WebとQAコンテンツの相互補完

高田　夏希†　　大島　裕明†　　田中　克己†

† 京都大学大学院情報学研究科　　〒 606–8501 京都府京都市左京区吉田本町

E-mail: †{takata,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　本研究では，解が複数考えうる質問に対し，与えられたいくつかの解情報に対する補完情報を含む Web ページや QA コンテンツを取得する手法を提案する．本研究は，ユーザが，自身の質問意図を表すキーワードクエリを入力する事を想定している．さらに，その質問意図に対する解情報を含む 1 つ以上の Web ページまたは QA コンテンツを入力する事を想定している．まず，入力されたクエリで収集した Web ページや QA コンテンツが解情報を含むかを判定する．次に，解情報を含むと判定された Web ページや QA コンテンツについて，解情報のみを表すベクトルを生成し，ベクトル間の類似度に基づいたクラスタリングを行う．

キーワード　QA コンテンツ，補完情報

1. はじめに

近年，Yahoo!知恵袋(注1)のようなコミュニティQA サイトが普及してきている．これまでは Web ページを閲覧することで Web から情報を取得するのが一般的であった．それが今日では，Web ページに加えて QA コンテンツ（QA サイトに投稿された，質問と 1 つ以上の回答の組）を閲覧する事でも情報を得ることができるようになった．QA サイトでは質問と回答をユーザ同士でやり取りする事ができる．それに加えて過去に投稿された質問と回答が蓄積されており，それらを検索・閲覧する事ができる．つまり，自分が疑問に思ったことが過去に他のユーザによって質問されていれば，その質問に対して投稿された回答情報を閲覧することで情報を得ることができるということである．Yahoo!知恵袋には，サービス開始から 2011 年 1 月現在までに投稿された質問約 5,300 万件およびそれに対する回答が蓄積されている．最近では，Web 検索エンジンにクエリを入力して得られる検索結果に Web ページだけでなく QA コンテンツが含まれることも多くなってきている．

ここで，例えば「二日酔いの解消法」についての情報を Web から取得する事を考える．まず，“ 二日酔い　解消法 ”といったキーワードクエリを Web 検索エンジンに与えることで検索結果を得られる．その検索結果を利用すれば,「二日酔いの解消法」について書かれた Web ページや QA コンテンツから「水を飲む」や「柿を食べる」などの様々な情報を得ることができる．このように，何か一つの検索質問に対しいくつか解情報が得られた状況を想定する．このとき，既に得られた解情報にはいくつかの情報が欠けていると思われる．まず，得られた解情報をより詳細に説明する情報や，得られた解情報の正しさを証明する情報が欠けていることが考えられる．例えば上記の例で言うと，QA コンテンツから得られる情報は「コーラを飲むとすっきりします」といった，QA サイトユーザの実体験に基づく解情報が得られやすい．しかし，その情報の根拠となる情報は示されていないことが多い．また，得られた解情報とは異なるが質問意図に対する解情報となる情報が他に存在する場合も考えられる．Web ページからは「シジミは肝臓の働きを活発にするため，二日酔いに効きます」といった，検索意図に対する詳細な解情報が得られやすい．しかし，Web ページからは QA コンテンツで得られるような実体験に基づく解情報は得にくいと思われる．

そこで我々は，一つの質問意図とそれに対する解情報が Web ページや QA コンテンツからいくつか得られている状況において，その解情報に不足する情報を Web ページおよび QA コンテンツを用いて補完する事を提案する．ここでは，解情報に対し補完する情報を「補完情報」とよぶ．さらに，解情報をより詳細に説明する情報および解情報を裏付ける情報を「追加情報」，解情報とは情報の内容が異なるが質問意図への解情報となる情報を「別解情報」とよび区別するものとする．

1.1　質問意図のタイプと補完情報

ここでは，質問意図のタイプ分類に関する先行研究と，本研究が補完の対象とする質問意図のタイプについて述べる．

QA サイトに投稿される質問の分類に関する研究はいくつか存在する．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 [1]．*Conversational* question とは質問者が回答者と「会話」「議論」をするためになされた質問を指す．*Informational* question は質問者が知識を得るために情報を集めることを目的とした質問を指す．*Conversational* question は質問への解を得るためではなく，回答者の意見を求めるための質問である．そのため，その解情報に対する情報補完の必要性は少ないと考えられる．Metzler らは場所や人物名などの事実に基づく解を求める質問を *fact-based* questions，質問者の問題を解決するような解情報を求める質問を *task-oriented* questions というように分類している [2]．*fact-based* question は質問に対する正しい解を得る目的があると考えられる．よって *fact-based* question に対し得られた解情報については，その証明となるような追加情報の補完が必要になると思われる．一方で「二日酔いの解消

(注1)：http://chiebukuro.yahoo.co.jp/

法」のような *task-oriented* question は，質問に関するより多くの解情報を得ることを目的としていると考えられる．よって別解情報の補完は *task-oriented* question に対する解情報が得られた際に必要となると思われる．

本稿では，*informational* question であり，かつ *task-oriented* question であるような質問意図に対する解情報が Web ページおよび QA コンテンツから得られた際に，その解情報に対する補完情報を Web ページおよび QA コンテンツから取得して補完することを目的とする．

