質問応答コンテンツに対する Web による回答補完

高田　夏希†　山本　祐輔‡　小山　聡‡　田中　克己‡

†京都大学工学部情報学科 〒606-8501 京都府京都市左京区吉田本町
‡京都大学大学院情報学研究科社会情報学専攻 〒606-8501 京都府京都市左京区吉田本町
E-mail: †takata@dl.kuis.kyoto-u.ac.jp, ‡{yamamoto,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　本研究では質問応答サイトに存在するコンテンツを分析し、そのコンテンツに欠けている情報を Web により補完する手法を提案する。質問応答サイトは、ユーザが質問を投稿しそれを見た別のユーザが自身の知識・考えを回答として提供する仕組みである。こうしてできる QA コンテンツは、質問者と同様の疑問を持った別のユーザにとって有益な情報になりうる。しかし、ある質問 Q に付けられた回答 A のほかにも質問 Q を解決するための別の回答 A'が存在する場合がある。また、回答 A が信用できる情報であるかの判断が難しい場合もある。回答の網羅性を上げるための別の回答や、回答 A の信憑性に関する補完情報を Web から収集し提示することで質問応答サイト利用者の疑問解消を補助することを目標とする。

キーワード　情報検索　コンテンツ補完

1. はじめに

近年、質問応答サイトの利用者が増加している。検索エンジンを用いる場合とは異なり、疑問を自然文で入力できることや自分の疑問に答えるユーザが存在することが利用者増加の要因として挙げられる。これに伴って質問応答サイトに存在する質問回答情報(質問 Q と、そこに付けられた 1 個以上の回答 A_m の組、以下 QA コンテンツ)の量は急激に増えており、それらの QA コンテンツから知りたい情報を得ようとする人もいる。現在 Yahoo!知恵袋[*1] には約 2331 万件の質問と約 6663 万件の回答が存在する。これらの質問・回答はだれでも閲覧することができる。また、知恵袋はユーザ登録を行うと質問や回答を投稿することができる。2009 年 2 月現在の登録ユーザ数は約 305 万人に及ぶ。これは知恵袋のような質問応答サイトの需要が高いことを示している。

ここで、Yahoo!知恵袋の仕組みについて説明する。質問に対しだれかが回答を投稿し、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(以下 BA)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。こうして一つの QA コンテンツが作られるが、質問の回答となりうる適切な知識を持ったユーザが回答しているとは限らず、また十分な回答が投稿されないまま期限が来たり BA が決まってしまう場合もある。さらに BA は、質問されている内容に詳しい専門家が決めるわけではなく、閲覧ユーザたちの投票で決定されたり質問者自身が決めたりする。よって結果として回答の信憑性の薄い QA コンテンツや質問に対する回答が不十分な QA コンテンツが生成されることもある。

本稿では上記のようにして生成された QA コンテンツに対し、Web 情報を用いてコンテンツを補完することを提案する。コンテンツを補完することで、QA コンテンツの利用者は回答の信憑性を確認したり、補足的な情報を得ることができるようになると考えられる。

QA コンテンツに対する補完情報は大きく 2 種に分けられる。まず質問 Q における回答 A_m の信憑性を判断するための情報を補完することが考えられる。つまり、回答 A_m で言われていることが Web 上ではどのように記述されているのかを調べて回答 A_m と比べるものである。このような補完は、質問に対し正解となる答えが決まっているような場合に必要と思われる。例えば、物の名称や由来、科学や物理など学問的な疑問を尋ねる場合である。2 つ目は回答 A において不足している質問 Q への回答情報を補完することである。質問 Q に対し Web ではどのようなことが言われているのかを調べて回答 A と比べ、回答 A の情報だけでは得られなかった質問 Q への回答となりうる情報を提示することがコンテンツに対する補完となると考える。この、回答 A からは得られない情報の補完は、質問に対し様々な回答が考えうる場合に有効である。レシピやダイエットの方法など、正解を求めるのではなく質問に則したあらゆる情報を求める場合が考えられる。

本稿では情報補完の中でも後者の補完情報を得る手法を提案する。質問 Q から生成したクエリを検索エンジンで検索すると、得られるページ集合の中には質問 Q に関する何らかの情報を含むページがあり、さらにその情報はクエリの周辺テキストに現れると考えられる。よってその周辺テキストを形態素解析し、得られる語集合の中には質問 Q に対する回答を象徴するような語が含まれる。また、そのような語は得られたページ集合中の複数のページに現れると考えることができる。このことを利用して、周辺テキストから得た語集合の中から回答を表す語を抽出する。回答 A に含まれない語が回答を表す語として抽出されれば、その語

