C-16

質問応答コンテンツにおける補完情報分析と検索質問生成

Complementary Information Analysis and Query Formulation for Question-Answer Contents

高田 夏希 †　　山本 祐輔 †　　小山 聡 †　　田中 克己 †
Natsuki Takata　Yusuke Yamamoto　Satoshi Oyama　Katsumi Tanaka

1. はじめに

疑問に思ったことを Web を用いて調べる方法として Yahoo!知恵袋[1]や教えて!goo[2]などの質問応答サイト（以下 QA サイト）を利用するユーザが増えている [1]。Web 検索エンジンなどを用いて疑問に思ったことを調べる場合、検索エンジンに投入するクエリをうまく選択しなければ望ましいページが得られない。一方、QA サイトは、ユーザが疑問に思ったことをそのまま自然文形式で投稿すると、それに関する回答を知っている別のユーザが自由に回答に答える、という形式を取っており、回答が容易に得られることから近年人気を博している。

しかし、QA サイトはユーザの疑問に対する回答を容易に取得できる一方で、得られる回答の質は必ずしも保証されていない。QA サイトにおける回答は不特定多数の QA サイト利用者が個々人の独自の判断で掲載した情報である。また Adamic や佐藤らの調査によると質問のトピックによって回答可能なユーザの数が少ない、数が多くともユーザの回答活動が活発でないことなどが報告されている [2][3]。このため、（１）質問に対して与えられた回答が正しいかどうかが分からない、（２）ユーザの回答が質問に対する回答として別解が不足している、根拠が不足しているなど十分な回答となっていない可能性がある。例えば、Yahoo!知恵袋では回答可能な期間が制限されており、質問者が回答を締め切るとたとえ有益な知識を持つユーザが存在しても疑問解決のための情報を収集することができないケースが存在する。

回答の質を評価するために、質問応答サイトの多くは質問者が自身の質問に対してなされた回答群から最も満足のできる回答をベストアンサーとして投票する、という仕組みを導入している。しかし、ベストアンサーは質問者の独断で選ばれるため一般的に質の高い回答がベストアンサーとして選ばれる保証はない。また、質問に対する回答としてベストアンサーのみで十分な情報を確保できるとは限らない。上記の問題は質問応答コンテンツ（以下 QA コンテンツ）に「コンテンツの正しさを保証する証拠」情報が欠けていることや、「回答となりうる情報がコンテンツ外に存在する」ことによって生じると考えられる。よってコンテンツに欠けた情報を補完することで上記の問題が解決するものと思われる。

そこで、本論文では質問応答コンテンツの情報補完について考察する。コンテンツに対しどのような情報の補完が考えられるかを考察すると共に、補完情報を Web から取得することを考えるときどのような検索質問（以下クエリ）を検索エンジンに与えるのが適切かを考える。

2. 質問応答サイトのコンテンツ

本論文では QA サイトとして Yahoo!知恵袋を対象としている。ここで Yahoo!知恵袋について説明する。

ユーザが質問を投稿すると、別のユーザがそれに対し回答を投稿する。その質問の回答期限が過ぎるか回答の中からベストアンサーが選ばれると、それ以上その質問には回答ができなくなる。ベストアンサーは質問者が決めるか、閲覧ユーザの投票で決まる。こうして一つの質問応答コンテンツが作られる。

一方、同様の QA サイトである教えて!goo のコンテンツについても説明する。こちらも、基本はユーザが質問・回答を行うものである。知恵袋と主に異なる点は 1. 質問者が各回答に補足を付与できる、2. ベストアンサーに当たる良回答のほか、次点としてもう一つ回答が選べる、3. 質問は質問者が締め切らない限り永久に回答受付中となるの 3 点である。

各サイトで質問応答コンテンツが生成される過程は異なるが、コンテンツが一つの質問 Q と、それに対する回答集合 $A_1,A_2,\ldots$ で成り立つことは共通である。