2. 関 連 研 究

本研究は，与えられた入力に対し，入力と関連する内容をもち，かつ入力とは異なる内容の情報を取得するというものである．これは，大島らの研究 [3] の考え方と類似する．彼らは入力された文書集合に対し，同一カテゴリに属するが入力された文書集合とは異なる内容を持つ文書を検索する手法を提案した．我々の研究 [4] では彼らの研究をベースとしていた．また，本研究と基本的に考え方が共通する研究に，Carbonell らの研究 [5] がある．研究 [5] は，検索クエリと，そのクエリで検索された文書集合からいくつか文書を選択したものに対し，検索クエリに類似しており，かつ既に選択された文書と似ていない文書を上位にリランキングするものである．我々の研究は別解情報（ある質問について既に選択された回答以外の回答情報）を取得する事を目的としており，Carbonell らの研究と目的が類似しているが，Carbonell らは文書のリランキングのみで内容ごとに分類するということは行っていない．

QA アーカイブにおける質問検索に関しては言語モデルや確率モデルを用いた Xue らの研究 [6] や Jeon らの研究 [7] が挙げられる．QA 検索には，意味的に類似した質問同士でも用いられる語彙が異なることで検索できないという問題が存在する．そこで質問文に基づく言語モデルと回答文に基づくモデルを組み合わせたモデルに基づいて類似質問を検索することを提案した Xue らの研究 [6] や，内容の類似する質問にそれぞれ与えられた回答同士の類似性を利用して意味的に類似した質問を集め，それを利用して質問文の言い換えを学習することで類似質問を検索することを提案した Jeon らの研究 [7] がある．Wang らも QA サイト内にある質問中から類似した質問を検索するといった研究 [8] を行っているが，この研究は，質問文を構文木に変換したものを利用して質問同士の類似度を測っている．これらの研究 [6]〜[8] は類似する質問や回答を対象アーカイブ中から検索するというものであり，入力に対し，類似する質問を含み，かつ，入力とは異なる回答を含む情報を検索する本研究とは異なる．

QA コンテンツを対象とした最近の研究動向としては，情報の質に着目した研究が数多く行われている．Harper らの研究 [9] では，様々な種類の QA サイトに着目し，高品質な回答にはどのような特徴があるかを推定しており，Google Answers のように，質問者が何らかの謝礼を支払う QA サイトでは，無料で質問可能なサイトよりもよりも高品質な回答が得られることや，Yahoo! Answers のようにコミュニティー自体が大きく，誰でも回答できるようなサイトの方が，特定の個人が回答するサイトよりも高品質な回答が得られることが明らかにされた．Agichtein らの研究 [10] でも，QA コンテンツの品質の推定を行うため，コンテンツそのものの特徴や，ユーザと質問，回答の関係性などを考慮して，高品質な質問や回答と，低品質なものを分類する手法を提案している．Bian らの研究 [11] では，*fact-based* question に対する回答が，質問に対する適合度と，品質をともに兼ね備えている必要があることを考慮したランキング手法を提案している．QA サイトのようなソーシャルメディアは，Web とは異なる構造を持っており，これまでの Web の分析手法をそのまま適用することができない．そこで，彼らは，ユーザや，質問，回答の様々な特徴を用いて，高品質で質問に適切に回答しているコンテンツを高位にランキングする手法を提案した．本研究では，QA コンテンツの品質には直接的には取り組まないが，*task-oriented* question の回答を検索する際には，回答となり得る情報をある程度網羅的に取得することによって，利用者に対して高品質な情報提示を行うことになるため，これらの研究は関連していると考えられる．しかし，本研究の，QA コンテンツや Web ページをもとに補完情報を含む QA コンテンツや Web ページを検索するものとは異なっている．

入力されたテキストに対し，関連するコンテンツを提示する研究としては，近藤らの研究 [12] がある．これは入力されたテキストから重要語を抽出し，重要語を動画検索 API や QA 検索 API に与えてテキストに関連するコンテンツを取得して提示するというものである．与えられた入力に関連する情報を QA 検索で取得するという点で本研究と類似するが，本研究では，入力された情報には含まれない情報を取得する事を目的としている点が，彼らの研究とは異なっている．また，小谷らの研究 [13] では，類似 Web サイト集合が与えられたときに，それらに共通する属性を表す語を HTML 構造や単語の出現頻度を利用して抽出し，それを用いて新しい類似 Web サイトを検索するという手法が提案されている．この研究 [13] は与えられた Web サイトの情報から Web サイトを検索するというものであり，与えられた QA コンテンツおよび Web ページの情報から，QA コンテンツおよび Web ページを検索するという本研究とは異なるものである．

3. 補完情報取得の概要

本節では，*informational* かつ *task-oriented* であるような質問意図を表すキーワードクエリと，質問意図に対する解情報を含む Web ページおよび QA コンテンツがユーザから与えられた際に，それらの解情報に対する補完情報を取得し，提示する手法についての概要を説明する．

まず，提案システムに与えられる入力について説明する．ユーザは一つの質問意図をもっているものとする．その質問意図は複数の解が考え得るものであるとする．ユーザは質問意図を表すキーワードクエリを提案システムに入力するものとする．また，ユーザは質問意図に対する解情報を含む Web ページや QA コンテンツをいくつか選択し閲覧しているものとする．ここで，

QA コンテンツは一つの質問と一つ以上の回答の集合であり，ユーザは QA コンテンツそのものでなく，そこに含まれるいずれかの回答を選択しているものとする．本論文では，ユーザが選択した回答は全て，その回答が含まれる QA コンテンツの質問と組になっているものとみなす．これを QA ペアとよぶ．この，ユーザが選択したいくつかの Web ページおよび QA ペアも提案システムの入力となる．

