QAサイトにおける専門用語を用いた最適な回答者提示

堀江 将隆　　山本 和英
長岡技術科学大学 電気系
{horie, yamamoto}@jnlp.org

1 はじめに

疑問に思っていることや悩みを質問し，他のユーザが回答することで知識を共有できるWeb上の質問応答サイト(以下，QAサイト)が存在する．QAサイトはWeb検索サイトでのキーワード検索と違い，文章を用いた柔軟な検索が可能であり，需要が高まっている．

QAサイトでは疑問を持っているユーザが質問を投稿し，その質問に対して回答可能な他のユーザが回答を投稿する．質問者の満足する回答がすぐに投稿された場合は問題ないが，一定期間が経過しても他のユーザから質問者が満足する回答が得られない場合がある．その為，得られた回答の中でどれが疑問を解決できる回答として正しいのか判断できない可能性があり，現在のQAサイトでは質問者の疑問や悩みを十分に解決できていないと考える．

そこで我々は質問者が最も満足できるような回答を投稿できる回答者を全ユーザから選出し提示することを目的とする．

本研究ではYahoo!知恵袋(1)の質問回答を対象に，ある質問に対してベストアンサー(以下，BA)に選ばれたユーザを最適な回答者と仮定し，そのユーザを選出する手法を提案する．

提案手法では，以下の2つの手順から最も適している回答者を提示した．まず，質問文から知識として専門用語を抽出する．次にその専門用語をもとに各回答者の質問回答履歴を参照し，最も適している回答者を選出する．

2 関連研究

関連研究として片山ら[1]，甲谷ら[2]のQAサイトにおける質問推薦が挙げられる．片山らは投稿された質問に対し回答が1件も投稿されないことを問題点として，回答者に対し普段回答している質問の内容に近い質問を推薦する手法を提案している．それに向けた履歴データの分析として，文章を表す索引語をもとに回答者のベクトル空間モデルと質問文のベクトル空間モデルを作成し，回答者と，実際に回答している質問とそれ以外の質問とのベクトル空間モデルでどれだけ類似しているかを実験している．

甲谷らは投稿された回答のみでは質問者が満足していないことを問題点として，良質な回答を増やす為にユーザ全員に質問を推薦する手法を提案している．複数の情報を用いてユーザの情報を推定した後，ロジスティック回帰モデルにて統合し，それを用いて質問推薦を行っている．

上記の研究は回答者に適した質問を推薦することが目的であり，質問者に対し最適な回答者を提示することを目的とする本研究とは異なる．

3 提案手法

3.1 回答に必要な情報

QAサイトにおける最適な回答者とは，以下に示す２つの条件を満たすユーザであると考える．

- 質問に対して不備無く回答できる
- 回答が正確であることを裏付けるような補足情報を付与することができる

質問者の観点から見れば，質問に関する知識をより多く持っているユーザに回答してもらえることが望ましい．例えば，パソコンに関する質問の場合，パソコンの知識を豊富に持っているユーザの回答は信頼性が高いと推測できる．よって，最適な回答者を提示する為には，対象の質問に関する知識を十分に持つユーザを見つけ出す必要がある．

本研究では，知識を持つユーザを探す手がかりとして，ユーザの回答履歴を用いる．回答履歴とは，ユーザが過去に回答した質問文とその回答文に関する情報のことである．この回答履歴を用いることで，ユーザ

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

がこれまでに回答してきた質問の傾向を知ることができ，そのユーザが持っている知識を推測できると考えた．実際に回答履歴を知識として利用する方法としては，文章を表現する語を用いることが挙げられる．文章を表現する語とは，文章内の内容語（動詞，名詞，形容詞）と考える．しかし，内容語全てを用いた場合，必要としている知識と関係の無い語を多く含むことになる．

ここで考える知識というのは質問がどういう分野に関係しているかを表現する語であり，それに相当するのは内容語の中でも専門用語であると考えた．専門用語は語単体でおおよその分野を表現することができる．よって，本研究では文章中に存在する専門用語をユーザの知識として扱い，最適な回答者を探す時に用いることにする．