3. QA コンテンツの種類と補完情報に関する考察

本節では QA コンテンツに対する情報補完を行うために、QA コンテンツの種類と補完すべき情報との関係を考察する。

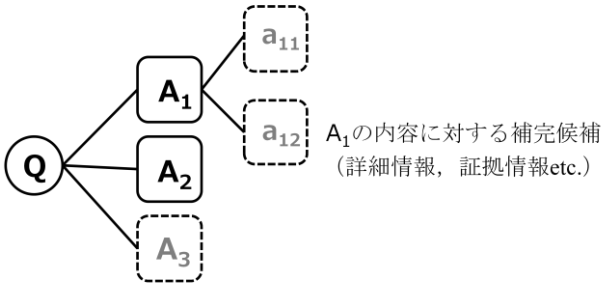

図 1：QA コンテンツの構成と補完情報

† 京都大学大学院情報学研究科

[1] Yahoo!知恵袋: http://chiebukuro.yahoo.co.jp/

[2] 教えて!goo: http://oshiete.goo.ne.jp/

3.1 補完情報についての考察

QA コンテンツに対し情報を補完するということは、コンテンツに不足している情報をコンテンツ外から取得し、その情報でコンテンツを補うことを意味する。

コンテンツの補完は 2 種類存在すると考えられる。QA コンテンツは一つの質問 Q とそれに対する回答 A_1、A_2, ... ,A_n から構成される（図 1）。QA コンテンツによっては、図 1 のように本来別解として考えられるべき回答 A_3 が記述されていないケース、回答 A_1 の a_{11},a_{12} のように回答についての説明や回答を裏付ける情報が不足しているケースが存在する。前者のケースにおける情報補完とは、ある質問内容について回答となりうる情報がすべてコンテンツに含まれるわけではないため、含まれなかった情報（図 1 における A_3）を補完し質問に対する回答の網羅性を向上させるということである。後者のケースに対する情報補完とは、一つの回答についてより詳細な補足情報や証拠情報（a_{11},a_{12}）を与える補完である。

3.2 質問応答コンテンツのタイプ分類

多くの QA サイトでは、質問応答コンテンツは質問の内容に基づきカテゴリ分けされている。さらにカテゴリは階層化されていることが多い。例えば Yahoo!知恵袋では「犬の無駄吠えをやめさせるには？」という質問は「暮らしと生活ガイド」カテゴリの「犬」サブカテゴリに分類されている。

本研究では質問のタイプに応じて補完すべき情報として「回答に対する補足情報」もしくは「質問に対する別解」のいずれかを選択し提示することを想定している。このとき提示すべき補完情報は、上記に揚げた質問の内容カテゴリよりも“色々な意見を求めた質問”“証拠が欲しい質問”といったように質問の意図タイプに依存すると考えられる。QA サイトにおける質問の意図分類に関する研究は近年注目されている。Harper らは QA サイトに存在する質問を conversational questions と informational questions の 2 種類に大別している [4]。Conversational question とは質問者と回答者が意見を求めるために質問者と回答者が「会話」「議論」をするようなコンテンツを指す。Informational question は質問者が知識を得るために情報を集めるための質問を指す。Conversational question は情報を得るための質問ではなく質問者と回答者だけの閉じた世界での意見交換用の質問であるため、その QA コンテンツに対する情報補完の必要性は少ないと考える。栗山らは質問意図を「情報検索型」「社会調査型」「非質問型」の観点から、それぞれをより詳細な意図に分類している [5]。情報検索型は「事実」「真偽」「定義・記述」「方法・手段」「原因・理由」「効果・結果」、社会調査型は「助言」「意見」「嗜好」「推薦」「経験」、非質問型は「主張」「理解不能」に分類される。栗山らが提案する情報検索型の質問は Harper らの informational questions とほぼ同様の定義である。本研究では、情報補完対象として情報検索型質問 (informational questions)に限定するが、その中でも「事実」「手順」型などの質問のタイプによって補完情報のタイプ、提示手法、補完情報の集め方を決定することが、満足度の高い情報補完を行う際必要である。