ユーザが入力したキーワードクエリを k とおく．k は例えば（二日酔い 解消法）など，いくつかの単語で構成されるクエリとする．QA コンテンツを c_n とおく．c_n の質問を q_n とし，また，それに対する回答集合を $\{a_{n1}, ..., a_{ni}\}$ とすると QA ペアは $t_{ni} = (q_n, a_{ni})$，QA コンテンツは $c_n = \{t_{n1}, ..., t_{ni}\}$ とおくことができる．ここで，t_{ni} の添え字 n は QA コンテンツを一意に特定するためのものであり，添え字 i は c_n の回答集合において回答を一意に特定するためのものである．ただし，今後，QA ペア集合の要素としての QA ペアは $t_m = (q_m, a_m)$ と表現し，添え字 m は QA ペア集合において QA ペアを一意に特定するためのものとする．ユーザが選択した QA ペアの集合を $S_t = \{t_1, ..., t_m\}$，Web ページの集合を $S_p = \{p_1, ..., p_l\}$（Web ページを p_l とする）とおくと入力は k および (S_t, S_p) となる．

次に出力について説明する．出力は入力に対する補完情報を含む Web ページおよび QA ペアの集合である．ここで，別解情報の補完においては，別解情報が何種類得られたかが重要であると考えられる．例えば「二日酔いの解消法」について,「水を飲む」という解情報を含む Web ページおよび QA ペアのみが大量に別解情報として取得されるよりも,「水を飲む」という内容を含む Web ページや QA ペア,「柿を食べる」という内容を含む Web ページや QA ペア，といったように様々な解情報が別解情報として取得される方が補完情報として価値が高いと思われる．また，入力された S_t 及び S_p の要素と同じ内容の解情報を含む Web ページおよび QA ペアから，入力に対する追加情報を得ることができると考えられる．よって，ユーザの質問意図に対する解情報を含む Web ページおよび QA ペアを取得し，それらを内容毎に分類して提示するものとする．つまり，同一内容の解情報を含む Web ページおよび QA ペアを要素にもつ集合を，内容の種類の数だけ出力するということである．

与えられた入力から出力を返す流れを以下に示す.

Step 1　入力された検索クエリを検索エンジンに与えて，Web ページおよび QA コンテンツ（QA ペアの集合とみなす）を収集する

Step 2　収集された Web ページおよび QA ペアが解情報を含むか判定する

Step 3　Step 2 で解情報を含むとされた Web ページおよび QA ペアについて，情報の内容に基づく分類を行う

Step 2 では，Step 1 で収集された候補 Web ページ集合 $S'_p = \{p'_1, ..., p'_j\}$ および候補 QA ペア集合 $S'_t = \{t'_1, ..., t'_k\}$ についてその各々の要素が質問に対する解情報を含む可能性の高さを表すスコアを付ける．スコアの計算は下記の関数で行うものとする．

$$IncAns(S_t, S_p, \alpha'_i) \tag{1}$$

ここで，α'_i は p'_j または t'_i を表すものとする．このスコアが閾値以上となる α'_i を解情報を含む Web ページまたは QA ペアと判定する．

Step 3 では，Step 2 で解を含むと判定された Web ページおよび QA ペアの分類を行う．分類は解情報の内容に基づいたものである．同一内容の解情報を含む Web ページおよび QA ペアを要素にもつ集合を $G^m = \{S_p^m, S_t^m\}$ とおく．ここで，同一内容の解情報を含む Web ページ集合を $S_p^m = \{p_1^m, ..., p_x^m\}$，同一内容の解情報を含む QA ペア集合を $S_t^m = \{t_1^m, ..., t_y^m\}$ と表す．別解情報を含むと判定された Web ページおよび QA ペアの集合を $S^{alt} = \{S_p^{alt}, S_t^{alt}\}$ とおき，その集合の要素を内容毎に分類する関数を以下のように表すものとする．

$$G = Classify(S^{alt}) \tag{2}$$

ただし，$G = \{G^1, ..., G^m\}$ である．m は質問意図に対し，取得する事が出来た解情報の種類数である．

次節ではそれぞれの関数において我々の提案する手法を説明する．

4. 補完情報取得の実装

4.1 解を含む可能性を表すスコア計算

本節では収集された S'_p および S'_t の各要素 α'_i について，α'_i がユーザの質問意図に対する解情報を含む可能性の高さを表すスコア $IncAns(S_t, S_p, \alpha'_i)$ の計算手法について述べる．まず，候補ページ p'_j および候補 QA ペア t'_k を表す特徴ベクトルの生成について説明する．

4.1.1 Web ページを表す特徴ベクトルの生成

本研究では，Term Frequency（TF) をもちいて Web ページの特徴ベクトルを表現する．Web ページ内のテキストを形態素解析して得られる語集合について，語 w の Web ページ内における出現回数を $tf(w)$ とおくと，ページ p'_j の TF ベクトル $\mathbf{v'_{p'_j}}$ は以下のように表すことができる．

$$\mathbf{v'_{p'_j}} = \{tf(w) | w \in p'_j\} \tag{3}$$

また，$\mathbf{v'_{p'_j}}$ は以下のように正規化するものとする．

$$v_{p'_j}(w) = \frac{v'_{p'_j}(w)}{v'_{p'_j}(w_{max})} \tag{4}$$

ここで，語 w_{max} はベクトル $\mathbf{v'_{p_i}}$ において最大の値を持つ要素の語であり，式（4）は $\mathbf{v'_{p_i}}$ 内の最大の値を持つ要素の値を 1 とする正規化を意味する．この正規化は，Web ページの大きさの違いによって特徴ベクトルの大きさにも差が出ることを防ぐためである．