*1: http://chiebukuro.yahoo.co.jp/

は回答Aとは異なった質問Qへの回答を表す語となる。この抽出した語を利用して質問Qに対するWeb上での回答を求める。

以下、本稿では補完情報を質問Qに対する回答Aからは得られない情報とし、そのような情報が得られるWebページを取得することを目的とする。また、Webには補完を行うQAコンテンツの存在する場でもある知恵袋そのものも含まれる。これに着目すると、回答Aから得られない情報は知恵袋内の他のQAコンテンツからも得られる可能性がある。なぜなら、知恵袋には質問内容の類似したQAコンテンツが複数存在するが、それぞれに付けられた回答は内容が異なる場合があるからだ。そこで、提案手法をWeb全体と、Webの中の知恵袋のみに対象を絞った場合の2種類にあてはめて補完情報を得ることを考える。得られた補完情報のWebページ(または他のQAコンテンツ)のURLを補完対象となるQAコンテンツ閲覧中のユーザに対し提示することでユーザの質問Qに関して得られる情報を増やす。続く2章では関連研究と参考文献を述べ、3章でYahoo!知恵袋に存在するQAコンテンツについての分析を行う。4章で補完情報をWebまたは他のQAコンテンツから取得する手法について説明し、5章で実験・考察を、6章でまとめを述べる。

2. 関連研究、参考文献

質問応答サイトの回答を対象にした研究には[1]や[2]があげられる。研究[1]はYahoo!知恵袋を対象にした研究で、質問に付けられた複数の回答から最も質問に適した回答(BA)を判定する方法を述べている。これは知恵袋の質問回答情報をクラスタリングし、クラスタごとに機械学習を行ってBAとなりうる回答を選ぶというものである。研究[2]は[1]とは逆に機械学習によって不適切な回答を分類することを目的とした研究である。どちらの研究も質問に対する回答の適合性を評価するものであり、本研究の回答にかけた情報を補完しようとする目的とは異なっている。

河重らの研究[3],[4]はともに情報補完に関する研究である。河重らは検索語とWeb検索で得られた文書における検索語の周辺テキストを用いてユーザの求める情報を得られるよう検索語を追加していくことを提案している。研究[3],[4]は、ユーザがWord文書などを閲覧中にその閲覧文書内の語に関する情報を求める場合を想定しているが、検索語追加の過程が本研究の回答を表す語の取得方法と似ているため参考にした。しかし、河重らの研究[3],[4]では、検索語の追加候補がユーザの閲覧文書内の語である点が異なっている。本研究では、取得する回答を表す語はQから生成したクエリでの検索結果の文書内から取得する。

本研究は質問応答サイトという、いわゆるWeb2.0コンテンツに対する情報の補完を目的としている。同じくWeb2.0コンテンツを対象として情報補完を行う研究は[5]や[6]がある。灘本らの研究[5],[6]は、コミュニティ型コンテンツであるSNSやBlogにおいてユーザの視点から外れてしまった、ユーザが気付いていないような情報を抽出しユーザに提示することを目的としている。この研究では、もともとコンテンツ内に存在するがユーザに気付かれていない情報を探し出し提示するもので、コンテンツからは得られない情報をWebから探し出し提示する本研究とはその点が異なっている。

3. Yahoo!知恵袋の分析

3.1. Yahoo!知恵袋の統計情報

1章でも述べているが、あらためて2009年2月現在の知恵袋に存在する質問数･回答数を表1にまとめる。また、表1には2008年9月の質問数･回答数も比較のために載せている。

	質問数	回答数
2009年2月	約2331万件	約6663万件
2008年9月	約1893万件	約5627万件

表 1：Yahoo!知恵袋内の質問数・回答数

表1より、約4ヶ月で質問数は約300万件・回答数は約750万件増加していることが分かる。このことから、知恵袋内のQAコンテンツが月を経るごとに急激に増加していることがうかがえる。

しかし、質問数が増えた分だけ質問の種類が増えているとは限らず過去に同様の質問があるにもかかわらず同じ内容を質問する人もいるため、知恵袋内の質問内容は重複が少なからず見受けられる。例えば、「英単語の覚え方」を問う内容の質問を想定して知恵袋のすべての質問文を対象に「英単語 覚え方」というクエリでAND検索すると、409件が該当した。だが、質問内容が同様であっても、質問に回答を投稿するユーザは質問毎に異なっており同様の質問でも回答の内容が異なる場合がある。上記の例で言うと、質問内容が同じ2つのQAコンテンツで、一方の回答者は「書いて覚えるのがいい」というのに対しもう一方の回答者は「声に出して覚えると覚えやすい」というように、同じ質問でもコンテンツによって回答が異なる場合である。このように、質問内容が重複することは必ずしも悪いこととは限らず、1章でも述べたように一つのQAコンテンツだけでは得られなかった知識を他のQAコンテンツから得られる場合もある。