4. 補完情報取得のためのクエリ生成

Web から補完情報を取得する際に必要なクエリ生成について考察する。

4.1 補完情報の種類の違いとクエリ生成

補完情報は前述のように 2 種類に分類できる。その種類ごとにクエリ生成の方法も変わるものと考えられる。

まず回答 A_x の詳細性を高める補完情報であるが、このような情報を取得するクエリは回答 A_x を踏まえる必要がある。よってクエリには A_x に含まれ、A_x の内容を象徴する語が必要である。次に質問 Q に対する回答の網羅性を上げる補完情報取得のクエリについて考える。この場合、補完の対象が個々の回答ではなくコンテンツ全体に対する補完となっている。そのためクエリに用いる語はコンテンツの内容を表す語となる。コンテンツは一つの Q とそれに結び付けられた回答集合 $\mathbf{A=\{A_1, A_2, \ldots, A_k\}}$ で構成されるため、コンテンツの主な内容は質問の内容となる。よってこちらの補完情報取得のクエリには質問を表す語を含める必要がある。

4.2 質問の型によるクエリ生成の違い

前章で述べたように QA コンテンツの質問は informational questions と conversational questions に大別される。本研究で扱う informational questions はすでに述べたように「事実」「真偽」「定義・記述」「方法・手段」「原因・理由」「効果・結果」に分類されるが、これらは疑問詞を用いた表現分類 5W1H と関連がある。なぜ・何・誰・いつ・どこ・どのように という質問に現れる疑問詞で分類するものであるが、この分類も補完情報取得のクエリ生成にかかわるものと思われる。

たとえば、「なぜ～ですか？」という質問は何かの理由を知りたいという意図が存在する。よってそのような質問を含むコンテンツの補完情報を得るクエリに“理由”という語を含めるとより補完情報を得やすくなると考えられる。他にも「どこで」という場所を尋ねる質問の場合には“場所”、「どうやって」という質問では“方法”という語をクエリに含めるなど、コンテンツの質問 Q に含まれない語でもクエリに加えると補完情報検索の再現率が向上する場合がある。

以上のように、質問応答コンテンツに対する補完情報を取得するためのクエリを生成するためには、まずそのコンテンツの内容や質問の型を判別する必要があることがわかる。クエリ生成のために、5W1H 以外にもどのような質問の形式があるか・その形式ごとにクエリに加えるべき語は何かを考察するとともに、コンテンツや回答 A_x を代表する語の取得方法などを今後考える必要がある。

4.3 クエリ生成の手順

補完情報の取得の流れは、クエリ生成を行い、そのクエリで Web 検索を行って得られる検索結果を用いて補完情報を得るというものである。検索結果から補完情報を得る方法として著者らは検索結果の各スニペットに出現する別解を表す語を抽出して別解情報の取得に用いるという手法を提案している[6]。

別解情報だけでなく詳細情報の取得にも言えるが、提案手法では検索結果の各スニペットに補完情報を表す語が含まれることが前提である。よって、そのような検索結果を得るためのクエリ生成についての手順を提案する。

1. QA コンテンツの Q から質問内容を表す語(q1,q2,…)を抽出する(手動)
2. 質問の型に基づき、「Q の質問型を表す語」(t_1,t_2,…)を決める
3. 回答を述べる際に重要な観点である語(a_1,a_2,…)をコンテンツの回答集合から拡張キーワードとして抽出する
4. 質問と回答の両方が含まれる Web ページを取得するために Web ページの構造を考慮したクエリを生成する

まず手順 1.について考察する。以下のようなヒューリスティックが考えられる。

- 質問文に「n_x と n_y の違い」や「n_x と n_y は何が違う」などの表現があればq=(n_x,n_y, 違い)となる。これは2つ以上のものに対する違いを尋ねる質問を想定したものである。
- 質問文に「n_xのn_yとは何(なに、なん)」、「n_x のn_yって何(なに、なん)」という表現があればq = (n_x,n_yとは)とし、上記の「n_xの」という部分がない表現ならばq = (n_yとは)とする。これは5W1H 型でいう What 型であり、何かについて詳細な説明を求める質問にあたる。

他にも、質問Qにおいて疑問詞や?マークを含む文に含まれる語などが質問内容を表す語として考えられるが、それをどのように抽出するかは今後の課題である。よって、現在この手順は手動で語を抽出するものである。

手順2. であるが、これはたとえば

- 質問文に「何故(なぜ)」、「何で(なんで)」、「どうして」という疑問詞が含まれるWhy型質問の場合はt=(理由)という語
- 質問文に「どうやって」、「どうしたら」、「どうすれば」などの疑問詞が含まれる How to 型の質問の場合は t=(方法)という語