4.1.2 QA ペアを表す特徴ベクトルの生成

Web ページの場合と同様に TF をもちいて特徴ベクトルを表現する．QA ペア t'_k の TF ベクトルは以下のように求められる．

$$\mathbf{v'_{t'_k}} = \{tf(w) | w \in t'_k\} \tag{5}$$

ここで，t'_k の質問 q'_m および回答 a'_m を表す TF ベクトルの定義も行う．

$$\mathbf{v'_{q'_m}} = \{tf(w)|w \in q'_m\} \quad (6)$$

$$\mathbf{v'_{a'_m}} = \{tf(w)|w \in a'_m\} \quad (7)$$

$\mathbf{v'_{t'_k}}$，$\mathbf{v'_{q'_m}}$，$\mathbf{v'_{a'_m}}$ も式 (4) を用いて正規化を行う．正規化後のベクトルをそれぞれ $\mathbf{v_{t'_k}}$，$\mathbf{v_{q'_m}}$，$\mathbf{v_{a'_m}}$ とおく．候補 Web ページを表すベクトルおよび候補 QA ペアを表すベクトルを総称して候補ベクトルと呼ぶ．この候補ベクトルを用いた $IncAns(S_t, S_p, \alpha'_i)$ の計算手法について以下に述べる．

4.1.3 解を含む可能性を表すスコアの計算手法

(i) 質問意図を表すベクトルと候補ベクトルとの類似度を用いて判定する手法

これは，候補 α'_i が質問意図を表す特徴ベクトルに似た部分をもてば α'_i に解情報が含まれる可能性が高くなるという考え方である．

質問意図を表すベクトルの生成には大島らの研究 [3] における共通ベクトルという考え方を用いる．共通ベクトルとは，ある文書集合が与えられたときに，その全ての文書に共通する部分を表すベクトルを指す．これを我々の研究に当てはめると，質問を表す文書の集合から生成される共通ベクトルがすなわち質問意図を表すベクトルと見なせると考えられる．質問を表す文書の集合としては，入力された S_t の質問のみを要素とする集合 S_q が考えられる．

S_q の共通ベクトル $\mathbf{c_q}$（質問の共通ベクトル）の生成方法について述べる．S_q の各要素を q_x とおく．q_x を表すベクトル $\mathbf{v_{q_x}}$ を以下のように表すものとする．

$$\mathbf{v_{q_x}} = \{tf(w)|w \in q_x\} \quad (8)$$

本稿では共通ベクトルの表現方法として以下の 2 種類の方法を試みるものとする．

i. $c_q(w)$ の値が $\mathbf{v_{q_x}}(w)$ $(1 \leqq x \leqq n)$ の平均値となるようなベクトル (ただし n は S_q の要素数)

ii. $c_q(w)$ の値が $df_{S_q}(w)$（S_q における w の Document Frequency に相当）となるようなベクトル

i. は，共通ベクトルにおいて大きな値を持つ語は各質問 q_x における出現頻度がある程度高く，かつ，特定の質問のみで出現頻度の高い語であるべきではない，という考え方であり，大島らの手法に沿うものである．一方 **ii.** は，共通ベクトルにおいて大きな値を持つ語は，質問集合においてより多くの質問に出現するはずである，という考え方である．**i.** の方法で生成される共通ベクトルを $\mathbf{c_q^{av}}$，**ii.** の方法で生成される共通ベクトルを $\mathbf{c_q^{df}}$ とおくと，$\mathbf{c_q^{av}}$，$\mathbf{c_q^{df}}$ の値はそれぞれ以下の式で表される．

$$\mathbf{c_q^{av}}(w) = \frac{\sum_{i=1}^{n} v_{q_x}(w)}{n} \quad (9)$$

$$\mathbf{c_q^{df}}(w) = df_{S_q}(w) \quad (10)$$

この共通ベクトル $\mathbf{c_q}$（ただし $\mathbf{c_q}$ は $\mathbf{c_q^{av}}$ または $\mathbf{c_q^{df}}$）を用いて，p'_j や t'_k が解情報を含む可能性の高さを表すスコアを計算する．本稿では式 (1) を，候補ベクトルと質問の共通ベクトルとのコサイン類似度で表すものとする．

二つのベクトル $\mathbf{v_1}$，$\mathbf{v_2}$ のコサイン類似度を $\cos(\mathbf{v_1}, \mathbf{v_2})$ とおく．式 (1) を $\mathbf{c_q}$ と α'_i を表すベクトル $\mathbf{v_{\alpha'_i}}$ とのコサイン類似度によって求めるものとする．

$$IncAns(S_t, S_p, \alpha'_i) = \cos(\mathbf{c_q}, \mathbf{v_{\alpha'_i}}) \quad (11)$$

ここで，質問の共通ベクトルと候補ページとの類似度計算の場合は $\mathbf{v_{\alpha'_i}} = \mathbf{v_{p'_j}}$ であり，質問の共通ベクトルと候補 QA ペアとの類似度計算の場合は $\mathbf{v_{\alpha'_i}} = \mathbf{v_{q'_i}}$ である．

