文書構造と言語表現の分析に基づく方法説明抽出

麻野間 直樹　　古瀬 蔵　　片岡 良治

日本電信電話株式会社　ＮＴＴサイバーソリューション研究所

{asanoma.naoki, furuse.osamu, kataoka.ryoji}@lab.ntt.co.jp

1　はじめに

何かの方法を調べようとする時、膨大な Web 上の情報のどこかに答えとなる方法の説明（以下、方法説明と呼ぶ）があるだろうという期待から、Web 検索エンジンを駆使してその答えを探すことはよく行われる。たとえば、グラタンの様々なレシピを探したい時、「グラタン」という主題の単語と「レシピ」という方法特有の語句を組み合わせて検索エンジンに入力することにより、いくつか所望のレシピを得られる可能性がある。

しかし既存の Web 検索エンジンは、方法説明に特化した検索ランキングや結果の提示方法を備えているわけではなく、実際には検索結果一覧の中の概要文やリンク先の文書を読んで、ユーザが役に立つ文書かを逐一判断することが必要となる。つまり、方法説明を検索する目的においては、既存の Web 検索とは別に、方法説明の抽出に特化した検索技術が必要であると言える。

近年では方法説明に特化した抽出技術や検索技術の研究が盛んになりつつある。たとえば、料理のレシピの手順を、抽出あるいは構造解析する技術[1][2]や、手順を示している箇条書きの自動判別を行う技術[3]が挙げられる。こういった先行研究は主に「手順」を対象としていた。

我々は、何らかの方法を問う質問に対して Web から情報を得ようとする時、既存の Web 検索よりも簡単かつ効率的に、その方法説明を抽出できる技術を目指している。本稿では、方法説明を抽出する仕組みを提案した後、方法説明のランキングに必要な「方法説明かどうか」を判別する手がかりを提案し、その検証実験について述べる。

提案する方法説明抽出の仕組みによって、既存の Web 検索では不十分だった、方法説明に特化した検索ランキングを可能にするとともに、従来技術が主に対象としていた「手順」だけでなく、選択肢が並ぶような方法説明の抽出も可能とする。

2　方法説明の分析

方法説明の文書の実態を把握することを目的として、方法説明を含む Web 文書にどのような特徴が見られるか調査した。

2.1　方法説明文書の収集

方法説明を含みそうな実際の文書を収集するため、何かの方法を問う質問 30 クエリ（例えば「梅干、漬け方」というクエリ）を、検索サイトgoo[1]　のウェブを検索」に入力し、各クエリに対する上位 90 件の検索結果である合計 2,700 ページのHTML形式の文書[2]を取得した（以下、取得した文書集合を取得HTML文書と呼ぶ）。各HTML文書の中で何らかの方法説明が書かれている部分があれば、その範囲を人手で特定した。

その結果、2,700 文書の中で方法説明を含んでいたのは 1,661 文書であった。つまり、既存の Web 検索を用いた場合、検索された文書の 62%にしか方法説明を含んでいないことになる。

2.2　方法説明の形態

前節の取得 HTML 文書を用いて、Web 上の方法説明は、ブラウザ上でどのような説明形態になっているかの予備調査を行った。10 クエリに対する各上位 10 件までの検索結果、合計 100 文書を調査対象とした。

取得HTML 文書には多くの箇条書きが含まれていたので、まず方法説明の形態を箇条書きに注目して分類する。箇条書きの構成形態は次の３種類に整理できる。

- 手順の箇条書き：複数の動作の１つ１つを順に箇条項目として記述している方法説明。箇条項目間に順序関係がある。（図 1 (a) の太枠）
- 選択肢の箇条書き：問題解決のバリエーションが複数羅列されている方法説明。箇条項目間の順序関係はない。（図 1 (b) の太枠）
- 箇条書きの一項目：説明部分が箇条書きの一項目のみに記述されている方法説明。（図 1 (c) の太枠）