のように、質問文には現れないが、検索クエリとして追加すると検索結果の適合率を上げるような語が質問の型によっては存在する。上記の例でいう Why 型における「理由」、How to 型における「方法」などのような語をそれぞれの質問型に基づいてあらかじめ定義するものである。

手順 3. について述べる。回答を述べる際に重要な観点である語(以下、観点語)とは、質問 Q に対する回答集合において回答を象徴するような語を表現するのに共通して用いられる語である。たとえば、質問 Q が「アボカドサラダのレシピを教えてください」というものであったとき、それに対する回答にはアボカドサラダを作る手順が現れると考えられる。その中で、回答 A_1 は「まず、レモンを切って～」という内容で回答 A_2 では「きゅうりを切って加えます」という内容であったとする。それぞれはレモンを用いるレシピときゅうりを用いるレシピであり、回答の内容は異なるが、それぞれの回答を象徴する「レモン」と「きゅうり」は“切る”という観点で結ばれるものである。

回答を表現する際に、上記の例では“切る”というような、回答を述べるにあたって必要な観点が存在すると考えられる。そのような観点語を抽出することが手順 3.の内容である。

抽出方法について述べる。観点語は、QA コンテンツの複数の回答に多く横断して出現するものと考えられる。そのため、コンテンツの回答集合に存在するすべての語について、語がいくつの回答に出現したかを表す answer frequency(af)を求め、その値の高い語を観点語(a_1,a_2,…)として抽出する。

手順 4. の Web ページの構造を考慮したクエリであるが、これはたとえば intitle: (q_1 + q_2 + t_1) intext: (a_1+a_2..)のように既存の検索エンジンの検索オプションを利用して、キーワードが Web ページのどこに含まれるかを指定するクエリを指す。このような検索手法を用いて検索精度を向上する研究として小山らの研究[7]がある。

著者らの提案手法[6]では、単純なキーワードの AND 検索を用いていたため、キーワードを含むページが多く hit し、求める情報を含む Web ページがその他のキーワードを含むだけのページに埋もれてしまうという問題があった。そこで、キーワードが Web ページに含まれる場所を考慮することを提案する。取得するべきページは、補完情報の中で別解情報を例にとると、質問内容に対する回答が含まれるページである。質問 Q を表す語を単純に AND 検索した場合、本文中には回答に結びつく情報が現れない可能性が高いが、タイトルに(質問 Q を表す語＋質問型を表す語)を含む Web ページならば、その本文で述べられていることが回答に結びつく情報である可能性が高くなると考えられる。また、その中で“本文中に観点語を含む”という条件を加えると検索結果のスニペットに別解情報が現れると思われる。観点語は先ほど述べたように「回答を述べるにあたって必要な語」であるので、別解を表す語も観点語周辺に現れると考えられるからである。

以上より、提案手法として、タイトルに手順 1, 2 で取得した質問 Q に関する語、本文に回答を述べる際に重要な観点語を含む Web ページを取得できるような構造化クエリ intitle: (q_1 + q_2 + t_1) intext: (a_1+a_2..)を別解情報を取得するクエリとして生成することを提案する。また、補完情報のうちの一つの回答 A_x に関する詳細情報や証拠情報の取得についてもこの構造化クエリを利用して、タイトルに回答を表す語を含み、本文に詳細情報や証拠情報を表す際に必要な観点語を含む Web ページを求めるクエリを生成することが考えられる。

この手順で生成したクエリを用いて各種の補完情報を取得することを提案する。

5. 補完情報提示の流れ

補完情報取得から提示までの流れを説明する。大まかには次のようになる。

1. 補完情報取得のためのクエリ生成
2. 生成したクエリで Web 検索（or QA サイト内検索）
3. 取得した補完情報の提示

本研究では、QA コンテンツに不足した情報を Web から補完情報として取得することを目的としている。そのため、

まずは情報取得のための検索クエリを注目している QA コンテンツから生成する必要がある。

次にそのクエリで Web 検索を実行し補完情報の候補を収集する。この検索は Google 検索[1]や Yahoo!検索[2]のように Web 全体を対象とした検索のほかに、QA サイト内に対象を絞った検索も考えられる。Web 上の情報は多種多様で量も多く、特定のコミュニティに情報が偏らないが、QA コンテンツのように質問と回答とある程度構造化された形式を取らない。一方で QA サイト内の情報は量は限定されるが同じ質問応答コンテンツであるので情報が扱いやすいというメリットがある。Web 全体もしくは QA サイトのみのいずれを情報ソースとして用い、どちらの方が補完情報の収集先として適しているかは今後検討が必要である。