この式 (11) の値が高くなるような α'_i を，解情報を含む Web ページまたは QA ペアとみなす．

(ii) 解表現に共通する部分を表すベクトルと候補ベクトルとの類似度を用いて判定する手法

本稿では，ある質問に対する解の集合において解を表現する際に共通して現れる語の存在を仮定している．たとえば,「二日酔いの解消方法は？」という質問に対する解情報の周りには二日酔いの原因である“ 酒 ”という語が出現することが考えられる．また，二日酔いの解消方法として「ウコンを飲む」,「大量の水を飲む」というものがあるがこれらの解には“ 飲む ”という語が共通して現れている．このような，解に共通する特徴を表すベクトルを生成し，候補 α'_i とそのベクトルとの類似度を調べることで α'_i に解情報が含まれるか否かを判定することができると考えられる．これはつまり，α'_i に，解に共通する特徴を表すベクトルと類似する部分が存在すれば α'_i は解情報を含む可能性が高くなるという考え方である．

ここで，ユーザの質問意図に対する解を述べる際に共通する特徴を表すベクトル（解の共通ベクトル）の取得方法について述べる．解の共通ベクトルにおいて大きな値を持つ語は，入力された S_t の回答 a_m のみを要素とする集合（回答集合 S_a）において，各回答における出現頻度がある程度高く，かつ，特定の回答のみで出現頻度の高い語ではないと考えることができる．また，共通ベクトルにおいて大きな値を持つ語は，解集合においてより多くの解に出現するはずである，という考え方も可能である．よって，質問の共通ベクトル生成と同様の手法で解の共通ベクトルを生成する．

$$\mathbf{c_a^{av}}(w) = \frac{\sum_{i=1}^{n} v_{a_x}(w)}{n} \quad (12)$$

$$\mathbf{c_a^{df}}(w) = df_{S_a}(w) \quad (13)$$

この共通ベクトル $\mathbf{c_a}$（ただし $\mathbf{c_a}$ は $\mathbf{c_a^{av}}$ または $\mathbf{c_a^{df}}$）を用いて，p'_j や t'_k が解情報を含む可能性の高さを表すスコアを計算する．

$$IncAns(S_t, S_p, \alpha'_i) = \cos(\mathbf{c_a}, \mathbf{v_{\alpha'_i}}) \quad (14)$$

ここで，回答の共通ベクトルと候補ページとの類似度計算の場合は $\mathbf{v_{\alpha'_i}} = \mathbf{v_{p'_j}}$ であり，回答の共通ベクトルと候補 QA ペアとの類似度計算の場合は $\mathbf{v_{\alpha'_i}} = \mathbf{v_{a'_i}}$ である．

この式 (14) の値が高くなるような α'_i を，解情報を含む Web ページまたは QA ペアとみなす．

(iii) 質問の共通ベクトル，解の共通ベクトルと候補ベクトルと

の類似度を両方用いて判定する手法

これは，候補 α'_i が質問意図を表す特徴と類似した部分をもち，かつ解表現に共通する特徴と類似した部分を持つ場合に，α'_i に解情報が含まれる可能性が高くなるという考え方である．

上記の **(i)** で述べた質問の共通ベクトル $\mathbf{c_q}$ （ただし $\mathbf{c_q}$ は式 (9) で計算される $\mathbf{c_q^{av}}$ または式 (10) で計算される $\mathbf{c_q^{df}}$）および，**(ii)** で述べた解の共通ベクトル $\mathbf{c_a}$ （ただし $\mathbf{c_a}$ は式 (12) で計算される $\mathbf{c_a^{av}}$ または式 (13) で計算される $\mathbf{c_a^{df}}$），それぞれと候補 QA ペアまたは候補ページを表すベクトル $\mathbf{v}_{\alpha'_i}$ との類似度を用いて，式 (1) を以下のように表すものとする．

$$IncAns(S_t, S_p, \alpha'_i) = \frac{\cos(\mathbf{c_q}, \mathbf{v}_{\alpha'_i}) + \cos(\mathbf{c_a}, \mathbf{v}_{\alpha'_i})}{2} \quad (15)$$

ここで，質問の共通ベクトルおよび解の共通ベクトルと候補ページとの類似度計算の場合は $\mathbf{v}_{\alpha'_i} = \mathbf{v}_{\mathbf{p'_j}}$ であり，質問の共通ベクトルと候補 QA ペアとの類似度計算の場合は $\mathbf{v}_{\alpha'_i} = \mathbf{v}_{\mathbf{q'_i}}$，回答の共通ベクトルと候補 QA ペアとの類似度計算の場合は $\mathbf{v}_{\alpha'_i} = \mathbf{v}_{\mathbf{a'_i}}$ である．