3.2 専門用語を用いた回答者選出手法

専門用語は一般的にある特定の分野でのみ使用される用語であるが，本研究では分野を表すような用語も専門用語として扱う．

専門用語を用いて最適な回答者を選出できるかどうかを調査する為人手で専門用語を抽出した．

以下に質問文から抽出された専門用語の例を示す．

・対象の質問文

> PC をリカバリしたいのですが、プロダクトキーが書かれた冊子が見当たりません。プロダクトキーの入力なしでリカバリできないものでしょうか？WinXP メーカーは hp です。ご教示、よろしくお願いします。

・抽出された専門用語

> PC, リカバリ, プロダクトキー, WinXp, hp

これらの抽出された専門用語は全て質問の分野を表現する為に必要な語であると考えた．複合名詞は分割した場合，それぞれ意味を持つ場合がある為，分割した語に関しても専門用語であるかを判断する．

以下に最適な回答者の選出手順を示す．

step1 最適な回答者を提示したい質問文 q から専門用語を全て抽出し知識集合 K を作成する

step2 ユーザ u_n 内の回答履歴 UI_n（q は含まない）に出現する，知識集合 K の異なり数をカウントし，それをユーザ u_n のスコアとする

UI_n はユーザ u_n が過去に回答した質問文集合 Q とその回答文集合 A で構成される．

$UI_n = \{Q, A\}$

step3 回答者集合 U の全てのユーザに対して step2 の処理を行う

step4 回答者集合 U からスコアが最も高いユーザを選出する

4 評価実験及び考察

4.1 実験方法

提案手法によって最適な回答者がどの程度選出できるか評価実験を行った．実験に使用したデータは Yahoo!知恵袋の「インターネット、ＰＣと家電」カテゴリ (以下，PC カテゴリ) において実際に投稿された質問とそれに対して回答したユーザの回答履歴 (最大で 3245 件) である．このカテゴリを選択した理由は，Yahoo!知恵袋において最も質問数が多い点と，専門用語を抽出する際に他のカテゴリよりも多く出現し直感的にも分かりやすいと考えた為である．本研究では問題の難易度を均一化する為，回答ユーザが 3 人の質問のみを対象にした．PC カテゴリの回答ユーザが 2 人以上の質問において回答ユーザが 2 人の質問数が最も多かったが，回答ユーザが 2 人では 2 択問題になり，最終目標である全ユーザから選出する環境とかけ離れると考えた．よって，次に質問数が多かった回答ユーザが 3 人の質問を選択した．対象の質問回答データセットは 30 件とした．

本研究では，投稿した回答が BA に選ばれたユーザを最適な回答者と仮定し，選出したユーザが BA 回答ユーザである場合を正解とする．これは，BA 回答ユーザは質問に対する回答が最も満足するという理由で選ばれたユーザであるからである．

比較の対象として，文中の内容語全てを知識として用いた選出実験も行った．専門用語は主観で専門用語だと思うものを人手で抽出した．内容語は，質問文中から形態素解析ツール MeCab(2) を用いて抽出した動詞，名詞，形容詞とする．ただし，非自立語，代名詞は内容語として扱わない．

4.2 実験結果

各選出実験の結果を表 1 に，結果から得られた知識の差異を表 2 に示す．

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

表 1: 選出手法の正解率

	内容語	専門用語
正解率	43.3 % (13/30)	43.3 % (13/30)

表 2: 内容語と専門用語の差異

	内容語	専門用語
選出ユーザが複数の質問	4 件	7 件
知識の平均語数	35.6 語	10.1 語

表 1 から両手法とも正解率が低いことが分かり，専門用語を用いた手法でも正解率が向上しなかったことから有効性を確認できなかった．内容語はツールにより自動で抽出できるが，専門用語の抽出は容易でない．どちらも同じ精度ならば抽出コストを考えた場合，内容語を用いたほうが有効である．しかし，両手法の結果では一部相違があり，それぞれの手法でのみ正解となる質問が確認できた．内容語を用いた手法では，質問の内容を表さない「当初」や「イマイチ」のような語の存在がスコア上昇の要因となり，正解となっているものが 3 件あった．これでは，必要な知識に基づいて選出したことにはならず，単純に過去の回答件数が多いユーザが有利になってしまう．その為，専門用語を用いた手法のほうが正しく知識を獲得出来ていると言える．

