5Q-3

Q&Aサイトのテキストデータを用いた検索キーワード候補の抽出とその評価

田中 友二† 高橋 寛幸* 徳永 幸生† 杉山 精‡
芝浦工業大学† ＮＴＴレゾナント株式会社* 東京工芸大学‡

1. はじめに

近年，大量の情報が配信されている World Wide Web 上から情報を効率的に入手する手段として，種々の検索エンジンが利用されている．しかし，検索者がいつでも検索目的に適した検索語を思いつくとは限らない．そこで，ユーザは検索エンジンで情報入手ができなかった場合，Q&A サイトでの質問の投稿や，質問回答データの閲覧を行う．

Q&A サイトの質問は情報検索型，社会調査型，非質問型の 3 つに分類することができる[1]（表 1）．その中で，客観的な情報を求める情報検索型質問は同様の情報を知りたい検索者にも有用であると考えられる．そこで，情報検索に利用することを念頭に，Q&A サイトの質問回答データの分析を進める．

本稿では，情報検索型質問のテキストデータから，検索エンジンへ入力される検索語に関連する語を抽出し評価する．具体的には，さらに，抽出に利用された質問回答文から，Q&A サイトの利用のされ方を考察し，情報入手の支援手法を提案する．テキストデータにおける単語の出現頻度や閲覧者の評価値を用いて関連語を自動抽出し，その関連語を既存検索エンジンのサジェスト語と比較評価する．

表 1. Q&A サイトにおける質問の分類

質問の型	詳細
情報検索型	客観的な事実や情報を求める質問 「人名」,「エラーの解決方法」など
社会調査型	個人的な助言・意見・経験などを求める質問 「推薦」,「助言」など
非質問型	・記述として何が書かれているのか 分析者に理解できなかった質問 ・質問者の主張に対する反応を求めている質問

Extraction of Web Search Keywords Using Q&A Text Data and Its Evaluation
†Yuji TANAKA(ma11105@shibaura-it.ac.jp)
*Hiroyuki TAKAHASHI（h-taka@nttr.co.jp）
†Yukio TOKUNAGA（tokunaga@shibaura-it.ac.jp)
‡Kiyoshi SUGIYAMA
†Shibaura Institute of Technology
*NTT Resonant Inc.
‡Tokyo Polytechnic University

2. 関連語の抽出アルゴリズム

Q&A サイトの質問回答データには閲覧者が役に立ったと評価したときにクリックするボタンが存在する．そして，各質問回答データごとにボタンを押された累計回数が明示されており，値が高いほど役に立っているといえる．そこで，本研究ではこの値を「役に立った件数」（閲覧者の評価値）として関連語の抽出に利用する．抽出は以下のステップで行い，算出したスコアの合計が高い語を有用な関連語とする．

①各質問文を形態素解析し名詞のみを抽出
②任意の検索語（単語）を含む質問文を抽出
③抽出した各質問文の単語に
　その質問回答データへ与えられている
　役に立った件数をスコアとして付与
④抽出した質問文において
　同じ単語同士のスコアの合計を算出

このアルゴリズムで抽出される語は役に立った件数が多い質問に含まれることとなり，閲覧者の評価が強く反映されていると言える．

3. 関連語の抽出

2 章のアルゴリズムを用いて関連語を抽出した．本研究では，Q&A サイトの 2001 年 11 月 11 日から 2012 年 1 月 31 日までに「国内旅行（全国）」「関東」「関西」カテゴリに投稿された 72,840 件の質問回答データから，まず情報検索型質問 9,242 件を自動抽出する．次に，これを関連語の抽出に用いた．質問の自動抽出は文献[2]の手法を用いた．

検索語を「飛行機」としたときの抽出結果（左列）と既存検索エンジンでのサジェスト語（右列）を表 2 に比較して示す．傾向として，既存検索エンジンのサジェスト語に比べ抽象的な語が並んでいる．例えば，「旅行」や「航空」などである．これは，知りたいことを十分に明確にできていない人が質問するためと考えられる．

一方で，「チケット」や「予約」など既存検索エンジンのサジェスト語と重複する関連語も存在した．これは，多くの人にとってニーズの強い話題であると考えられる．

表 2.　抽出した関連語とサジェスト語

東京	予約
旅行	チケット
航空	格安
チケット	手荷物
国内線	時刻表
空港	運行状況
機内	持ち込み
予約	乗り方

4. 評価された質問回答データの分析

3 章より，Q&A サイトの質問回答データを用いると主に抽象的な関連語が抽出された．

そこで，抽出に利用した閲覧者の評価値が高い質問を分析して，Q&A サイトと検索エンジンの利用のされ方の違いを明らかにし，検索者の支援手法を検討する．

4.1 質問内容の分析

分析対象は検索語（飛行機，ANA，JAL，新幹線，バスの 5 語）とその関連語（上位 8 語）を含む評価された質問（上位 10 件）の合計 400 件の中で，重複などを除く 164 件とした．

