ソフトウェア機能を対象とした質問応答

竹形誠司　藤井敦

筑波大学大学院図書館情報メディア研究科

{takegata,fujii}@slis.tsukuba.ac.jp

1. はじめに

情報技術の発達に伴い、ソフトウェアの機能が多彩かつ高度になっている。そこで、マニュアルの記述が膨大かつ複雑になり、ユーザが知りたい情報を効率よく探す方法が必要である。しかし、ユーザとマニュアルの間に生じている「不一致」のために、目的の情報を探すことが困難な場合がある。

不一致の 1 つは、ユーザが思い浮かべる言葉とマニュアルで使われている用語の不一致である。たとえば、Microsoft Word ではページ上部の欄外を「ヘッダー」と呼ぶ。そこで、ユーザが「欄外」という言葉で探しても、ヘッダーに関する機能が見つからない可能性がある。

もう 1 つの不一致は、ユーザが思い浮かべる機能とソフトウェアが装備している機能の不一致である。Microsoft Word ではワードラップ機能の副作用により、文字間隔が不自然に広がってしまうことがある。そこで、ユーザが文字の間隔を設定する機能を探しても問題が解決しないことがある。

Wilensky ら[1]は、知識ベースと推論機構によって UNIX に関するユーザの質問に答えるシステムを構築した。清田ら[2]は、マイクロソフトの技術文書を用いてユーザの質問に答えるシステムを構築した。これらの研究に共通する問題点は、ユーザとマニュアルの間の不一致を解消するために、知識ベース、ルール、辞書などを人手で作成しなければならない点である。

本研究は、ソフトウェアに関するユーザの疑問が 1 つ以上の機能名や設定項目名で解決することが多い点に着目し、「ユーザが使用する平易な言葉」による質問に対して、「ソフトウェアの機能または項目の名前」を回答する質問応答手法を提案する。また、提案手法を計算機上のシステムとして実装し、実験によって精度を評価する。

2. 質問応答の手法

2.1 概要

本研究で提案する質問応答システムは、ソフトウェアの使い方に関する質問を入力し、ユーザの要求を満たすソフトウェア機能の順位付きリストを回答として出力する。本研究では、回答の単位となるソフトウェアの機能や設定項目を「機能項目」と総称する。本研究で提案する質問応答システムの概要を図 1 に示す。

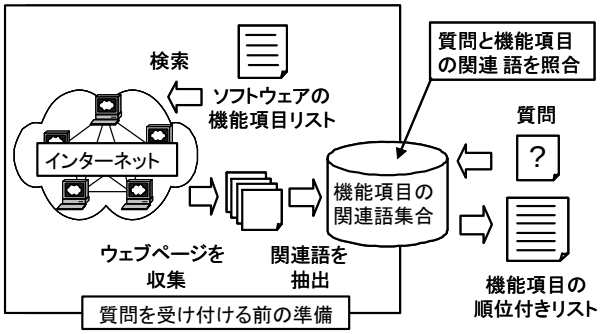

図 1: 質問応答システムの概要

図 1 について説明する。まず、Microsoft Word などの対象ソフトウェアに関する機能項目のリストを人手で作成する。次に、それぞれの機能項目に対してインターネット上のウェブページから関連語を抽出する。抽出した関連語と各機能項目を関連付けて索引付けを行う。関連語を中継することで、質問と機能項目を柔軟に照合し、「不一致」問題を解消する。

ユーザの質問を受けると、質問と機能項目の関連度を計算して、関連度が高い順に機能項目の順位付きリストを作成し、出力する。

2.2 機能項目リストの作成

本研究では、対象ソフトウェアのメニューに表示される項目を機能項目として扱う。さらに、ダイアログボックス内に表示されている項目も機能項目として扱う。これらの機能項目は階層的に配置されている。図 2 に Microsoft Word に関する機能項目の階層を一部示す。

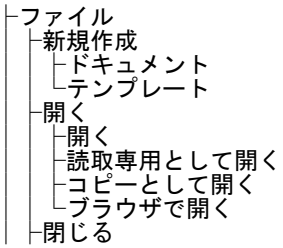

図 2: Microsoft Word における機能項目の階層（一部）

本研究では、図 2 の階層構造を図 3 のようなメニューから機能項目へ至る経路として表現し、各経路を 1 つの機能項目として扱う。1 つの機能項目名はソフトウェアの操作手順を表している。

1:ファイル
2:ファイル⇒新規作成
3:ファイル⇒新規作成⇒ドキュメント
4:ファイル⇒新規作成⇒テンプレート
5:ファイル⇒開く
6:ファイル⇒開く⇒開く
7:ファイル⇒開く⇒読取専用として開く
8:ファイル⇒開く⇒コピーとして開く
9:ファイル⇒開く⇒ブラウザで開く
10:ファイル⇒閉じる