最終ステップでは取得した補完情報をコンテンツ閲覧ユーザに提示する。補完情報の提示の仕方には 2 つのものがあげられる。一つは補完情報が載っている Web ページの URL を提示するというものである。ユーザは補完情報を得るために URL をクリックしなければならないが、詳細なテキスト情報が必要またはテキスト情報以外に画像情報が補完情報として必要といったケースのように、情報量が重要な場合に有効である。二つ目の提示方法は、Web ページから補完情報が記載された部分を抽出したものを要約して提示することが考えられる。これは補完情報を文章単位で扱うものである。このためには、補完情報が載っている Web ページを取得した後に、自然言語処理などを用いて情報を抽出その後情報を要約する必要があるが、閲覧ユーザにとっては URL をクリックすることなく補完情報を簡潔に見られるという利点がある。

以上が本研究の補完情報提示の流れである。

6. まとめ

質問応答コンテンツの補完には種類があり、補完情報取得のためのクエリ生成手法も補完対象のコンテンツによって異なることを考察した。また、intitle:など、既存の検索エンジンの検索オプションを用いた補完情報取得クエリの生成手法を提案した。今後の研究では補完情報取得のうち、「手段・方法型」の質問に対する補完を考察していく。方法を問う質問は知恵袋にも数多く投稿されており、その内容もレシピ、ペット、ダイエットなど多岐にわたる。手段・方法型の質問に対する回答は一つに定まらず、いろいろな回答が考えられる。レシピ投稿サイトやブログの普及から、手段・方法型質問を含むコンテンツの回答集合には含まれない回答情報が Web 上には多く存在する。そのような情報を補完情報として取得する手法や、クエリ中での質問を表す語の自動抽出法、取得した補完情報の提示の仕方などが今後の課題である。

謝辞：

本研究の一部は京都大学グローバル COE プログラム「知識循環社会のための情報学教育研究拠点」, 文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」,および,計画研究「情報爆発に対応するコンテンツ融合と操作環境融合に関する研究」（研究代表者：田中克己、課題番号 1809041）, ならびに,NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」によるものです。ここに記して謝意を表すものとします。

[1] Google 検索: http://www.google.co.jp/

[2] Yahoo!検索: http://search.yahoo.co.jp/

参考文献

[1] Zoltan Gyongyi, Georgia Koutrika, Jan Pedersen, Hector Garcia-Molina, Questioning Yahoo! Answers, Proc. of the 17th international conference on World Wide Web (WWW2008), pp.665-674, 2008.

[2] Lada A. Adamic, Jun Zhang, Eytan Bakshy, Mark S. Ackerman, Knowledge Sharing and Yahoo Answers: Everyone Knows Something, Proc. of the 17th international conference on World Wide Web (WWW2008), 2008.

[3] 佐藤弘樹, 島田諭, 福原知宏, 佐藤哲司, コミュニティの活性度評価に関する一検討, 知識共有コミュニティワークショップ, pp.31-38, 2008.

[4] F. Maxwell Harper, Daniel Moy, Joseph A. Konstan, Facts or Friends? Distinguishing Informational and Conversational Questions in Social Q&A Sites, Proc. of the 27th international conference on Human factors in computing systems (CHI2009), pp.759-768, 2009.

[5] 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-データベースシステム (DBS), 2009.

[6]高田夏希,山本祐輔,小山聡,田中克己,質問応答コンテンツに対する Web による回答補完,DEIM2009

[7]小山聡,田中克己, 質問の階層的構造化を用いたWeb検索手法の提案, DBSJ Letters Vol.1,No.1