D-026

# Q＆Aサイトを対象とした地域情報の抽出法とその応用

## Extraction Method and Its Application of Regional Information for Q & A Text Data

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

### 1. はじめに

近年，大量の情報が配信されるWeb上から情報を効率的に入手する手段として，一般的にキーワードを入力する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，質問回答ログの閲覧などの利用が期待される．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ナレッジコミュニティである．このQ&Aサイトでは具体的に求める情報を言葉として表現しているため，同様の情報を求める他の人にとっても，わかりやすく有用な情報となることが多い．

そこで，本稿ではグルメ・交通・観光など地域に関する情報を対象に大量の質問回答データから有用な地域情報の自動抽出を試みた．さらに，そのデータを用いて地域情報を提示する利用法について提案する．

### 2. 有用な地域情報の自動抽出

利用価値のある地域情報は，Q&Aサイトにある一部の有用な質問回答データに含まれており，それには以下の性質があると考えられる．

・多くの人が疑問に思うことについて質問と回答がなされている
・質問の意図がわかりやすい
・回答が的確である
・回答に根拠がある

また，定量的な値としては以下のものが利用できる．

・ベストアンサー（以下，BA）の有無
（BAとは質問者が最も満足した回答に付与する指標）
・閲覧者が評価した件数
・閲覧者の閲覧数

そこで，有用な地域情報が含まれている質問回答データとして，「タイトル文に地名を含んでいる」かつ「BAが付与されている」ものを自動抽出することとした．

### 3. 抽出実験

質問回答データの地域情報を提示する際，古いデータではお店などが変わっている場合がある．従って，可能な限り新しいデータであることが望まれる．そこで，任意の地名で有用な地域情報が含まれる質問回答データを 10 件取得するために必要な期間を調査した．具体的には，ある投稿日時から遡り，10 件の質問回答データが抽出されるまでの期間を出し，更に，その後 10 件抽出されるまでの期間を出すことを 5 回分繰り返した．これにより，質問回答データを 10 件取得するために必要な平均的な期間を調査した．また並行して，抽出した質問回答データの内容の分析も行った．

†芝浦工業大学, Shibaura Institute of Technology
‡東京工芸大学, Tokyo Polytechnic University

調査結果を表 1 に示す．なお，本研究では，Q&A サイト[1]の 2001 年 3 月 14 日から 2012 年 4 月 30 日までに「国内旅行（全国）」「関東」「遊園地・テーマパーク」「食べ歩き（全国）」「お酒」「お茶・ドリンク」「レストラン・ファミレス」「カフェ・喫茶店」「その他（料理・グルメ）」「季節の行事」「正月・年末年始」のカテゴリに投稿された約 12 万件の質問回答データを用いた．また，任意の地名として「東京」「横浜」「新宿」「池袋」「銀座」の 5 ヶ所を対象とした．

表 1.　10 件の質問回答データを得るのに要する期間
（単位：月）

| 地名＼件数 | 1～10 | 11～20 | 21～30 | 31～40 | 41～50 | 平均 |
|---|---|---|---|---|---|---|
| 東京 | 1 | 1 | 1 | 1 | 1 | 1 |
| 横浜 | 3 | 2 | 4 | 1 | 2 | 2.4 |
| 新宿 | 4 | 3 | 2 | 2 | 3 | 2.8 |
| 池袋 | 8 | 6 | 4 | 4 | 3 | 5 |
| 銀座 | 6 | 9 | 6 | 5 | 6 | 6.4 |

表 1 より過去 1 年のデータを対象とすれば BA が付与された質問回答データを 10 件以上抽出できる見通しを得た．

次に，抽出した質問回答データを話題に基づいて分類した．その結果を表 2 に示す．また，各話題で多かった質問回答のパターンを表 3 に示す．

表 2.　話題による質問回答データの分類

| 地名＼話題 | グルメ | 経路 | 観光 | 場所 | 宿泊 | テーマパーク | その他 |
|---|---|---|---|---|---|---|---|
| 銀座 | 38 | 10 | 1 | 1 | 0 | 0 | 0 |
| 池袋 | 27 | 6 | 4 | 4 | 0 | 2 | 7 |
| 新宿 | 21 | 13 | 2 | 7 | 0 | 1 | 6 |
| 横浜 | 14 | 17 | 10 | 5 | 1 | 0 | 3 |
| 東京 | 6 | 11 | 10 | 5 | 8 | 4 | 6 |
| 合計 | 106 | 57 | 27 | 22 | 9 | 7 | 22 |

表 3.　各話題で多かった質問回答のパターン

| | 質問 | 回答 |
|---|---|---|
| グルメ | ある地名周辺における<br>オススメのお店を知りたい | 複数のお店を推薦 |
| 経路 | ある地点からある地点へ<br>の行き方を知りたい | その行き方を説明 |
| 観光 | ある地名における<br>観光の計画を<br>立ててほしい | 観光地を巡る<br>スケジュールを提案 |
| 場所 | 駐車場やカラオケなどの<br>有無や場所を知りたい | その有無や行き方を<br>説明 |
| テーマパーク | あるテーマパーク<br>（ディズニーランドなど）内<br>を回る計画を立ててほしい | テーマパーク内を<br>回るスケジュールを<br>提案 |
| 宿泊 | ある地名周辺における<br>オススメの宿泊施設を<br>知りたい | 複数の宿泊施設を<br>推薦 |

全ての地名でグルメ，経路に関する質問回答データが 5 件以上抽出された．グルメに関する情報はクチコミサイトでも入手可能であるが，Q&A サイトでは質問者が自由な言葉で条件を設定できるため，多く質問されていたと考えられる．また，経路に関する質問回答データは徒歩やバスを利用した移動方法や所要時間など，既存の路線検索サイトでは入手困難な情報について質問されていた．