図 3: Microsoft Word の機能項目リスト(一部)

図 3 において、行頭の数字は機能項目に 1 から順に割り当てた識別番号である。

2.3 ウェブページの収集

2.2 節で作成した機能項目のリストを使用して、それぞれの機能項目に関係するページ集合をウェブから収集する。たとえば、Microsoft Word の「表示⇒ヘッダーとフッター」という機能に関するページを収集する場合は、「Word 表示 ヘッダー フッター」を検索質問として使用する。

図 4 に Microsoft Word を対象としたウェブページの収集方法を示す。図 2 の「ファイル」、「新規作成」、「開く」など、階層構造の各ノードは機能項目である。階層構造のルートはソフトウェア名である「Word」とする。ウェブから検索されたページは、機能項目の関連語を抽出するために使用する。

本研究では NTCIR ウェブコレクションを対象とした検索エンジン[3]を実験に使用した。この検索エンジンでは、検索モデルとして Okapi BM25[4]を使用している。

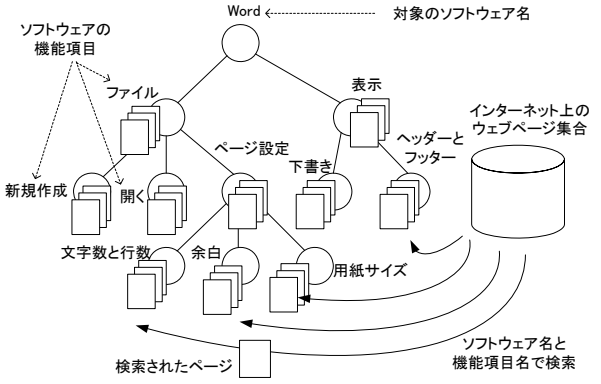

図 4: 機能項目に関するウェブページの収集

2.4 ウェブページからの重要箇所抽出

2.3 節の処理によって収集されたウェブページには、広告、ウェブサイトのメニュー、著作権表示のように、本文の内容とは関係ない部分がある。また、掲示板のページでは、検索キーワードに関連する記述は限られており、ほとんどが検索質問と無関係であることが多い。

ページ本文の内容と無関係な箇所に含まれる語を関連語として使用すると、質問と機能項目の照合に悪影響を及ぼす。そこで、機能項目に関連する重要箇所だけを抽出する。

収集したページから重要箇所を抽出するために、まず、ウェブページ中の HTML タグを削除する。次に、図 5 のようにウェブページを固定長（D）の断片に分割する。各断片を「パッセージ」と呼ぶ。

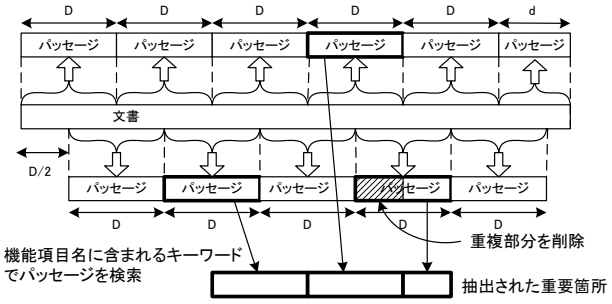

図 5: 重要箇所の抽出

作成したパッセージ集合から、重要な関連語を多く含むパッセージだけを検索する。パッセージ境界の前後に重要な関連語が分散していると、前後どちらのパッセージも検索されない可能性があるため、半分ずつ重なるようにパッセージを作成する。

ここで、2.3 節でウェブページの収集に使用した検索エンジンを使用してパッセージを検索する。たとえば、Microsoft Word の「ファイル⇒ページ設定⇒余白」という機能項目について収集したページの中から重要箇所を抽出する場合は、「ファイル ページ 設定 余白」という質問で検索をする。連続するパッセージが検索された場合は、重複部分があるため、後続するパッセージの前半部分を削除する。

2.5 索引付け

2.4 節で説明した手順によって機能項目ごとに検索されたパッセージに対して関連語の索引付けを行う。索引の単位は、単語および単語バイグラムである。単語バイグラムとは、隣り合う 2 つの単語を結合した索引語である。単語の特定には ChaSen を使用する。索引語の重み付けには Okapi BM25 を使用する。そこで、重みが高い索引語が結果として機能項目の関連語として働く。すなわち、関連語抽出を明示的に行う訳ではない。

