスタごとに，(3) を実行することにより，説明語の抽出やランキングの精度，ユーザにとってのわかりやすさを向上させることが可能であると考えている．

提案手法の応用としては，共通する構成要素を抽出することで簡易的な要約が可能になると考えている．これは野中らの定義する見やすさの概要把握に役立てることもできる．逆に，共通しない構成要素はそのハウツー情報にとって特徴的であると考えることもできる．これは，野中らの定義する見やすさの詳細把握に関係すると考えている．

さらに，本研究は，構成要素を抽出するため，ハウツー情報間の手順の違いをより詳細に理解することができる．たとえば，ユーザが最低限の手順が書かれているページを閲覧中にもう少し詳細に書かれたページを閲覧したいと考えたとする．このような場合に，ページ内での構成要素の位置および構成要素間の対応関係がわかるため，より詳細に書かれたページの対応する部分へのナビゲートが可能になる．このように，ユーザがハウツー情報を探すための検索用途だけではなく，ハウツー情報を理解するための閲覧用途において有用であると考えている．

2.3 関 連 研 究

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する手法を提案している[7]．材料によるアイテムのモデル化はクラスタリングにも適していると考えられるが，本研究で扱う構成要素の抽出の目的は，クラスタリングだけではなく，手順の要約も含む．そのため，材料だけではなく，動作も用いたモデル化が必要であると考えている．

また，浜田らはレシピから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や事前知識を用いることで高い精度でデータフローグラフを生成することに成功している．このデータフローグラフは素材 (本研究で「対象」と呼ぶもの) と動作をノードとしており，本研究で扱う構成要素をより正確にモデル化したものと考えることもできる．しかし，この手法はドメインに特化にした辞書の作成や事前知識が必要となるため，ハウツー情報全般に適用することは難しい．本研究は，さまざまなドメインに対応することを重視するため，モデルを単純化することで，ドメインに特化した辞書や知識を使わない方法を採用している．

3. 提 案 手 法

3.1 構成要素のモデル化

野中らはハウツー情報は手順のリストとして構成されるとしている．一方，浜田らはハウツー情報の一種であるレシピをデータフローグラフで表現する手法を提案している[8] が，その中で手順が必ずしも順序によらないとしている．そこで，本研究では，ハウツー情報を手順の集合で表現する．また，手順 s を動作を表す語 a と対象を表す語 t のペアとして表現する．

$$s = (a, t) \tag{1}$$

なお，以下では，i 番目の手順 s_i を $s_i = (a_i, t_i)$ と表記する．また，ハウツー情報の手順を $S = \{s_1, s_2, \cdots, s_n\}$ と表記する．

3.2 構成要素の抽出

構成要素は以下の順に抽出する．

(1) 文への分割
(2) 説明文かどうかの判定
(3) 動作と対象の抽出
(4) 省略された対象の補完
(5) 手順の分割

なお，各 Web ページのハウツー部分は与えられているものとする[⋆1]．

3.2.1 文への分割

句点 (「。」,「.」) およびそれに準じた記号 (「!」,「?」など) によって文を分割する．この段階では,「野菜を切り，炒める」のような複文を残したまま処理し，後の処理で文を分割するものとする．

3.2.2 説明文かどうかの判定

手順についての説明を行っている文を説明文と定義する．ハウツー情報に対する説明文ではない文としては，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，疑問文などがある．これらに対応する文末を**表 1** に定義する．文末がこれ

⋆1 ハウツー部分の抽出の具体的な方法は文献[2] を参照のこと

らに一致する文は以降の処理対象から除き，残った文のみを説明文とみなし，以降の処理の対象とする．

3.2.3 動作と対象の抽出

文ごとに形態素解析を行い，末尾の形態素から遡って，動作と対象を抽出する．この段階では，対象は集合 T_i として扱い，省略された対象の補完を行った後に分割する．

まず，動作は自立語となる動詞とする．ただし，該当する動詞が存在しない場合および動詞が「する」の場合は，文末に最も近い位置にある非自立語 (「こと」など) でない名詞を動作とする．例えば,「たまねぎをみじん切りにする」の動作は「みじん切り」である．

次に対象を発見する．対象は助詞の「は」または「を」の直前に出現する名詞とする．これは,「たまねぎ を みじん切りにする」は「たまねぎ は みじん切りにする」とも書けるためである．なお,「と」や「,」などのような並列の場合は複数の名詞を対象とし，それぞれを構成要素とする．たとえば,「肉と野菜を炒める」という文からは，(炒める, { 肉, 野菜 }) を抽出する．