他に箇条書きではない形態が２種類ある。

- 見出し文始まり：説明部分が見出し文から始まっている方法説明。（図 1 (d) の太枠）

---

[1] http://www.goo.ne.jp/

[2] 文献[4]で定めた 30 クエリ、および取得したHTML文書と同一である。

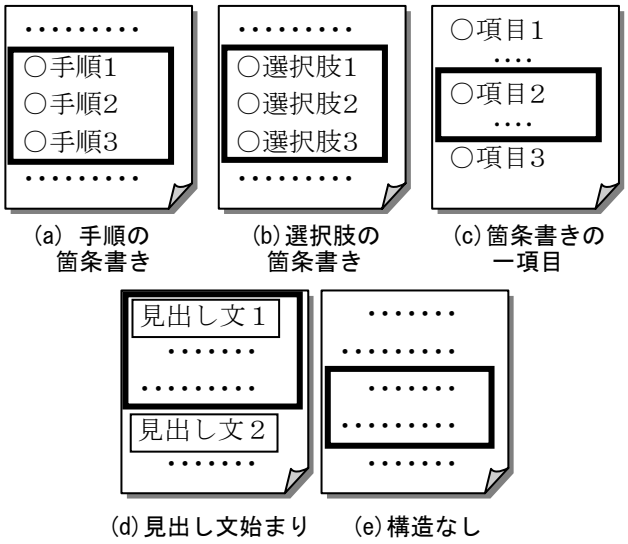

図１：方法説明の形態

表１：方法説明の形態とその分布

箇条書き			見出し文始まり	構造なし
手順	選択肢	一項目		
79	33	18	17	19
51.6%	21.6%	13.7%	11.1%	12.4%

- 構造なし：構造をまったく持たず、文中に現れる方法説明。（図 1 (e)の太枠）

本稿では、「見出し文」を、箇条書きの一項目ではなく、本文と区別して強調されている文と定義する。

調査対象 100 文書から、方法説明のまとまりとして、152 箇所を人手により特定した。表 1 は、各形態の、方法説明と特定された部分の件数と割合を示したものである。一箇所の方法説明に 2 種類の方法説明の形態を含むものは 14 件あり、その 14 件全ては何らかの箇条書きの形態をとっていた。

その結果、方法説明のうちの 76%（116 件）は箇条書きを構成していることがわかった。「見出し文始まり」は、見出し文と本文部分という構造を持っていると見做せるので、方法説明の 89%（133 件）には何らかの文書構造を含んでいることになる。さらに箇条書きの内訳を見ると、「手順」以外の「選択肢の箇条書き」や「箇条書きの一項目」も比較的件数が多いこともわかった。

本稿では、先行研究の多くが対象としている「手順」の方法説明に限定せず、文書構造を持つ全ての方法説明を抽出の対象とする。

3 方法説明抽出の仕組み

本節では方法を問う質問文を入力とし、与えられた文書の中からその回答となる方法説明を抽出する一連の流れを説明する。

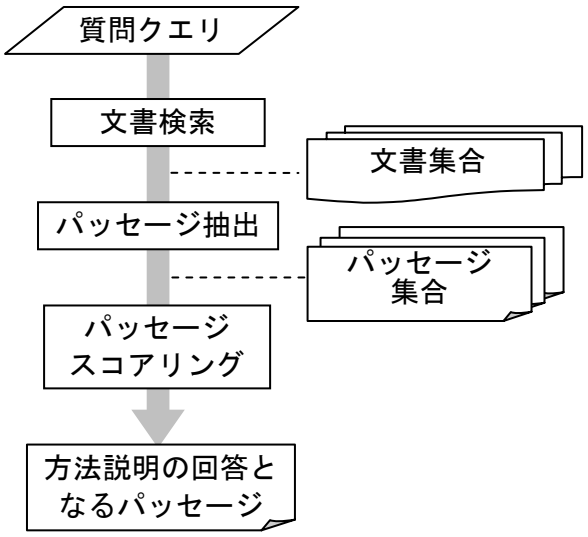

図２：方法説明抽出の流れ

3.1 抽出フロー

提案する方法説明の抽出手順を図２に示す。

文書検索では、通常の Web 検索の枠組みで、質問クエリから関連する文書集合を取得する。

パッセージ抽出では、取得した各文書から、説明回答の候補となるパッセージの集合を抽出する。本稿では、パッセージとは文書中から抜き出した一部分であり、見出し文から、次の同レベルの見出し文の直前までの範囲、もしくは箇条書きの一項目をパッセージの一単位とする。箇条書きの一項目の中に箇条書きがさらに含まれることもある。

パッセージスコアリングでは、質問クエリに関連していて、かつ方法説明がなされているパッセージが高い値となるよう、パッセージにスコアを付与する。さらにこのスコアを基にして、質問クエリに対する回答を抽出する。