不正解となった質問の回答者に着目すると，BA 回答ユーザの回答履歴が少ない場合が多かった．今回使用したユーザの回答履歴の回答件数の平均は 1,926 件であり，その 1 割にも満たない回答件数のユーザ，つまり Yahoo!知恵袋での利用経験が浅いユーザが BA に選ばれた質問は 4 件あった．これは，QA サイトで過去に残してきた知識の量が少ない為，本手法は適用できなかったと考える．本研究で選出したユーザの回答を確認すると，BA ではないが質問に対して十分に満足できる回答であると思えるものが 5 件あった．以下に例を示す．

・質問

EXCEL について教えてください。例えば SHEET １から SHEET ３を作っている EXCEL のファイルがあります。それぞれのファイルには計算したデータがあります。sheet １で出した数字と、sheet2 で出した数字の合計数を、sheet3 に自動的に関数を利用して反映させる方法はありますか？同じ sheet 内で、SUM を利用する方法は分かるんですが、sheet 間をこえてもできるのでしょうか。教えてください。

・BA に選ばれた回答

出来ます。sheet3 のセルに = を入力して、該当する sheet のセルにをクリックして、計算式を入力すればよいです。ん～。文字だけの説明は難しい…。

・選出したユーザの回答

各シート同じセル番地（範囲）なら【URL】 各シート表の構造が違う場合【URL】 またこの場合 2 シートなので例 Sheet3 のセルで＝と入力し下のシート見出し Sheet1 をクリック→計算したいセル A1 クリック→+入力→シート見出し Sheet2 をクリック→計算したいセル C1 クリック→Ｅｎｔｅｒ =Sheet1!A1+Sheet2!C1 こんな式で Sheet1 の A1 と Sheet2 の C1 の足し算 またこんな風に …

=SUM(Sheet1!A1:A10)+SUM(Sheet2!D10:D15)

この例から，選出したユーザの回答のほうが丁寧で詳しく書かれているように感じる．しかし，BA の定義は質問者が最も満足した回答であるので，ここでの最適な回答者とは，BA 回答ユーザである．この場合，最適な回答者の提示はできていないが適した回答者の提示はできていると言える．

質問文が少ない語数で構成されている場合，抽出できる知識の数が少なくなり，手法を適用するのが困難である．この為，複数のユーザの回答履歴において知識の異なり数が等しくなり，最も良いスコアを持つユーザが複数になる場合がある．つまり最適なユーザ一人を選出できていないと言える．表 2 より最も良いスコアを持つユーザが複数いた為，正解となった質問は内容語を用いた手法で 4 件，提案手法で 7 件確認できた．差が生じた理由は表 2 の知識の平均語数より，専門用語のほうが抽出できる語数が少ない為である．

そこで，この問題を解決する為に専門用語の拡張実験を行った．

4.3 専門用語の拡張実験

前節で挙げた問題を改善する為に専門用語を拡張して実験を行った．専門用語を拡張する方法として，Yahoo!検索の関連検索ワード Web API(3) を用いた．関連検索ワードとは，検索で使用されたキーワードをもとに，指定されたキーワードとよく組み合わせて検索されるキーワード等のことである．専門用語の関連検索ワードは，同じ分野を表現する語であることが多いと考えた．実際に取得した関連検索ワードの例を以下に示す．

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

・「メモリ」の関連検索ワード
メモリ増設, 仮想メモリ, バッファロー, ＵＳＢメモリ, 物理メモリ, 増設メモリ, ガイアメモリ, エルピーダメモリ, フラッシュメモリ, パソコン
・「ＰＣ」の関連検索ワード
ＤＥＰＯＴ, ＰＣゲーム, ＰＣボンバー, 自作ＰＣ, ＰＣ工房, １００円ＰＣ, Ｗａｔｃｈ, タブレットｐｃ, ｐｃマックス, モバイルＰＣ