また、株式会社インタースコープが2005年に全国

の 15～59 歳の男女計 4151 人を対象に行った情報メディアに関する調査[7]によると、質問応答サイトの信頼度について「非常に信用できる」、「おおむね信用できる」を選択した人が全体の 48.5%であった。次に質問応答サイトの利用度であるが、調査対象者の 63.7%が「非常によく利用する」、「良く利用する」または「時々利用する」と答えている。以上から、約半数の人が質問応答サイトを信頼していることや過半数の人が質問応答サイトを利用していることがわかった。だが、サイト内の QA コンテンツの情報は信憑性が薄かったり、情報が欠けている可能性がある。よって、何かを検索していて検索結果として質問応答サイトのコンテンツが現れた場合や自らが質問者となり何か知りたいことを質問応答サイトに投稿した場合など、QA コンテンツを利用する状況は様々であるが、そこから得られる情報を 100%信用してしまうのは危険である。にもかかわらず半数近くの人が質問応答サイトを信用しているという調査結果が出ているため、質問応答サイトに存在する QA コンテンツの情報を補完し、コンテンツを見るだけでは得られない情報や本論文では触れていないがコンテンツの信憑性を判断するための情報を与えることは有意なことであるといえる。

3.2. 質問の分類と、補完の対象とする質問の種類

質問の分類方法についての研究には[8]がある。研究[8]は質問がどのようなことを尋ねているかで Description Questions や Method Questions など大きく 5 つに分類している。このような研究からわかるように、質問には様々な種類がある。

1 章でもふれたように QA コンテンツの質問 Q が求めている回答によって補完できる情報に差が出ると考えられる。QA サイトの質問を「意見を求める質問」や「方法を問う質問」など、どんな情報を求めるかで分類した研究には[9]がある。Marciniak の研究[9]では Yahoo!Answers(海外における Yahoo!知恵袋)の質問を YES/NO,FACTOID,INFORMATIONSEEKING,PROCED-URAL,OPINIONSEE-KING,SURVEY/POLL という風に分類している。筆者は質問 Q が求める情報(回答)の種類によって、質問をさらに大まかに 2 種類に分けることができると考える。一つ目は回答者の意見、主観を求める質問である。例えば「大根を使った料理を教えて」や「太ももを細くするには？」、「大阪でお勧めのたこ焼き屋はどこですか？」など、回答者ごとに色々な情報が得られる場合である。二つ目は回答者の知識を求める質問である。つまり、質問に対し何か正解の情報が存在するような状況であり、例をあげるとものの名称や由来、自然現象の疑問、2 つ以上のものの違い(ex.プラズマテレビと液晶テレビ)などである。この場合、回答者は自らの意見ではなく知っていることを回答として投稿する。

補完情報は 1 章でも述べているが「質問 **Q** における回答 A_m の信憑性を判断するための情報」と「回答集合 **A** には現れないが質問 **Q** への回答となりうる情報」の 2 種類に分けられる。本研究では後者の、質問 Q へのいわゆる別解を取得することを目的としている。ここで、質問が二つ目に述べた種類のものである場合、そこに付けられた回答によって別解が存在するかしないかが左右される。つまり、ある QA コンテンツの回答集合 A が質問 Q への正解を網羅していれば Web 上には回答集合と同様な情報しか存在しないし、網羅されていなければ Web 上にはその回答に現れない情報が存在する。回答で間違った情報が与えられている場合は Web 上には回答と異なる情報が存在するが、これは前者の回答の信憑性に関する補完情報であるといえる。質問が回答者の意見を求めるものである場合は、Web 上には様々な別解が存在すると考えられる。質問が求める情報に正解が決められていないため、Web 上にある質問 Q に対する様々な情報が別解情報になりうる。そこで本論文では提案手法で Web による補完を行う際に対象とする QA コンテンツを、質問 Q が回答者の意見を求めているコンテンツに限定する。

3.3. 回答文に含まれる URL

本研究では、補完情報を得られる Web ページの URL を示すことによって知恵袋を閲覧しているユーザを支援する方法をとっている。そこで、現在の Yahoo!知恵袋において回答文中に Web ページの URL が示されているものの数やその URL のドメインの種類などを調べた。ランダムに 5000 件の QA を取得し、その回答に含まれていた URL のドメインの種類数・回答に現れた URL の総数・回答文中に URL を用いた QA の数を調べた結果を表 2 に示す。

ドメインの種類	１２０３種
出現 URL の総数	２０４０個
URL を用いた QA の数	１１９３件

表 2：5000 件の QA に含まれる参照 URL

表 2 において一番用いられた URL は Wikipedia[*3] で 98 回現れており、次いで YouTube[*4] で 86 回であった。表 2 において出現 URL の総数と URL を用いた QA の数が一致しないのは、一つの QA 中に複数個の URL が表れる場合もあるためである。また表 2 より、5000 件の QA 中約 25%が回答に URL を含めていることがわかる。よって本研究にて得られた補完情報の提示の仕方として URL を用いるということは知恵袋のユーザにとって違和感のあることではなく、また逆に言うと約 75%の QA には参考となる URL の情報がないとも言えるため、そういった QA に対して補完情報の URL を