この式 (15) の値が高くなるような α'_i を，解情報を含む Web ページまたは QA ペアとみなす．

(iv) 質問，解，入力された Web ページの共通ベクトルと候補ベクトルとの類似度を全て用いて判定する手法

上記の **(i)**〜**(iii)** は，入力された QA ペア集合がユーザの質問意図および，その質問に対する解表現に共通する性質を表すはずであるという考え方に基づいたものである．ここで，QA ペア集合だけでなく，入力された Web ページについても同様のことが言えると考えられる．つまり，質問意図を表す特徴，解表現に共通する特徴に加え，入力された Web ページに共通する特徴と類似した部分を持つ候補 α'_i が，解情報を含む可能性が高くなるという考え方である．しかし，Web ページは QA ペアとは異なり，質問と回答が明確に分けられていない．そこで，入力された Web ページに共通する部分を表すベクトルは，ユーザの質問意図に共通する特徴および，それに対する解表現に共通する特徴を共に表すベクトルとみなす．Web ページの共通ベクトルも，質問や解の共通ベクトル生成と同様の手法で生成する．

$$\mathbf{c_p^{av}}(w) = \frac{\sum_{i=1}^{n} v_{p_x}(w)}{n} \quad (16)$$

$$\mathbf{c_p^{df}}(w) = df_{S_p}(w) \quad (17)$$

質問の共通ベクトル $\mathbf{c_q}$ （ただし $\mathbf{c_q}$ は $\mathbf{c_q^{av}}$（式 (9)）または $\mathbf{c_q^{df}}$（式 (10)）），解の共通ベクトル $\mathbf{c_a}$ （ただし $\mathbf{c_a}$ は $\mathbf{c_a^{av}}$（式 (12)）または $\mathbf{c_a^{df}}$（式 (13)））および，入力された Web ページ集合の共通ベクトル $\mathbf{c_p}$ （ただし $\mathbf{c_p}$ は $\mathbf{c_p^{av}}$ または $\mathbf{c_p^{df}}$）の各々と候補 QA ペアまたは候補ページを表すベクトル $\mathbf{v}_{\alpha'_i}$ との類似度を用いて，式 (1) を以下のように表すものとする．

$$IncAns(S_t, S_p, \alpha'_i) = \frac{\cos(\mathbf{c_q}, \mathbf{v}_{\alpha'_i}) + \cos(\mathbf{c_a}, \mathbf{v}_{\alpha'_i}) + \cos(\mathbf{c_p}, \mathbf{v}_{\alpha'_i})}{3} \quad (18)$$

ここで，質問の共通ベクトル，解の共通ベクトル，および入力された Web ページ集合の共通ベクトルと候補ページとの類似度計算の場合は $\mathbf{v}_{\alpha'_i} = \mathbf{v}_{\mathbf{p'_j}}$ であり，質問の共通ベクトルと候補 QA ペアとの類似度計算の場合は $\mathbf{v}_{\alpha'_i} = \mathbf{v}_{\mathbf{q'_i}}$，回答の共通ベクトルと候補 QA ペアとの類似度計算の場合は $\mathbf{v}_{\alpha'_i} = \mathbf{v}_{\mathbf{a'_i}}$，Web ページ集合の共通ベクトルと候補 QA ペアとの類似度計算の場合は $\mathbf{v}_{\alpha'_i} = \mathbf{v}_{\mathbf{t'_i}}$ である．

この式 (18) の値が高くなるような α'_i を，解情報を含む Web ページまたは QA ペアとみなす．

4.2 分　　類

本節では，4.1.3 節で得られた解情報を含む Web ページまたは QA ペアを解の内容毎に分類する手法について説明する．

本論文では，ベクトル間のコサイン類似度を用いたクラスタリングで分類を行うものとする．ここで，解を含む Web ページ p_n^{ans} および QA ペア t_m^{ans} を TF ベクトルで表すと，そのベクトルは p_n^{ans} や t_m^{ans} に含まれる質問およびその解を表すベクトルとなる．しかし，解情報を含む Web ページや QA ペアを表す TF ベクトルが質問の特徴を含むと，ある問題が生じる．すなわち，質問の特徴によってベクトル間のコサイン類似度が高くなり，Web ページや QA ペアに現れる解の内容が類似していなくても，類似した情報として同一クラスタに分類されてしまうということが起こりうる．これは，解の内容に基づくクラスタリングを実現していないことになる．よって，解を含む Web ページ p_n^{ans} および QA ペア t_m^{ans} から，解情報のみを表す TF ベクトルを生成する必要がある．そのような TF ベクトルの生成手法を以下に提案する．

(1) 質問に共通する部分を表すベクトルを除いたものをクラスタリングする手法

解を含む Web ページを表すベクトル $\mathbf{p_n^{ans}}$ や解を含む QA ペアを表すベクトル $\mathbf{t_m^{ans}}$ から質問に共通する部分を表すベクトルを除去する事で解のみを表す TF ベクトル $\mathbf{u_i}$ を生成することが考えられる．ベクトル $\mathbf{u_i}$ の値を以下の式で計算する．

$$u_i(w) = max(\alpha_i^{ans}(w) - c_q(w), 0) \quad (19)$$

式 (19) は，減算の結果がマイナスの値をとるものは，その値を 0 に置き換えることを意味する．$\mathbf{c_q}$ は式 (9) または式 (10) によって生成される，質問の共通ベクトルを指す．Web ページのベクトル生成の場合は $\alpha_{\mathbf{i}}^{\mathbf{ans}} = \mathbf{v}_{\mathbf{p_i}}^{\mathbf{ans}}$ であり，QA ペアのベクトル生成の場合は $\alpha_{\mathbf{i}}^{\mathbf{ans}} = \mathbf{v}_{\mathbf{a_i}}^{\mathbf{ans}}$ である．

この，解を含む Web ページまたは QA ペアを表すベクトルから質問の共通ベクトルを引いたベクトルをクラスタリングの計算の対象とする事を提案する．このベクトルの集合を U^q とおくと，式 (2) は $G = Classify(U^q)$ となる．