次節では、方法説明の抽出において重要な役割を果たすパッセージスコアリングの実現法について説明する。

3.2 パッセージのスコアリング

質問クエリに対し、適切かつ方法説明らしいパッセージを選び出すことを目的として、次の２つの指標によりパッセージのスコアを測る。

質問関連性：質問クエリの内容と関連しているかどうかを示す指標

方法説明性：方法説明に関する内容かどうかを示す指標

この２つの指標の両方が高いパッセージが、求め

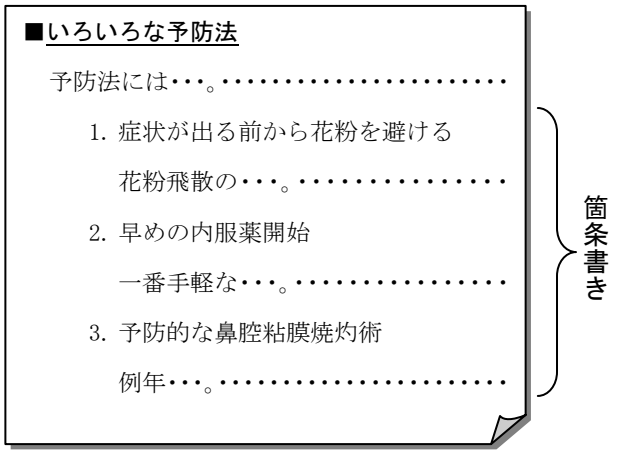

図３：方法説明パッセージの例

るべき方法説明の回答となる。

質問関連性は、従来の情報検索で用いられている質問と文書、もしくは質問とパッセージとの関連性に相当する。たとえば、「グラタンのレシピ」の質問文に対して、梅干の漬け方の記述は質問文とは直接関連する内容ではないので、質問関連性は低い。一方、マカロニグラタンのレシピの記述は、質問文と直接関連する内容なので質問関連性は高い。

質問関連性を示す数値指標は、BM25[5]やベクトルの類似度といった従来の検索モデルが適用可能である。

方法説明性は、文書そのものの方法説明らしさを示す。方法説明性が低い文書の代表的な例としては、単なるリンク集や商品の紹介だけが並ぶようなWebページが挙げられる。たとえば、「グラタンのレシピ」という質問クエリに対して、グラタンの歴史の記述は、質問文と関連する内容だが方法説明ではないので方法説明性は低い。一方、グラタンの調理手順の記述は、方法説明を記述しているので方法説明性は高い。

方法説明性は連続値を取る指標であるが、本検討では、方法説明性を方法説明であるかないか（方法説明性の有無）の２値で考え、方法説明性を測る問題を方法説明性の有無を判別する問題に置き換える。

4　方法説明性判別の手がかり

質問関連性は従来技術で測ることが可能だが、方法説明性については検討が進んでいない。これ以降は方法説明性に絞って検討を行う。本節では、方法説明性の判別の手がかりとなる見出し文と文末表現について説明する。

4.1　見出し文の分析

２節で示したように、方法説明のほとんどは、箇条書きや見出し文といった文書構造を持っている。

一般的に見出し文には、その後に続く内容をまとめた表現が現れる。特に、ある見出し文のすぐ後の文書に方法説明があったとすると、その方法説明を示唆するような表現が見出し文に現れる可能性がある。図３は花粉症の予防・対策の方法説明を記述したパッセージであり、予防法に関する選択肢の箇条書きを含んでいる。パッセージの先頭にある見出し文（図３の下線）に「予防法」という方法説明を示唆する語句が出現している。

このように箇条書きの前方にある見出し文に注目して、見出し文に現れる特徴を方法説明性判別の手がかりとする。

表２：設定した方法説明の文末表現

様相	文末表現の標準形
依頼	～下さい
勧誘	～ましょう
命令	～なさい
必要	～必要がある
良好	～ほうがよい
希望	～ほしい、～いただきたい
十分	～てみればよい、～十分だ

4.2　文末表現の分析

方法説明の中には、ある問題を解決するために相手に動作を促すような文末表現が用いられる、という仮説から、文末表現を方法説明と判断する手がかりとする。

このような依頼、勧誘、命令などの様相を含む文末表現をあらかじめ方法説明の文末表現として設定した。本検討で設定した９つの方法説明の文末表現を表２に示す。たとえば「ダイアログの OK ボタンを押してください」では「ください」という語が、「前もって医者にかかっておきましょう」では「ましょう」という語が、方法説明と判断する手がかりとなる。

5　方法説明性の判別実験

方法説明性の有無を判別する評価実験により、前節で提案した手がかりの有効性を検証した。

5.1　実験データ

２節で述べた 2,700 の HTML 文書から方法説明の判別実験データを取得した。具体的には、箇条書き部分に加えて、箇条書きの前方に現れる見出し文、および箇条書き前後の文脈を含めたパッセージを、実験データとして機械的に収集した。