メモリの関連検索ワードでは，多くの語が関連する専門用語と考えてもよいが，PC の関連検索ワードでは専門用語である語は少数であった．これは検索クエリを使った手法であることから，一般的に多く使用される語ほど，様々な語と共起することが多い為と考える．逆に専門用語の中でも特定の分野でのみ使用される語ならば，関連する語は専門用語である可能性が高いと考える．

専門用語を拡張した為，抽出できる知識が少なく回答者を提示できなかった質問に対して，最適なユーザーを提示しやすくなると考える．この実験ではそういった問題を改善することが目的だが，それ以外の質問でも表 2 より内容語と比べて専門用語は少ない為，知識の数が増えることにより選出するユーザが最適な回答者である可能性が高くなることが期待できる．

提案手法の評価実験と同様の方法で，知識集合に専門用語の各関連検索ワード最大 50 語を追加して実験した．30 件の質問で実験した結果と，提案手法で複数ユーザを選出して正解となった 7 件の質問に対しての結果を表 3 に示す．

表 3: 関連検策ワードを用いた選出手法の正解率

	全質問	選出ユーザが複数の質問
正解率	40.0 % (12/30)	42.8 % (3/7)

関連検索ワードを用いた手法では全ての質問に対し，選出するユーザを一人に絞ることができた．これにより，知識の数が少ないことによる問題を改善できたと言える．表 3 より関連検索ワードを用いた手法は，提案手法と比較して正解数 1 件のみの減少となった．複数のユーザを選出している質問が 0 件になったことを考慮すると，精度が良くなっていると考える．

提案手法の複数ユーザを選出して正解となった質問の正解率は 42.8 %と約 4 割の正解率となっている．これは，提案手法では質問に回答する為の情報が足りていなかったことを示している．その他の 23 件の質問に対しては，提案手法では不正解となったが関連検索ワードを用いた手法では正解となっている質問が 5 件あった．これは，回答する為に必要な知識が増えたことにより，質問に適したユーザを選出しやすくなったと考える．

5 おわりに

本研究では，質問文から抽出した専門用語を適するユーザを探す為の知識として扱い，最適な回答者を選出する手法を提案した．しかし，回答者選出の正解率が低い結果となった．要因として質問文から抽出できる知識が少ないことが挙げられる．この問題を改善する為に専門用語の関連検索ワードを収集し，知識を拡張して追加実験を行った．結果より多少の改善は見られたが，精度が低くこのままでは本研究の目的を達成出来ないことが分かった．

今後の課題は，大量の質問を対象に実験を行うことが挙げられる．次に人手で抽出した専門用語を自動で抽出すると共に，専門用語として適する語のみを拡張する手法を考える．そして新たに知識とは何かを考える必要がある．他に知識を表現する方法としては，係受けを用いる方法が挙げられる．単語単位では知識としての基準を満たしていない内容語が，最適な回答者を提示する為に必要な知識となり得る．

使用した言語資源及びツール

(1) Yahoo!知恵袋. http://chiebukuro.yahoo.co.jp/

(2) 形態素解析器 MeCab, Ver.0.9.1. 京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクト, http://mecab.sourceforge.jp/

(3) Yahoo!JAPAN デベロッパーネットワーク関連検索ワード Web API. http://developer.yahoo.co.jp/

参考文献

[1] 片山亮, 川村秀憲, 鈴木恵二. QA サイトにおける質問推薦へ向けた履歴データの分析. 電子情報通信学会信学技報 Vol.109,No.394,pp.11-16,2010.

[2] 甲谷優, 岩田具治, 塩原寿子, 藤村考. QA コミュニティにおける複数情報源を用いた効果的な質問推薦. 情報処理学会論文誌 : データベース (TOD)Vol.3,No.4,pp34-46, 2010.

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.