提示することはより QA コンテンツの内容を充実させることにつながると考えられる。

次に、質問が分類されている各カテゴリで参照 URL に差が出るのかを調べた。2004 年 4 月から 2005 年 10 月に知恵袋に投稿された QA コンテンツ約 311 万個のデータを解析し、「エンターテインメントと趣味＞音楽」,「暮らしと生活ガイド＞料理、グルメ、レシピ＞レシピ、調理法」(大カテゴリ＞中カテゴリ＞小カテゴリ) において出現回数の多い参照 URL のサイト名・出現数を出現回数順に 3 つ示したものが表 3,4 である。

表 3,4 から、カテゴリによってよく参照されるサイトに大きく差が出ることがわかる。また、表 2 は最近の QA コンテンツも含めてデータを収集したため YouTube という動画サイトが多く参照されているという結果が出ているが表 3 は 2004 年から 2005 年 10 月までのデータであるため音楽カテゴリでよく参照されている URL には YouTube は現れなかった。このように、QA コンテンツが生成された時期によっても参照 URL に変化が出ることがうかがえる。

サイト名	出現回数
うたまっぷ.com	726
Amazon	266
Yahoo!MUSIC	257

表 3:音楽カテゴリの参照 URL

サイト名	出現回数
Yahoo!グルメ	496
Cookpad	424
AllAbout	340

表 4:レシピ・調理法カテゴリの参照 URL

4. QA コンテンツに対する補完情報取得

本研究は Web 情報を用いて QA コンテンツを補完することを目的としている。今回補完を行う対象となる QA コンテンツについて、3.2.でも述べたとおりそのコンテンツの質問 Q が回答者の意見を問うものに限定している。また、QA コンテンツの補完に用いる Web 情報として「質問 Q における回答 A_m の信憑性を判断するための情報」と「回答集合 A には現れないが質問 Q への回答となりうる情報」の 2 種類を補完情報と呼び、本論文では後者の情報を取得するための手法を提案する。

4.1. Web からの補完情報取得方法

提案手法の処理の流れを以下に示す。

(1) 質問 Q からクエリ $\boldsymbol{q}=\{q_1,q_2,\ldots\}$を生成

(2) 生成したクエリで Web 検索を行いページ集合 $\boldsymbol{P}=\{p_j\}$(j=1,…,300)を取得

(3) 取得したページ集合 $\boldsymbol{P}$ の各ページ p_j のスニペット t_j からクエリ以外の語を収集

(4) 収集したすべての語から重複を省き、さらに回答集合 A に含まれる語を除いた一つの語集合 $N=\{n_1,n_2,\ldots\}$を得る

(5) (4)で得られた語集合 N の各要素 n_i について、(2)で得たページ集合 $\boldsymbol{P}$ において n_i を含むページの数(取得したページ集合における DF 値)を調べる

(6) (4)で得られた語集合 $\boldsymbol{N}$ の各要素 n_i について、(1)で生成したクエリ $\boldsymbol{q}$ との共起度 $\boldsymbol{Co}$ を調べる

(7) (5)で得た各語の DF 値を(6)で得た共起度 $\boldsymbol{Co}$ を用いて重み付けし、重み付け後の値の高い順に n_i を 30 個取得する(語集合 $\boldsymbol{N'}$とする)

(8) (2)で取得した各ページ p_j のスニペット t_j が語集合 $\boldsymbol{N'}$中の語をいくつ含むか調べる

(9) (8)の結果、語集合 $\boldsymbol{N'}$中の語を多く含んだページ p_j を補完情報を示すページとして提示する

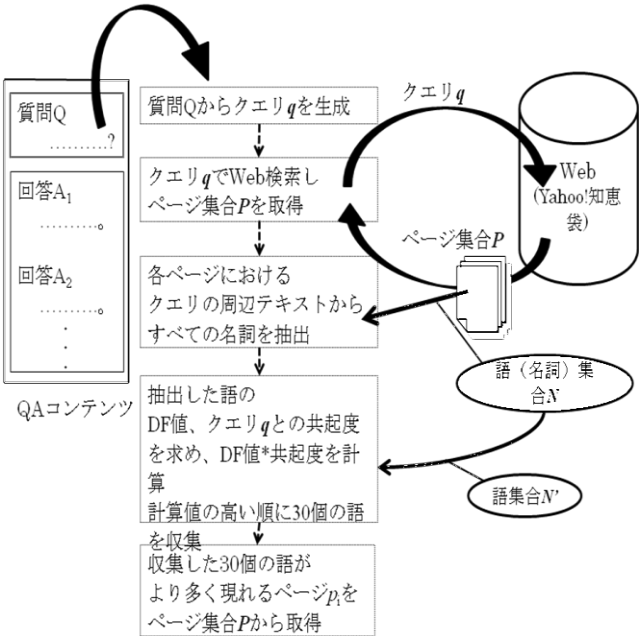

図 1:提案手法の流れ

4.2. 知恵袋からの補完情報取得

知恵袋内には質問の内容が類似した QA コンテンツが存在する。しかし、質問の内容が同じコンテンツでも、そこに付けられた回答集合から得られる情報は異なる場合がある。そこでこの節では、Web の中でもさらに Yahoo!知恵袋の中から補完情報を得る手法を説明する。大まかな流れは前節と同様である。

