D-013

Q&A サイトにおける情報検索型質問の自動抽出とクラスタリング

Automatic Extraction and Clustering of Factual Information in Q&A Site

田中 友二†　　徳永 幸生†　　杉山 精‡
Yuji Tanaka†　Yukio Tokunaga†　Kiyoshi Sugiyama‡

1. はじめに

近年，大量の情報が溢れる World Wide Web(以下，WWW)上から情報を効率的に入手する手段として，WWW 検索エンジンを利用することが多い．そのため，WWW 検索者を支援する様々な研究が行われている．その 1 つに検索ログの分析に基づいた研究がある．しかし，検索ログは単語の集合であり，分析できる情報が少ない．また，WWW 検索者が検索行動を終了した際に，検索結果に満足しているか不明である．これらのことから，情報要求を検索ログから推測することは困難である．

一方で，WWW 検索エンジンを利用しない情報の入手方法として Q&A サイトがある．Q&A サイトとは，質問者が自然文で書いた質問に対して回答者が自然文で回答する，人同士の知識の共有をベースにしたサービスである．質問者は情報要求を質問として自然文で表現するので，検索語に比べて意図を表現しやすい．また，あいまいな表現や誤った言葉を用いてしまった場合でも，回答者が適切に解釈し回答することが期待できるので，質問者が満足する回答を得られることが多い．さらに，質問者は回答に評価を行うため，満足した回答か否かを明解に判断することができる．

WWW 検索エンジンを利用する場合と，Q&A サイトを利用する場合では，入出力される情報の形式は異なる．しかし，どちらも情報入手の手段であり，その情報入手の手順が共通する．このことから，検索語を入力し検索結果を得ることと，質問し回答を得ることは，対応付けられる．そのため質問者の情報要求や満足した回答の実例を分析することで，WWW 検索者を支援するための知見が得られると考えられる．

本稿では，Q&A サイトの質問回答ログから情報検索型質問を自動抽出し，クラスタリングする．そして，情報検索型質問から WWW 検索エンジンでの検索が試行錯誤する理由と提示するべき情報を探り，WWW 検索者への支援が必要な情報要求を明らかにする．

2. 質問の分類

Q&A サイトには様々な種類の質問が存在する．そこで，栗山ら[1]の分類に従い表 1 の 3 種類に質問を分類した．

3 種類の質問の中で，情報検索型質問では質問者が WWW 検索エンジンでは調べられないことについて質問している．そこで，情報検索型質問とそれに対する回答から，試行錯誤する理由と提示するべき情報を探ることにより，WWW 検索者を支援するための知見が得られると考えられる．

†芝浦工業大学, Shibaura Institute of Technology

‡東京工芸大学, Tokyo Polytechnic University

表 1.　Q&A サイトにおける質問の分類

質問の種類	詳細
情報検索型	検索エンジンや図書館のレファレンス・サービスなどを利用して回答を探すことが可能な質問 「人名」，「エラーの解決方法」など
社会調査型	特定の個人あるいは集団に対してアンケート調査を行うことで回答を得るような質問 「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

3. 情報検索型質問の自動抽出

情報検索型質問を抽出するために，回答文における評価をあらわす表現（以下，評価表現）の割合と特有のキーワードの有無の 2 つの要素を用いる．

3.1　評価表現の割合

情報検索型質問に対する回答には，「きれい」「役立つ」などの評価表現が使われにくいと考えられる．そこで，回答文に対して，MeCab[2]を用いて形態素解析を行い，回答 a の形態素群 X_a と評価表現辞書[3]の形態素群 Y を利用し，式(1)で得られる S_a の値が低い質問を情報検索型質問とした．

$$S_a = \frac{|X_a \wedge Y|}{|X_a|} \quad \cdots (1)$$

3.2　キーワードの有無

情報検索型質問に使われやすい，または使われにくい単語があると考えられる．そこで，目視で分類したデータ 100 件から頻出語を抽出し 3 種類の質問に特有のキーワードがないか調査した．その結果，質問文に「おすすめ」を含む質問とお礼文に「思い」を含む質問は情報検索型質問でない場合が多いということが分かった．そのため，それらを含む質問を情報検索型質問から除外した．