解析中に読点が出現した場合，その直前の形態素の種類によって処理を変える．名詞の場合 (例：「肉，野菜を炒める」の下線の部分) は並列であるとし，その直後の名詞と同様に扱いをする．すなわち，直後の名詞が対象の場合は，対象とみなし，それ以外の場合は無視する．動詞の連用形 (例：「野菜を 切り，炒める」) もしくは接続助詞 (例：「野菜を切って，炒める」) の場合は，複文であるとみなし，それ以前の部分について,「説明文かどうかの判定」からやり直す．上記以外の場合 (例：「3 分たった ら，できあがり」) は複文だが，説明文ではないとみなし，その読点以前の部分については処理しない．

3.2.4 省略された対象の補完

すべての文について動作と対象を抽出した後，対象が抽出されなかった場合，$T_i = \emptyset$ と表記する．$T_i = \emptyset$ の場合，$T_i = T_{i-1}$ とすることで，省略された対象を補完する．

上記の手順によって，抽出される構成要素の例を示す．以下のような文を考える．

たまねぎをみじん切りにし，炒める．

表 1 説明文の判断基準となる文末

種類	文末
過去形	た
状態動詞	いる/います，ある/あります，だ/です
口語表現	しょう/よ/ね/ー
疑問形	?/？

この文では，後半部分から (炒める, ∅) が，前半部分から (みじん切り, { たまねぎ }) が抽出される．その後，(炒める, ∅) の対象部分が，(みじん切り, { たまねぎ }) によって補われ，(炒める, { たまねぎ }) となる．すなわち，構成要素としては，(みじん切り, { たまねぎ }), (炒める, { たまねぎ }) が得られる．

3.2.5 手順の分割

対象の部分が集合になっている手順を，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．

4. 実　　験

4.1 実験方法

構成要素の抽出方法の評価のため，実験を行った．実験には検索クエリを「肉じゃが」,「豆腐ハンバーグ」として Bing Search API[⋆1]で検索したページのうち，野中らの手法でハウツー情報であると正しく判定されたページの上位 5 件ずつを用いた．使用したページの URL を**表 2** に示す．

著者が各 Web ページのハウツー部分に対して単文ごとに動作と対象を手作業で抽出したものを正解セットとした．たとえば,「たまねぎをみじん切りにし，炒める」という文であれば,「たまねぎをみじん切りにし」と「炒める」のそれぞれについて動作と対象を抽出する.「たまねぎをみじん切りにし」では (みじん切り, たまねぎ) が正解とする．また,「炒め

表 2 実験に使用したページ

ID	URL
肉じゃが 1	http://cookpad.com/recipe/5629
肉じゃが 2	http://erecipe.woman.excite.co.jp/features/okazu/o16.html
肉じゃが 3	http://recipe.gourmet.yahoo.co.jp/T001001/
肉じゃが 4	http://allabout.co.jp/gm/gc/42805/
肉じゃが 5	http://shiori22.exblog.jp/6273614/
豆腐ハンバーグ 1	http://cookpad.com/recipe/423284
豆腐ハンバーグ 2	http://cookpad.com/recipe/306187
豆腐ハンバーグ 3	http://recipe.gourmet.yahoo.co.jp/T002084/
豆腐ハンバーグ 4	http://aiaicafe.exblog.jp/4997504
豆腐ハンバーグ 5	http://www.royalqueen.jp/recipe/recipe4305.htm

⋆1 http://www.bing.com/toolbox/bingdeveloper/

る」は省略されている目的語を補い，(炒める, たまねぎ) を正解とする．評価には，以下のように定義する適合率 (P)，再現率 (R), F 値 (F) を用いた．

$$P = \frac{|A \cap T|}{|T|} \tag{2}$$

$$R = \frac{|A \cap T|}{|A|} \tag{3}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{4}$$

ただし，A は正解セットに含まれる手順の集合，T は提案手法で抽出された手順の数である．また，形態素解析器は MeCab[⋆1]を用いた．

4.2 結果と考察

実験結果を**表 3** に示す．手順の数が少ない，すなわち，簡潔に書かれたページについては比較的良好な結果が得られているが，手順の数が多いページについては極端に結果が悪いものがあるため，この原因について考察する．