以上のように，具体的に求める情報を言葉として表現できる Q&A サイトの利点を活かした質問を抽出できた．これらは質問に対して質問者が満足した回答が付与されており，同様の情報を求める他の人にとっても，わかりやすく有用な地域情報を含む質問回答データとなっていた．また，回答には参考となる URL が記載されていることもあり，それらを閲覧することでさらに詳細な情報が得られる場合も多かった．このことは BA が付与されたデータのみを抽出したことが寄与していると考えられる．

## 4. 質問回答データの利用法の提案

地域情報を求めている場合，その地域名を入力することは容易である．そこで，地名を入力することで有用な地域情報が含まれる質問回答データを提示する手法を提案する．その際，3 節で述べたように質問回答データでは様々な話題が扱われているため，その話題ごとに分類して提示することでユーザが求めている情報に辿り着きやすくなると考えられる．これを考慮した地域情報の入手手法として，以下の手順を提案する．また，その①～③までを実装したプロトタイプシステムを試作した（図 2）．

①地名の入力
・ユーザが求めている地域名（例：横浜）を入力
②話題の選択
・ユーザが求めている話題（例：グルメ）を選択
③質問回答データの選択
・システムは入力された地名において選択された話題に関する質問回答データのタイトル文を提示
・ユーザはタイトル文から興味のあるものを選択
④地域情報の検索
・ユーザは選択したタイトル文の質問と回答を閲覧し興味のある地域情報（例：飲食店）を Web 検索
⑤情報入手
・ユーザは検索結果から詳細な地域情報を入手

提案システムは，ユーザが①と②を行うことで該当する質問回答データのタイトル文の一覧を表示する．さらに，ユーザはこの一覧の中から興味のあるものを選択することで，有用な地域情報を含む質問回答データを閲覧することができる．その店舗名などの情報をキーワードとして Web 検索することで，ユーザは店舗の位置情報やさらなるクチコミ情報などを入手できると考えられる．このように，提案システムは十分に知りたいことを具体化できていないユーザが対象となる．また，タイトル文は質問回答データの内容を要約した自然文と考えられ，既存のクチコミサイトやブログなどよりわかりやすく情報入手ができる．

同種の地域情報サービスとして，クチコミ情報や質問回答データの地域情報を横断検索できる「NAVER スポット」[2]，Q&A サイトの質問回答データを人手でまとめて新たなコンテンツとする「知ってトクする！べんり Q&A」[3]などがある．

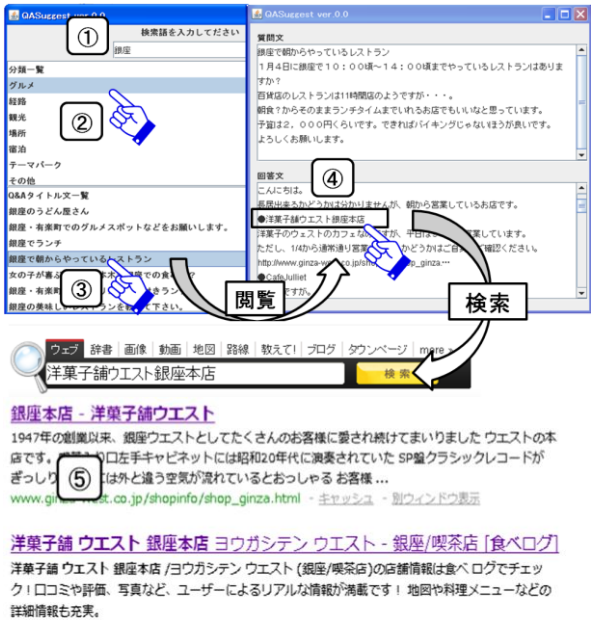


図 2. プロトタイプシステムと情報入手の流れ

「知ってトクする！べんり Q&A」は，カテゴリの選択または検索キーワードを入力し，人手でまとめた質問回答データの情報を入手する．このサービスは，提供者が手動でライフや旅行・レジャーなどのコンテンツを作成しているため，テーマに沿った情報が集められているが作成に手間がかかると考えられる．一方，提案システムでは自動で有用な地域情報を含む質問回答データを収集するため手間が少ない．そのため，十分な量の質問回答データがあれば容易にコンテンツを生成できる．

しかし，提案システムの検討課題は二つある．

一つ目は，収集した質問回答データを話題ごとに自動分類することである．サービスとして提供する際に，質問文などを利用して適切な話題ごとに自動でクラスタリングすることを検討する必要がある．

二つ目は，十分な量の質問回答データを収集することである．今回は，首都圏の地名を対象としたが，他の地名では質問回答データの量が十分に集められない可能性がある．対策としては，対象とする Q&A サイトやカテゴリを増やすことが考えられる．

## 5. おわりに

本稿では，質問回答データを対象として「地名」と「BA」の有無を利用して有用な地域情報の自動抽出を試みた．その結果，質問者が満足した回答が付与されており，誰にとっても，わかりやすく有用な地域情報を含む質問回答データを抽出できた．さらに，そのデータを話題ごとに提示し，地域情報を与える手法を提案した．

今後は，今回対象とした地名以外での抽出実験や提案したシステムの改良と評価を行う．

――― 参考文献 ―――

[1] 教えて！goo，http://oshiete.goo.ne.jp/
[2] NAVER スポット，http://spot.naver.jp/
[3] 知ってトクする！べんり Q&A，http://oshiete.goo.ne.jp/benriqa/
[4] 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とクラスタリング：FIT2011 第 10 回情報科学技術フォーラム，D-013(Sep. 2011)
[5] 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