(2) 質問に共通する部分および解に共通する部分を表すベクトルを除いたものをクラスタリングする手法

p_n^{ans} や t_m^{ans} には，解を表現する際に共通する語が含まれていると考えられる．「二日酔いの解消法」に対する解情報で「肝臓の働きを活発にするしじみが良い」という解と「水を飲めば肝臓の負担を減らすことができる」という解がある場合を考える．これらの解は，内容としてはそれぞれ「しじみ」を用いる解と「水」を用いる解であり，別々の内容である．しかし，そ

表 1　実験に用いた検索クエリ

	クエリ		クエリ
1	英単語　覚え方	6	眠気　覚ます
2	二の腕　細く　方法	7	髪の毛　さらさら　方法
3	京都　縁結び　神社	8	米　おいしく　炊く
4	犬　無駄吠え　しつけ	9	切花　長持ち　方法
5	花粉症　効く　食べ物	10	二日酔い　解消法

の解を説明する際に「肝臓」という語が共通して現れている．このような語の存在によって $\mathbf{p_n^{ans}}$ および $\mathbf{t_m^{ans}}$ のコサイン類似度が高くなり，内容の異なる解情報を含むベクトル同士が同一クラスタに分類されてしまうことも起こりうると考えられる．そこで，解を含む Web ページを表すベクトル $\mathbf{p_n^{ans}}$ や解を含む QA ペアを表すベクトル $\mathbf{t_m^{ans}}$ から質問に共通する部分を表すベクトルと，解を表現する際に共通する部分を表すベクトルを除去する事で，その解に特有な部分を表す TF ベクトル $\mathbf{u_i}$ を生成することが考えられる．ベクトル $\mathbf{u_i}$ の値を以下の式で計算する．

$$u_i(w) = max(\alpha_i^{ans}(w) - (c_q(w) + c_a + c_p), 0) \qquad (20)$$

$\mathbf{c_q}$ は式 (9) または式 (10) によって生成される，質問の共通ベクトルを指す．また，$\mathbf{c_a}$ は式 (12) または式 (13) によって生成される解の共通ベクトルを指し，$\mathbf{c_p}$ は式 (16) または式 (17) によって生成される入力 Web ページ集合の共通ベクトルを指す．Web ページのベクトル生成の場合は $\alpha_\mathbf{i}^\mathbf{ans} = \mathbf{v_{p_i}^{ans}}$ であり，QA ペアのベクトル生成の場合は $\alpha_\mathbf{i}^\mathbf{ans} = \mathbf{v_{a_i}^{ans}}$ である．

式 (20) で計算されるベクトルの集合を U^{qap} とおくと，式 (2) は $G = Classify(U^{qap})$ となる．

5. 評価実験

5.1 テストセット

入力された QA ペア集合や Web ページ集合に対する補完情報取得手法についての評価実験を行った．実験のために，*task-oriented* 型質問を表す検索クエリを 10 個人手で用意した．このクエリはキーワードクエリとする．表 1 に用意したクエリを示す．

さらに，そのクエリで QA サイト内検索および Web 検索を行い，一つのクエリに対しそれぞれ QA コンテンツと Web ページを 3 個ずつ収集した．この QA コンテンツおよび Web ページは，検索クエリが表す質問の内容に沿うものであり，かつそれに対する解情報が含まれるものを著者が選択した．また，それぞれに含まれる解情報の内容が重複しないように選択した．

QA コンテンツについては，QA コンテンツの質問 q が一つの事を尋ねているものという条件を満たすものとする．これは，質問 q が検索クエリが表す質問に加えて別の事を尋ねている場合，QA コンテンツ内の回答がどの質問内容に対して回答しているかの判別が必要になるためである．

Web 検索結果に QA コンテンツを含む Web ページが含まれることが予想されるが，本研究は QA コンテンツに対し，別解情報をもつ QA コンテンツ以外の Web ページを取得する事を目的とするため，Yahoo!知恵袋，教えて!goo(注2)，人力検索はてな(注3)，OKWave(注4) のサイト内の Web ページは実験の対象としないものとする．

以上のクエリと QA コンテンツ集合および Web ページ集合を入力として提案手法の評価を行った．

5.2 Web ページおよび QA コンテンツが解情報を含むかどうかのスコアの評価

式 (11),，式 (14)，式 (15) および式 (18) の値の妥当性について評価を行った．この値は Web ページまたは QA ペアが入力に対する解情報を含むかどうかを表すものである．

入力として与えられた検索クエリ k を Web 検索エンジンおよび QA サイト内の検索エンジンに与えて検索結果を取得した．Web 検索エンジンは Yahoo!検索 Web API(注5)を，QA サイト内の検索エンジンは Yahoo!知恵袋 Web API(注6)を利用した．検索エンジンの出力順に検索結果から上位 100 件までの Web ページおよび上位 50 件までの QA コンテンツを収集した．QA コンテンツは質問と回答の QA ペアの集合とみなす．

収集した Web ページ集合の各 Web ページ p'_j について，式 (4) を用いてページ p'_j の特徴ベクトル $v'_{p'_j}$ を作成する．また，収集した QA ペア集合の各 QA ペア t'_k について，式 (6)，式 (7) および式 (5) を用いて質問，回答の特徴ベクトル $v'_{q'_m}$，$v'_{a'_m}$，および組の特徴ベクトル $v'_{t'_k}$ を作成する．