(1) 質問 Q からクエリを生成して Web 検索を行う。この際に検索対象のドメインを Yahoo!知恵袋に絞る。得られた検索結果をページ集合 $\boldsymbol{P}$ とする

(2) ページ集合 $\boldsymbol{P}$ の回答文から名詞を収集し、そこから前節の手法と同様に語の重複や補完対象となる QA コンテンツの回答集合 A に含まれる語を除いた語集合を得る。

*3: http://ja.wikipedia.org/wiki/
*4: http://jp.youtube.com/

(3) 得られた語集合の各語について、ページ集合 $\boldsymbol{P}$ において語を回答文に含むページの数($\boldsymbol{P}$ における DF 値) を調べる

(4) 以下、共起度の計算からページ取得までは前節と同じである。

以上の、Web または知恵袋の他の QA コンテンツからの補完情報取得の流れを図 1 に示す。

4.3. 質問文からのクエリ生成

4.3.1. クエリ生成に関する制限

まず質問文 Q からクエリ $\boldsymbol{q}$ を生成する。ここで、クエリ生成時に回答の情報を用いないのは、本研究が求める補完情報が回答で言われていることとは異なる情報、つまり質問に対する別解としているからである。クエリはプログラムが自動で生成する場合と人間の手入力によるものの 2 種類が考えられる。現段階ではクエリ $\boldsymbol{q}$ は手入力であり、質問の内容を表していると考えられるいくつかの語をクエリ $\boldsymbol{q}$ としている。また、知恵袋に存在する質問文の中には 1 つの質問文中に聞きたいことを複数個含めているものもあるが、今回は QA コンテンツ閲覧中のユーザが、一つの質問内容に対して補完を求めるものとする。

クエリ生成が手入力であることの理由は、クエリが質問を表すものでなければそのクエリでの検索結果に質問に関連するページが含まれなくなるためである。

補完を行う状況によってクエリの生成方法に違いが出ると考えられる。本研究は QA コンテンツを閲覧中のユーザに対し、そのコンテンツの内容を Web 情報で補完するというものである。ユーザがそのコンテンツを閲覧するに至った経緯が、検索エンジンか知恵袋の検索機能で何かを調べていてその QA コンテンツの URL をクリックしたという状況ならば、質問 Q の内容を表すクエリは{そのユーザが検索エンジンまたは知恵袋の検索機能に与えたクエリ ― 回答文に含まれる語}であると言える。なぜなら、ユーザが何かを知りたいと思い作ったクエリでの検索結果にその QA コンテンツが含まれ、さらにそこに自分の知りたい情報があるかもしれないと思いそのコンテンツを閲覧するのだから、そのユーザが作ったクエリはコンテンツの内容を象徴していると考えられる。この場合、質問からのクエリ生成は人間の(コンテンツ閲覧中のユーザの)手入力に値し、今回の実験でのクエリの手入力に近い状況である。

ユーザが偶然そのコンテンツを発見した場合や、どこかからリンクを辿ってそのコンテンツを見るに至った場合は、その QA コンテンツを補完する際の質問 Q からのクエリ生成はユーザにクエリを入力してもらうという方法もあるが、やはりプログラムが自動で生成するのが望ましいといえる。以下に質問 Q からクエリを自動生成する方法を述べる。

4.3.2. 質問からのクエリの自動生成

クエリの自動生成は以下の手順で行う。なお、この節では様々な質問内容からのクエリ生成を考えるために QA コンテンツの質問の種類をこの章の最初に書いたようには限定しない。

- 質問文に「n_x と n_y の違い」、「n_x と n_y って(*)違」、「n_x と n_y は(*) 違」という表現があれば $\boldsymbol{q}$={n_x,n_y, 違い}とする。(*) はどんな語句が来ても、また何も語句がなくてもマッチするようにしている。これは「ってどう違う？」、「は何が違いますか？」、「って違うの？」というような表現に対応するためである。また、文末を「違」としているのは「違いますか？」や「違うの？」という表現の違いに対応させるためである。このクエリは 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。知恵袋の質問文に”の違い”というフレーズを含むものは約 1 2 万個存在する。
- 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 $\boldsymbol{q}$ = {n_x,n_y とは}
- 上記の、「n_x の」という部分がない表現ならば $\boldsymbol{q}$ = {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。「とは何」や「って何」というフレーズを含む What 型の質問数は約 1 7 万件である。
- それ以外の質問表現ならば質問文を形態素解析し、tf-idf 法で値の高い語を最大 3 語クエリとする。(tf = 形態素(名詞、形容詞、動詞)が質問文とすべての回答文中で何回現れるか、df = いくつの Yahoo!知恵袋内の QA コンテンツがその形態素を含むか)
- 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場合は上記の tf-idf 法で得たクエリに{理由}という語を追加する。これは Why 型質問のクエリ生成方法であり、「何故」や「どうして」を含む質問は約 113 万件ある。