主な原因として文内で省略されている対象の取得に失敗していることがあげられる．対象が省略されている場合は直前の対象で補うという方針をとっているが，直前ではなく，さらにその前の対象で補う必要があるものが数多くあった．特に,「加える」,「入れる」などの動作の以降の手順は「加える」や「入れる」の対象だけではなく，その直前の手順の対象も対象になるが，それに対応できておらず，対象を正確に抽出することができていなかった．

表 3 実験結果

ID	$\|A\|$	$\|T\|$	$\|A \cap T\|$	P	R	F
肉じゃが 1	12	18	11	0.61	0.92	0.73
肉じゃが 2	36	42	27	0.64	0.75	0.69
肉じゃが 3	29	44	17	0.39	0.59	0.47
肉じゃが 4	30	40	20	0.50	0.67	0.57
肉じゃが 5	40	50	20	0.40	0.50	0.44
豆腐ハンバーグ 1	31	33	8	0.24	0.26	0.25
豆腐ハンバーグ 2	21	28	18	0.64	0.86	0.73
豆腐ハンバーグ 3	41	14	4	0.29	0.10	0.15
豆腐ハンバーグ 4	21	15	9	0.60	0.43	0.50
豆腐ハンバーグ 5	19	21	12	0.57	0.63	0.60
平均				0.49	0.57	0.51

⋆1 http://mecab.sourceforge.net/

これに対しては,「加える」,「入れる」などについてはハウツー情報に共通して使用される表現なので別途ルールを設定することが妥当であると考えている．

また，対象そのものか対象の部分かについての判別を行っていなかったことも原因としてあげられる．たとえば,「たまねぎは皮をむき，芯をとる」という文において,「芯をとる」の対象が「たまねぎ」であることが抽出できなかった．これについては，レシピなどのように材料がまとめられている箇所があるものについては，その箇所を分析することで，あらかじめ，対象の候補を取得したり，他のハウツー情報との比較を行うことによって候補を修正するなどの対策が考えられる．

その他の原因としては，動作の抽出において慣用句に対応していなかったことがあげられる．たとえば,「冷ます」ことを説明するのに,「あら熱をとる」と表現していたため，動作として「とる」が抽出された．これに対しては，シソーラスを利用したり，他のハウツー情報と比較して同義表現を発見するなどの対策が考えられる．

また，口語調で書かれていたり，ひらがなで書かれていたために形態素解析に失敗していたものも数件あった．

さらに，複数の対象が連続して省略された場合，ある時点で対象の推定を失敗すると，それ以降の対象の推定も失敗するため，これが精度を低下させていた．これについては，上記の問題を解決することによって解消されると考えている．

5. おわりに

本研究では，ハウツー情報の構成要素を手順とし，これを動作と対象のペアとしてモデル化した．そして，その抽出方法を提案した．手順の抽出は，まず，文末表現を手がかりとして，過去形，状態の説明，口語表現，疑問文を排除することにより，手順を説明する文のみを発見する．次に，それらの文に対して，動詞と目的語を発見して，動作と対象とみなす．最後に，対象が省略されている文に対しては，前の文から対象を補い，対象が複数のものについては手順を分割する．

提案手法の評価を行ったところ，平均で適合率が 0.49, 再現率が 0.57, F 値が 0.51 という結果であった．この原因として，対象の補完の不備，対象そのものと部分の判別への未対応，慣用句による動作の表現への未対応，口語表現に対する形態素解析の問題などが考えられる．今後は，以下によって抽出性能の向上を目指す．

- 汎用的なシソーラスの利用およびハウツー情報に共通した辞書の構築
- 他のハウツー情報の比較による動作および対象の推定

謝辞 本研究の一部は，平成 23 年度科研費若手研究 (B)「把握容易性に基づく手法掲載ページの検索とランキング」(課題番号：22700108) によるものです．ここに記して謝意を表すものとします．

参 考 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) 吉田光男, 山本幹雄, “教師情報を必要としないニュースページ群からのコンテンツ自動抽出”, 日本データベース学会論文誌, 日本データベース学会, vol.8, no.1, pp.29–34 (2009).
6) Deng Cai，Shipeng Yu，Ji-Rong Wen，Wei-Ying Ma，“VIPS: A VIsion based Page Segmentation Algorithm”，Microsoft Technical Report (2003).
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) 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).
10) 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).