分析データを質問者の質問動機は何かという観点で分類した．その結果，6 種類の動機に基づいて質問をしていることがわかった（表 3）．

表 3.　質問動機に基づく分類結果

	質問の動機	件数
①	あいまいな表現でしか言語化できず 検索語を思いつけない	22
②	適切な検索語を思いつけず検索で調べることが困難	42
③	検索しても Web 上から十分に情報が得られない	18
④	複数の知りたいことがあり検索では時間がかかる	28
⑤	検索できないが適切な機関に問合せればわかる	23
⑥	規約違反の可能性があり適切な機関に 問合せできない	13
⑦	その他	18

最も多かったのは②である．ここに分類された質問は知りたいことが一般的な検索語では得られないため，質問していると考えられる．例えば，「高松と東京間の新幹線料金を知りたい」という質問があった．この場合，高松に新幹線は開通しておらず，別手段で岡山に行きそこから新幹線に乗車することとなる．この要求を「高松　東京　新幹線」として検索すると，飛行機を利用した行き方や新幹線の割引情報のみが上位の検索結果に表示され，必要な情報を入手できない．しかしながら，下位の検索結果まで閲覧すると質問内容に合致する質問回答データの Web ページに辿り着け情報入手ができる．

4.2 情報要求の 4 階層

前節の質問動機と検索エンジンの利用のされ方を Taylor の情報要求の 4 階層（図 1）[3]に対応づけ，情報入手手段を整理する．

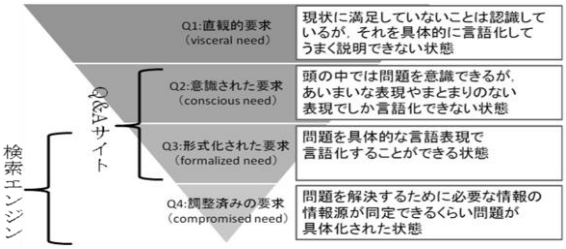

図 1.　情報要求の 4 階層

表 3 の動機の①は意識された要求（図 1 の Q2）に相当し，②～⑥は形式化された要求（同 Q3）に相当すると考えられる．そのため，Q&A サイトの利用者は Q2 や Q3 の状態で利用していると言える．一方，検索エンジンは形式化された要求（同 Q3）または調整済みの要求（同 Q4）に相当する．

人間の情報要求は直感的要求（同 Q1）から調整済みの要求（同 Q4）に向かって徐々に段階的に具体化されるが，人間の情報要求を満たすためには，問題の言語化ができる Q2 までの情報要求に対応した情報入手支援が必要と考えられる．そのため，既存検索エンジンで情報入手ができなかった人には，抽出した関連語と質問タイトルを利用し，質問回答データを提示することで，自然文による検索者の要求の具体化を支援することが有効だと考えられる．そこで，質問回答データを用いた情報入手支援システムを構築した（図 2）．本システムでは，自然文の情報から検索者の要求が具体化されることを期待できる．

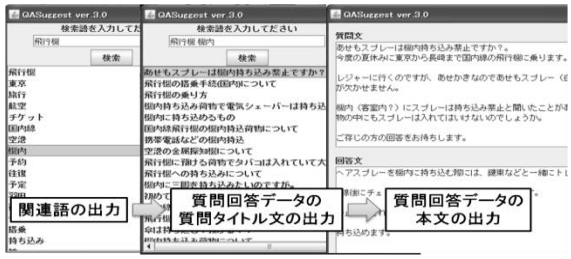

図 2.　情報入手支援システム（飛行機の例）

5.おわりに

本稿では，検索エンジンに入力される検索語と関連する語の抽出及びその評価を行ったが，抽象的な語が抽出された．そこで，関連語の抽出に利用した質問を分析し，Q&A サイトでの質問の動機を明らかにし，情報要求の 4 階層を参考に，情報入手の支援システムを構築した．これにより，情報入手の支援を自然文を用いて行える見通しを得た．

――― 参考文献 ―――

[1] 栗山和子，神門典子：Q&A サイトにおける質問と回答の分析，情報処理学会研究報告，Vol.2009-FI-95 No.19

[2] 田中友二，ほか．Q&A サイトにおける情報検索型質問の自動抽出：第 74 回情報処理学会全国大会,2012 年

[3] S.Taylor , Question-Negotiation and Information Seeking in Libraries. : College & Research Libraries, 178-194, 1968.