何かの違いを説明するページや、物事の理由を述べるページ内に「違い」や「理由」という語が必ずしも現れるとは限らないが、クエリに{違い}や{理由}を追加することで、取得するページ集合 $\boldsymbol{P}$ での適合率を上げることができると思われる。

しかし、質問からのクエリ自動生成は現段階ではあまりうまくいくとは言えない。例えば質問文が「睡眠不足は体にどんな影響がありますか？」や「バイクのカウルって何ですか？」など、比較的簡単な文章では先ほど説明したアルゴリズムでクエリを生成するとそ

れぞれ{睡眠 不足 影響}と{カウルとは バイク}などとなり、直観的にだが質問文を表すクエリが作れている。だが、「京都で有名な温泉はどこですか？和歌山でゆう城崎温泉みたいな…定番を 2.3 こ教えて頂けませんか？」という質問文でクエリを生成すると{温泉 有名 和歌山}となった。この例でクエリを作るなら{温泉 有名 京都}などがふさわしいと思われる。和歌山がクエリに含められた理由として、質問者が兵庫県の城崎温泉を和歌山であるといったため回答者からの指摘が加わって和歌山という語の tf 値が上がってしまい tf-idf 法で高く評価されたためと考えられる。ほかに形態素解析の誤りなどでうまくいかない例も多々あった。質問者が聞きたいことの本質以外の付加情報を質問文に含めることで質問文が長くなっている場合などにクエリが作りにくくなる傾向があった。また、極端な例えだが、言語パターンによる生成手法も取り入れているため、What 型を想定した「～の…とは何」というフレーズは「あのひとは何で」など What 型でない質問にも無条件で当てはまってしまうのも問題である。

4.4. 質問 Q に対する回答を象徴する語

質問文から作成したクエリ $\boldsymbol{q}$ を Web 検索エンジンで AND 検索する。この検索結果で得られるページ集合 $\boldsymbol{P}$ には質問文で聞かれていることに関する情報のページが含まれると思われる。(なお、今回は取得するページ数は 300 としている)質問の回答となりうる情報を象徴する語（今回は名詞のみを対象とした）は、ページ p_j においてクエリ $\boldsymbol{q}$ の周辺テキストに含まれると仮定する。そこで、取得した上位 300 件の各ページ p_j のスニペット t_j をクエリ $\boldsymbol{q}$ の周辺テキストとして抽出した。この t_j を形態素解析し名詞のみ抜き出す。これを t_1 から t_{300} まで繰り返し、得られた名詞集合から重複を除いた語集合 $\boldsymbol{N}=\{n_1,n_2,\ldots\}$ を質問に対する Web での回答を表す語の候補集合とする。

4.4.1. 語の DF 値

もし n_i が質問の解を表すような語ならばページ集合 $\boldsymbol{P}$ の中で n_i が表れる p_j の数は多くなると考えられる

よって得られた $\boldsymbol{N}$ に含まれるすべての n_k について、

$$m_j(n_k)=\begin{cases}1 & n_k \text{が} P_j \text{に現れる}\\ 0 & n_k \text{が} P_j \text{に現れない}\end{cases}$$

を調べ、語 n_i のページ集合 $\boldsymbol{P}$ における DF 値 df_i を

$$df_i=\sum_{j=1}^{100} m_j(n_i)$$

で求める。上記の考えより、n_i が Web での(質問に対する)解を表す語であるならば df_i の値は高くなると思われる。

4.4.2. DF 値の重み付け

DF 値が高い語には、一般的な語も多く含まれる。たとえば「サイト」という語や「関係」という語、または形態素解析にて名詞と判断された「心」や「番」などの漢字 1 文字の語などは状況を限らず良く使われる語であるので必然的に上記の式で得られる DF 値も高くなる。しかし、その一般的な語が質問への回答情報を表す語であるとは限らないため、得られた DF 値を重み付けする必要があると考えられる。そこで、それぞれの語と生成したクエリとの共起度 $\boldsymbol{Co}$ を求め、それらを併用した値を

$df'_i = df_i * \boldsymbol{Co}_i$

と定義し、df'_i の高い語を質問 Q への回答を表す語とする。

次に語とクエリの共起について説明する。語 n_i が質問の回答に関する語ならば、質問から生成されたクエリ $\boldsymbol{q}$ との共起度合いは高くなると考えられる。ここでいう語とクエリの共起度とは、それらが同時に出現する割合のことであり、語とクエリの関連性の強さを示す指標である。本論文ではこの共起度を Jaccard 係数を用いて求める。Jaccard 係数を用いた共起度は以下のように表わされる。

$$Co_i=\frac{|q\cap n_i|}{|q|+|n_i|-|q\cap n_i|}$$

ここで、$|\boldsymbol{q}|$ は検索エンジンでクエリ $\boldsymbol{q}$ に含まれる語をすべて AND 検索して得られる検索結果数、$|n_i|$ は語 n_i の検索結果数、$|\boldsymbol{q}\cap n_i|$ はクエリ $\boldsymbol{q}$ に含まれる語と語 $\boldsymbol{n}_i$ を AND 検索した検索結果数を表す。$\boldsymbol{Co}_i$ はクエリ $\boldsymbol{q}$ が現れるページ集合と語 n_i が現れるページ集合の重なりの度合いを示すものである。