2.6 質問応答処理

質問応答処理は、ユーザの質問を入力して、関連する機能項目の順位付きリストを出力する。質問応答処理の概要を図 6 に示す。

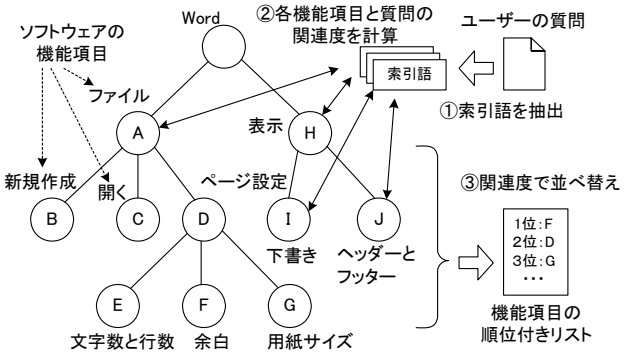

図 6: 質問応答処理

質問応答処理では、ユーザの質問から ChaSen を使って索引語（単語および単語バイグラム）を抽出する。ここで得られた索引語と、それぞれの機能項目に関連付けられた索引語を照合することによって、質問文と各機能項目の関連度を計算する。関連度の計算に使用する検索モデルは Okapi BM25 である。関連度が大きい順に機能項目をソートし、機能項目の順位付きリストを作成してユーザに提示する。すなわち、本研究では、質問応答という問題を「機能項目を検索する」という問題に置き換えた。

3 評価実験

3.1 実験に使用したデータ

実験対象のソフトウェアとして、Microsoft 社の Word 2000、Excel 2000、PowerPoint 2000 を使用した。これらのソフトウェアに関する書籍が多く出版されていることと、関連のウェブページや掲示板が多いためである。各ソフトウェアは最新バージョンではない。しかし、発売からの期間が長い分だけ、関連する書籍やウェブページは多く存在する。

実験対象のソフトウェアについて、Q&A 形式の書籍[5,6,7,8,9]とインターネットの掲示板[1]から質問と正解を収集した。質問と正解を収集した情報源を表 1 に示す。

表 1: 質問と正解を収集した情報源

ソフトウェア	情報源	質問数	正解数の平均
Word	書籍	50	2.1
	掲示板	72	1.19
Excel	書籍	158	6.23
	掲示板	134	1.16
PowerPoint	書籍	57	1
	掲示板	27	1

実験に使用した質問と正解の例を表 2 に示す。いずれも書籍から収集した質問である

[1] http://www.moug.net/

表 2: 収集した質問と正解の例

ソフトウェア	質問	正解
Word	ページをまたがった表に自動的に見出し行を表示させたい[5]	罫線⇒表のプロパティ⇒行⇒オプション⇒各ページにタイトルを表示する
Excel	入力後にセルの移動する方向を右に変更したい[7]	ツール⇒オプション⇒編集⇒設定⇒入力後にセルを移動する⇒方向⇒右
PowerPoint	すべてのスライドの書式を簡単に揃えたい[9]	表示⇒マスタ⇒スライド マスタ

3.1 実験結果

実験の結果を表 3 に示す。評価尺度は、入力した質問に対して、システムが出力した機能項目リスト中に正解が現われた順位の平均である。正解が複数存在する質問については、最初に正解が現われた順位を正解の順位とした。

機能項目名のみで関連度を計算した場合（関連語無）と本手法で収集した関連語を使用して関連度を計算した場合（関連語有）を比較すると、全てのソフトウェアで関連語有の場合に平均正解順位が向上した。

表 3: 平均正解順位の比較

ソフトウェア	機能項目数	平均正解順位	
		関連語無	関連語有
Word	4324	828.9	281.1
Excel	1749	397.0	171.6
PowerPoint	729	208.3	95.0

Word、Excel、PowerPoint の質問における正解順位の分布を表 4 に示す。Excel と PowerPoint では、関連語を使用することによって 1 位～10 位に正解が現われる質問がそれぞれ 88 件から 116 件、21 件から 29 件へと増えた。Word では、1 位～10 位に正解が現われる質問が 37 件から 36 件へと減った。しかし、11 位～100 位と 101 位～1000 位に正解が現われる質問が増え、1001 位以下に正解が現われる質問が減った。以上より、本研究で提案した関連語に基づく質問応答の有効性を確認することができた。

表 4: 正解順位の分布

ソフトウェア	Word		Excel		PowerPoint	
関連語の有無 / 順位	無	有	無	有	無	有
1～10	37	36	88	116	21	29
11～100	29	35	77	87	32	32
101～1000	30	43	75	78	31	23
1001～	26	8	52	11	－	－

表 5 に、関連語によって正解順位が変化した質問の数を示す。全ての場合において、関連語によって正解順位の向上した質問の数が、低下した質問の数を上回った。

表 5: 関連語によって正解順位が変化した質問の数

ソフトウェア	向上	変化なし	低下
Word	60	6	56
Excel	173	26	93
PowerPoint	54	4	21

関連語によって、正解順位が 612 位から 3 位に向上した Word の質問と正解を図 7 に示す。順位が向上した理由は、「赤」、「緑」、「波線」といった言葉で Word の文書校正機能が説明されているページがウェブから収集したページの中に含まれていたためである。