3.3　抽出実験

上記の，2 つの要素を用いて情報検索型質問が抽出できるか評価実験を行った．Q&A サイトには予め提供者が用意したカテゴリが存在する．今回は，交通情報に関する質問が投稿されている「国内旅行（全国）」カテゴリを対象にする．WWW 検索エンジンの利用目的として，交通情報を求める場合が多く存在するからである[4]．実験のために，「国内旅行（全国）」カテゴリから 100 件無作為に質問を抽出した．そして，目視で分類した質問を正解データとして，自動分類の結果を評価した．その際の評価値として，適合率と再現率を用いた．適合率と再現率は式(2)のように定義する．実験結果を表 2 に示す．また，質問回答ログは 2010 年 5 月 23 日から 6 月 13 日までに投稿されたデータ 100 件を用いた．

$$(\text{適合率}) = \frac{R}{N} \quad , \quad (\text{再現率}) = \frac{R}{C} \quad \cdots (2)$$

R：適切に分類された情報検索型質問の件数
N：情報検索型質問と分類された質問の件数
C：評価に用いた情報検索型質問の件数

表 2.　抽出実験結果

	キーワードを含む質問を除外した場合			キーワードを含む質問を除外しない場合		
	適合率	再現率	F値	適合率	再現率	F値
S_a<0.005	0.867	0.464	0.605	0.867	0.464	0.605
S_a<0.010	0.889	0.571	0.696	0.889	0.571	0.696
S_a<0.015	0.905	0.679	0.776	0.833	0.714	0.769
S_a<0.020	0.885	0.821	0.852	0.636	0.750	0.689
S_a<0.025	0.767	0.821	0.793	0.558	0.857	0.676
S_a<0.030	0.571	0.857	0.686	0.455	0.893	0.602

S_a の値が 0.02 未満でキーワードを含む質問を除外した場合が最も高い調和平均値（F 値）となった．

適合率が低くなる抽出ミスの原因として，情報検索型以外の質問において，回答者が参考となる URL の紹介による回答を行ったため評価表現が少ないことが考えられる．また，再現率が低くなる抽出漏れの原因として，情報検索型質問において個人的な意見を回答に付加したため評価表現が多くなったことが考えられる．

4. 質問のクラスタリング

抽出した情報検索型質問において類似した内容ごとに分類するためにクラスタリングを行う．湯浅ら[5]の手法を用いて質問文の特徴ベクトルを生成し，Repeated Bisection 法を採用しているクラスタリングツール“bayon”[6]を用いる．また，Q&A サイトの質問には予めタグが付与されている．同じタグが付与されている質問は類似しているため，同じタグをもつ質問に対してクラスタリングを行った．これにより，詳細な内容で分類ができる．

5. クラスタと検索ログの分析

「国内旅行（全国）」カテゴリで，「飛行機」「新幹線」「バス」のタグが付与されている質問に対してそれぞれクラスタリングを行った．「飛行機」の質問におけるそのクラスタリング結果を図 1 に示す．