5. 実験・考察

QA コンテンツは大きく分けて、その質問内容が回答者の知識を求めるものと回答者の意見を求めるものの 2 種類がある。3.2.の最後にも述べたが、今回の実験は後者の回答者の意見を求める質問内容の QA コンテンツを対象に Web 情報を用いたコンテンツへの情報補完を行う。また、QA コンテンツを Yahoo!知恵袋内の他の QA コンテンツを用いて情報補完した結果も示す。実験に用いた QA コンテンツ数は 25 個で、何かの作り方や何かをする方法、お勧めのものなど回答者毎に色々回答があり得る質問の内容を「レシピ・もの・場所・体・方法」の 5 分野でそれぞれ 5 つずつ著者が適当に選んだものである。

5.1. 評価実験

4 章で述べた方法で、25 個の QA コンテンツに対する Web 上または知恵袋内の補完情報と思われるペー

ジを取得した。質問 Q に対する回答を象徴する語集合 $\boldsymbol{N'}$(DF 値*クエリとの共起度の値の高い順に取得した 30 個の語)の要素をより多く(スニペットに)含むページを 10 個集め(語の含まれる個数が同じ場合ははじめにクエリで検索した時の検索結果順)、そのページの内容が、補完対象の QA コンテンツの質問文 Q に対する回答となる情報が記載されていて、さらにその情報は QA コンテンツの回答集合 A からは得られない、新しい回答(別解)情報であったかどうか<1>・その中で何種類の補完情報が得られたのか(内容の重複するものを除いた数)<2>・実際に役立つ情報か<3>を人手で評価した。結果を以下に示す。なお、<3>の評価は、取得した情報が実際に活用できるかを判断し、あまり現実的でない情報(ダイエットなど体関連の質問でエステの情報など)や取得したページそのものから情報が得にくい場合(レシピの質問で、回答に関連するリンク集のページ)を省いた情報の数となっている。

5.2. 考察

Web からの補完情報としては平均 3 種類の補完情報が取得でき、実際役立つ情報は平均 2 個であった。質問の種類別に見ていくと、体関係の質問はあまり有益な情報が得られなかった。質問は、ウエストを細くする方法(23)やクマの取り方(24)などを用いたが、エステや整形外科の広告などが混じり、あまり現実的でない解決法が多かった。レシピ関係の質問に関しては、全て 1 つ以上の補完情報が取得できた。もの関係でも、たとえば質問 6 はぬる燗に合う辛口の日本酒が知りたいという質問であり、補完情報として色々な種類のぬる燗に適した辛口の日本酒を見ることができた。また、単純に検索サイトで{ぬる燗 辛口 日本酒}で検索した時の検索順位と比較すると、取得した 8 個のページ中 3 つは 200 位以上のページであった。このことから、質問に対する答えを表す語(語集合 $\boldsymbol{N'}$)を含むか否か調べることで、クエリを検索エンジンに与えるだけでは得にくかった質問に対する回答情報を得やすくなったと言える。

提案手法は、回答を表すような語として集めた語集合 $\boldsymbol{N'}$が重要なポイントとなる。語集合 $\boldsymbol{N'}$が質問 Q と全く関係のない語の集まりとなった場合、補完情報がとれなくなる。たとえば、目の下のクマの取り方の場合だが、$\boldsymbol{N'}$の要素を見ると「熊」や「葉」という、動物のクマに関する語が含まれていた。そのため、結果の中に熊に関するページが含まれてしまったようだ。本手法では語集合は DF 値とクエリとの共起度を掛けて出した値をもとに算出している。一般的な語の DF 値が高くなることに対して共起度の重み付けを取り入れたが、その場合 DF 値が高くない語が共起度によって重要な語とみなされる場合もあるため(熊という語は DF 値が低いが共起度ははじめに取得した語集合の中で 14 番目に高かった)、DF 値と共起度の両方の特性を損なうことのない値の重み付けの方法を考えなければならない。

さらに、結果が質問から生成したクエリに依存することについても改良が必要となる場合がある。クエリは質問文から生成するため、質問者の独特な表現によってあまり適切でないクエリになることがある。たとえば、大根を用いたオカズを教えてくださいという質問文(3)で、クエリにオカズという語を含めたが、それをひらがなのおかずに変えると約 308,000 件も結果数が増加した。結果数が増えることが必ずしも補完情報取得にいい影響を与えるわけではないだろうが、最初に生成したクエリでより質問 Q に関する情報を取得するため(再現率を向上させるため) にも、クエリについても考察が必要である。