質問	文字の下に赤や緑の波線が表示されてしまうのですが
正解 1	ツール⇒オプション⇒文章校正⇒スペルチェック⇒自動スペルチェック
正解 2	ツール⇒オプション⇒文章校正⇒文章校正⇒自動文章校正

図 7: 関連語によって正解順位が向上した Word の質問例

関連語によって、正解の順位が 35 位から 1480 位に低下した Word の質問と正解を図 8 に示す。正解の順位が低下した理由は、質問に含まれる索引語が、他の機能項目で索引付けされた関連語として大きく重み付けされていたことである。

質問	文字を入力すると黒以外の色で表示されてしまいます。作成した文書を開いて文字を入力すると、設定したフォントの色以外で文字が入力され、下線が表示されてしまうのですが
正解	ツール⇒変更履歴の作成⇒変更箇所の表示⇒編集中に変更箇所を記録する

図 8: 関連語によって正解順位が低下した Word の質問例

関連語によって正解順位が 1749 位から 6 位に向上した Excel の質問と正解を図 9 に示す。以下、順位が変動した理由は Word の場合と同じである。

質問	ダブルクリックしてファイルが開けなくなってしまった
正解	ツール⇒オプション⇒全般⇒設定⇒他のアプリケーションを無視する

図 9: 関連語によって正解順位が向上した Excel の質問例

関連語によって正解の順位が 23 位から 157 位に低下した Excel の質問と正解を図 10 に示す。

質問	数値を入れたいが日付の表示になってしまった
正解	書式⇒セル⇒表示形式⇒分類⇒標準

図 10: 関連語によって正解順位が低下した Excel の質問例

関連語によって正解順位が 729 位から 2 位に向上した PowerPoint の質問と正解を図 11 に示す。

質問	インターネットでスライドショーを公開したい
正解	ファイル⇒Web ページとして保存

図 11: 関連語によって正解順位が向上した PowerPoint の質問例

関連語によって正解の順位が 138 位から 615 位に低下した PowerPoint の質問と正解を図 12 に示す。

質問	ヘッダーとフッターのフォントサイズ変更について。ヘッダーとフッターのフォントの種類及びサイズの変更について、変更項目が見当たらないのでやり方を教えて下さい。
正解	表示⇒マスタ⇒スライド マスタ

図 12: 関連語によって正解順位が低下した PowerPoint の質問例

おわりに

本研究は、ユーザの質問に対して、関連語による柔軟な照合によって、ソフトウェアの機能を回答する手法を提案した。機能項目名は操作手順を表しているので、ユーザはマニュアルや技術文書などの解説を読む必要がない。しかし、機能項目名だけでは情報が不足する場合は、関連情報へのリンクを提示するなどの対策を取る必要がある。

評価実験によって、関連語の効果を確認することができた。しかし、本手法では「複数の機能項目を組み合わせて対処する必要がある質問」は対象外とした。この問題に対処するには、機能項目の組み合わせと順序を自動的に決定する必要がある。

参考文献

[1] Robert Wilensky, David Chin, Marc Luria, James Martin, James Mayfield, and Dekai Wu. The Berkeley UNIX Consultant project. Computational Linguistics, Vol.14, No.4, pp.35-84, 1988.

[2] 清田 陽司, 黒橋 禎夫, 木戸 冬子. 大規模テキスト知識ベースに基づく自動質問応答－ダイアログナビー. 自然言語処理, Vol. 10, No. 4, pp.145-175, 2003.

[3] Atsushi Fujii, Katunobu Itou, Tomoyosi Akiba, and Tetsuya Ishikawa. Exploiting anchor text for the navigational Web retrieval at NTCIR-5. In Proc. of the Fifth NTCIR Workshop Meeting, pp.455-462, 2005

[4] S. Robertson, S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. Proc. of 17th ACM SIGIR, pp. 232–241 , 1994.

[5] マイクロソフトが答えるWord Q&A集. マイクロソフト編. 日経BP ソフトプレス, 2005.

[6] ワードの裏技・便利技. 荻原 洋子, 貝原 典子, 加藤 多佳子. 新星出版社, 2004.

[7] マイクロソフトが答えるExcel Q&A集. マイクロソフト編. 日経BP ソフトプレス, 2005.

[8] エクセルで困ったときの基本技・便利技. 小濱 良恵. 技術評論社, 2006.

[9] PowerPoint で困ったときの基本技・便利技. AYURA. 技術評論社, 2006.