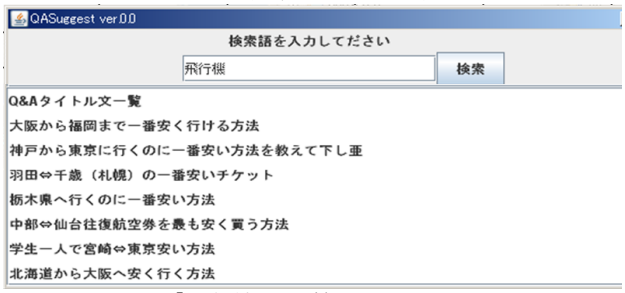

図 1.　「飛行機」の質問におけるクラスタ

これらの，交通手段に関する質問においては「ある地点からある地点まで安く行きたい」というクラスタが共通して存在することがわかった．

さらに，WWW 検索エンジンで試行錯誤する理由や求めている情報を調査した．「ある地点からある地点まで安く行きたい」という質問群において，最適な交通手段が調べられないと記述されている質問が「飛行機」「バス」の質問に多く見られた．また，割引の情報を提供している回答が「飛行機」「新幹線」の質問に多く見られた．最適な交通手段が不明な質問の割合を表 3 に示し，割引情報を求めている質問の割合を表 4 に示す．これらのことから，質問者の多くは WWW 検索エンジンで最適な交通手段が調べられないため質問を投稿し，割引を利用した交通手段で安く行ける情報を求めていると考えられる．

表 3.　最適な交通手段が不明な質問の割合

	「飛行機」の質問	「新幹線」の質問	「バス」の質問
最適な交通手段が不明という質問 / 「安く行きたい」という質問の数	14/21	2/22	25/35

表 4.　割引情報を求めている質問の割合

	「飛行機」の質問	「新幹線」の質問	「バス」の質問
割引の情報を提供している回答 / 「安く行きたい」という質問の数	15/21	17/22	3/35

次に，実際に WWW 検索者が格安の交通手段を求めて試行錯誤を行っているか 2010 年 3 月 1 日から 7 日までの検索ログを調査した．ここで，WWW 検索者が 1 時間以内に検索語の組み合わせを変更し 3 回以上入力を行っていることを試行錯誤とする．WWW 検索者の試行錯誤の例を表 5 に，調査した結果を表 6 に示す．

表 5.　試行錯誤をしている WWW 検索者の例

検索回数	検索日時	検索語		
		1語目	2語目	3語目
1回目	03-04 21:25:49	新幹線	格安	
2回目	03-04 21:26:09	新幹線	格安	京都
3回目	03-04 21:36:31	新幹線	格安	京都
4回目	03-04 21:39:23	格安新幹線	京都往復	
5回目	03-04 21:41:27	格安新幹線	京都往復チケット	

表 6.　試行錯誤をしている WWW 検索者の割合

	3/1（月）	3/2（火）	3/3（水）	3/4（木）	3/5（金）	3/6（土）	3/7（日）	合計
試行錯誤人数 / 格安の交通手段を求めている人数	6/22	6/16	5/8	3/8	0/7	5/9	1/6	26/76
割合	0.273	0.375	0.625	0.375	0.000	0.556	0.167	0.342

平均 3 割程度の WWW 検索者が試行錯誤していることがわかった．以上のことから，交通手段に関する情報を求めている WWW 検索者に対して支援を行う必要があり，割引情報を提供することが有効であると考えられる．

6. まとめ

本稿では，Q&A サイトの質問を分類抽出し，クラスタリングすることで質問を構造化した．中でも交通手段に関する質問を取り上げ，そのクラスタを分析することで質問者が試行錯誤する理由や求めている情報がわかった．そして，Q&A サイトから得られた知見の確度を調べるために検索ログを調査することにより，実際に試行錯誤している WWW 検索者が存在することを示した．

謝辞

本研究の遂行にあたって御指導いただいた，NTT レゾナント株式会社の望月崇由氏，松田達樹氏に厚くお礼申し上げます．

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] MeCab, http://mecab.sourceforge.net/
[3] 評価表現辞書, http://www.syncha.org/
[4] インターネット白書 2010, 資料 7-1-3 インターネットの利用目的
[5] 湯浅夏樹, 上田徹, 外川文雄：大量文書データ中の単語間共起を利用した文書分類, 情報処理学会論文誌, Vol.36, No8, 1995.
[6] bayon, http://code.google.com/p/bayon/