他の QA コンテンツからの補完情報としては、平均 3 種、有益な情報 2 個が取得できた。QA コンテンツには店の広告などが含まれないが、同じ回答者が別の質問に対し全く同じ回答文を張り付けている場合がしばしば見受けられた。また、質問 12 はセキセイインコにおしゃべりを教える方法なのだが、質問内容が具体的だと別解を含む可能性のある QA コンテンツを多く取得できない場合があった。

語集合に関して、Web からの情報補完でも言えることだが例えば猫のしつけ(質問 11)に関して取得した語集合に犬という語が含まれたり、博多の観光スポット(質問 19)に関しての語集合に長崎という語が含まれる場合が多く見受けられた。猫と犬という動物同士や博多と長崎という地名同士が同一のページに現れることが多いため共起度が高くなり語集合に含まれたものと考えられるが、そのような同じ種類の別の語が含まれることで猫のしつけや博多の観光スポットを知りたいという状況に対し、犬のしつけや長崎の観光スポットについてのページが多く得られてしまうこともあった。このような、クエリと分類が同じである別の語を判別して除く必要があるかもしれない。

6. まとめ

本稿では知恵袋の QA コンテンツに対する Web 情報を用いた補完の手法を提案した。

今後は本研究の目的である、QA コンテンツの A とは異なる Q への回答情報の取得の精度を上げていかなければならない。今回は提案手法に従って重み付けされた n_k のうち、重みの大きいものが回答を象徴する語と位置付けている。しかし重みの付け方に関して 5.1. で述べたような問題点があるため、重みの大きい語が回答を表す語であるとはまだ言えない。よって重み付

けの方法の改良をしなければならない。また、提案手法で取得した URL が補完情報であるかどうかの評価方法も考案しなければならない。

		Web			QA コンテンツ		
	質問	<1>	<2>	<3>	<1>	<2>	<3>
レシピ	1	2	2	2	5	4	4
	2	8	8	6	8	5	4
	3	2	2	1	4	4	2
	4	1	1	1	1	1	0
	5	3	3	2	8	8	4
もの	6	8	8	8	5	5	3
	7	5	5	2	0	0	0
	8	0	0	0	3	3	2
	9	3	3	3	2	2	2
	10	1	1	1	0	0	0
方法	11	3	2	0	2	2	2
	12*	4	4	1	6	5	5
	13	9	7	3	7	3	3
	14	3	3	2	2	2	1
	15	1	1	1	1	1	0
場所	16	5	4	3	7	5	5
	17	4	4	2	1	1	0
	18	5	5	2	1	1	0
	19	4	4	3	2	2	1
	20	5	4	0	6	3	2
体	21	1	1	0	1	1	1
	22	5	5	0	8	4	3
	23	6	6	1	4	3	2
	24	2	2	1	3	3	2
	25	2	2	0	8	2	2

表 5:提案手法で取得した 10 件の URL の評価
<1>取得した補完情報の数、<2>重複を除いた補完情報の種類数、<3>有効な補完情報の数
***12 の QA コンテンツに関してのみ 8 件中の数**

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。

文　献

[1] 西原陽子,松村真宏,谷内田正彦:QA サイトにおける質問に適した回答の判定, 言語処理学会 NLP 若手の会第 2 回シンポジウム, Sep.2007.

[2] 小林大祐,松村真宏,木戸冬子,石塚満:知識検索サイトにおける不適切な投稿の分類, 第 21 回人工知能学会全国大会, Jun.2007.

[3] 河重貴洋,大島裕明,小山聡,田中克己: 検索語の閲覧文書と検索結果における文脈を利用した質問修正, 第 16 回データ工学ワークショップ, 2005.

[4] 河重貴洋,大島裕明,小山聡,田中克己: 質問修正と再ランキングを用いた文脈依存 Web 検索, 第 17 回データ工学ワークショップ, 2006.

[5] 灘本明代,阿辺川武,荒牧英治,村上陽平: コミュニティ型コンテンツのコンテンツホール抽出手法の提案, 日本データベース学会 Letters Vol.6,No.2.

[6] 荒牧英治,灘本明代,阿辺川武,村上陽平: コンテンツホール検索のためのコミュニティ型コンテンツの対話解析, 第 19 回データ工学ワークショップ, 2008.

[7] 株式会社インタースコープ:情報メディアに関する調査,2005

[8] Rodney D. Nielsen,Jason Buckingham,Gary Knoll,Ben Marsh,Leysia Palen: A Taxonomy of Questions for Question Generation,Workshop on the Question Generation Shared Task and Evaluation Challenge,2008.

[9] Tomasz Marciniak:Language Generation in the Context of Yahoo! Answers,Workshop on the Question Generation Shared Task and Evaluation Challenge,2008