実験データとして収集したパッセージは下記のすべての条件を満たすテキスト範囲とする。

・見出し文から始まり、次の見出し文の直前で終わるか、もしくは<li>で始まる１つの箇条項目の

範囲。見出し文は HTML の<h1>～<h6>で囲まれる文､あるいは箇条項目の場合はその第１文とする。
- パッセージの中には、<ul>もしくは<ol>の HTML タグで囲まれている部分を含む。
- パッセージ内の本文テキストが５文以上である。
- 方法説明である部分とそうでない部分が混在しない。

結果としては、2,700 文書から 10,673 のパッセージを取得し､そのうち､方法説明であるものが 1,820、方法説明でないものが 8,853 であった。図３で示すパッセージは、実験データ中の方法説明であるパッセージの一つとなっている。

5.2 実験方法

実験データのパッセージに対して、方法説明性の有無の判別実験を行った。用意した実験データのタグを除く本文部分を形態素解析した上で、素性を各単語（形態素）とし、パッセージ内に出現しているか否かの２値を素性値とする。判別は２値分類器であるSVM（Support Vector Machine）を用いた。SVM の実装にはTinySVM[3]を利用し、SVMのカーネル関数には２次の多項式関数を用いた。

用いる素性の集合を以下に定義する。

F1：本文中に出現する各単語の有無
F2:見出し文もしくは箇条項目の第１文で出現する各単語の有無
F3：箇条書き中の各文末表現の有無

F2、F3 が前節で提案した特徴を反映した素性であり､F1 は単純に方法説明に現れる単語の分布を反映したベースラインの素性とみなせる。これらの素性の組み合わせを考え、“F1”、 “F1+F2”、 “F1+F3”、“F1+F2+F3”の４条件で､SVM による学習と判別を行い、提案する特徴の有効性を検証する。判別実験については、５分割の交差検定を行った。

5.3 実験結果

表３は各条件における、方法説明であるパッセージの適合率 P、再現率 R、F 値を示している。F 値は以下の式で表される。

$$F = \frac{2PR}{P+R}$$

見出し文の特徴を特別扱いとした条件“F1+F2”は、単に単語の分布を反映した条件“F1”より、わずかではあるが適合率を向上させることができた。しかし再現率は若干低下しており、見出し文の有効性を明らかに示すものとは言いがたい。

一方、条件“F1+F3”と条件“F1”との比較、あるいは条件“F1+F2+F3”と条件“F1+F2”との比較を見ると、すべての値が向上していることがわかる。つまり箇条書き内の文末表現に重きを置くことで、より的確に方法説明らしいパッセージを判別できることがわかった。

表３：方法説明性の判別結果

条件	適合率 P	再現率 R	F 値
F1	85.3%	77.7%	81.4%
F1+F2	85.8%	76.8%	81.0%
F1+F3	86.8%	78.3%	82.3%
F1+F2+F3	87.4%	78.4%	82.7%

6 まとめ

方法を問う質問に対して、Web からその回答を抽出することを目的として、箇条書きの前方に来る見出し文と、箇条書き中に現れる方法説明特有の言い回しに着目することによって､文書の方法説明性(方法説明らしさ）を測る手法を提案した。

方法説明性の有無を判別する実験では、単に単語の分布の特徴を用いた場合よりも、適合率を 87%まで向上させることを確認でき、方法説明の抽出技術の実現に近づくことができた。

方法説明性判別の手がかりとして、文末表現のほかにも、意思性のある動詞などの言語表現の有効性の検討を行う予定である。

また本稿では、提案した方法説明の抽出フローの中で根幹となる方法説明性の有無の判別に絞って手法の評価を行ったが、全体の方法説明の抽出性能を測る総合的な評価も必要であるため、質問関連性とあわせたパッセージのスコアリングについても検討を行っていく予定である。

参考文献

[1] 田島，奥村．Web 上の料理レシピの抽出とその利用．第 11 回言語処理学会年次大会発表論文集, pp.65-68，2005.

[2] 浜田，井出，坂井，田中．料理テキスト教材における調理手順の構造化． 信学論（D-II），Vol.J85-D-II, No.1, pp.79-89, 2002.

[3] 武智, 徳永, 松本, 田中. WWW ページからの手順に関する箇条書きの抽出. 情処学論:データベース, Vol.44, No.SIG12, pp.51-63, 2003.

[4] 麻野間，古瀬，片岡．How-to 型質問応答の実現に向けた質問回答文書の特徴分析．信学技報，NLC2005-1, pp.55-60, 2005.

[5] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford. Okapi at TREC-3. Proceedings of the 3rd Text Retrieval Conference (TREC-3), 1994.

[3] http://www.chasen.org/~taku/software/TinySVM/