我々は 4.1.3 節において，ユーザが持つ質問意図に対する解情報を含む可能性を表すスコアの計算手法について以下の 4 種類の手法を提案した．

(Q) 入力された QA ペア集合の質問集合の共通ベクトルを用いる手法（式 (11)）

(A) 入力された QA ペア集合の回答集合の共通ベクトルを用いる手法（式 (14)）

(QA) 入力された QA ペア集合の質問集合の共通ベクトルと回答集合の共通ベクトルを用いる手法（式 (15)）

(QAP) 入力された QA ペア集合の質問集合の共通ベクトルと回答集合の共通ベクトル，および入力された Web ページ集合の共通ベクトルを用いる手法（式 (18)）

また，共通ベクトルの生成方法について，

(ave) ベクトル集合の平均値を共通ベクトルの値とする方法（式 (9)，式 (12)，式 (16)）

(df) ベクトル集合における DF 値に相当する値を共通ベクトルの値とする方法（式 (10)，式 (13)，式 (17)）

の 2 種類の方法を提案した．これらを組み合わせた **(Qave)**，**(Qdf)**，**(Aave)**，**(Adf)**，**(QAave)**，**(QAdf)**，**(QAPave)**，**(QAPdf)** の 8 種類の手法について，$IncAns(S_t, S_p, \alpha'_i)$ の値を収集した Web ページおよび QA ペアについて計算し，値が降順になるように Web ページ集合の各ページおよび QA ペア集合の各 QA ペアを並び替えた．並び替えた後の上位 10 ペー

(注2)：http://oshiete.goo.ne.jp/
(注3)：http://q.hatena.ne.jp/
(注4)：http://okwave.jp/
(注5)：http://developer.yahoo.co.jp/webapi/search/
(注6)：http://developer.yahoo.co.jp/webapi/chiebukuro/

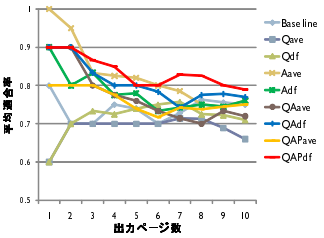

図 1　解を含む Web ページの平均適合率

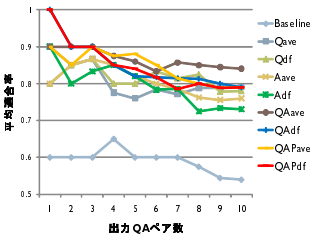

図 2　解を含む QA ペアの平均適合率

ジ，10QA ペアについて，解情報を含むかどうかを調べた．なお，解情報を含むかどうかの判定は著者らが行った．解情報を含むページを正解ページとみなし，並び替えた後の上位 x ページ（$1 \leqq x \leqq 10$）における適合率を各手法について計算した．この操作を用意した 10 個のクエリについて行った．また，比較のため，Web 検索エンジンの返す検索結果順の上位 10 ページおよび，QA 検索エンジンの返す検索結果順の上位 10 ペアについても同様に正解ページ数を調べた．この結果を Baseline と呼ぶ．各手法について，10 個のクエリについてそれぞれ求めた適合率の平均をとった結果を図 1 および図 2 に示す．

図 1 より，Baseline を上回る結果となったのは手法 **(QAdf)** と手法 **(QAPdf)** のみであった．また，4 種類の提案手法を比較すると共通ベクトルの作成方法については，**(A)** のみが **(ave)** の方法での共通ベクトル生成が **(df)** を上回ったがそのほかの手法では **(df)** の方法で共通ベクトルを生成する方が良い適合率が得られることが分かった．

図 2 より，解を含む QA ペアの取得においては提案手法の全てが Baseline を上回る結果となった．この要因の一つは，Yahoo!知恵袋 API の返す検索結果は検索クエリを含む QA コンテンツを投稿日時順に並べたものであるためである．しかし，提案手法は Baseline を上回る以外にも，結果の上位 10 件までで適合率が 0.7 以上を維持しているなど，解を含む QA ペアの取得において提案手法には一定の効果があるといえる．

今回は実験数が上位 10 件と少なく，解を含む Web ページの取得においては，提案手法が良いとは一概に言えなかった．また，解を含む QA ペアの取得においても，どの提案手法が良いかを決めることはできないという結果になった．今後実験数を増やすことで提案手法が有効であったかどうかをより詳細に調べたい．

5.3　解情報のクラスタリングの評価

まず，解情報を含む Web ページ集合および QA ペア集合のクラスタリングを行うにあたり，注意すべき点について述べる．解を含む Web ページや QA ペアは，一つの解情報を含むとは限らない．つまり，解の内容に基づいてあるクラスタ G^x に分類された p_n^{ans} や t_m^{ans} が，同時に別のクラスタ G^y の要素となる場合もある．よってクラスタリングはハードクラスタリング（要素が一つのクラスタに属する事のみ許容）ではなくソフトクラスタリング（要素が一つ以上のクラスタに属する事を許容）を用いるべきと考えられる．

また，別解情報が何種類あるかを事前に知ることは困難と思われる．そのため，あらかじめクラスタ数を設定する非階層的クラスタリングではなくクラスタ数を設定する必要のない階層的クラスタリングを用いるべきであると思われる．

ただし，本稿におけるクラスタリングの実装はファジー c 平均法を用いるものとする．これは，クラスタ数を事前に設定する必要があるが，ソフトクラスタリングを実現するものである．

4.2 節で述べた 2 通りのクラスタリング手法について考察を加える．4.2 節では，以下のようなクラスタリング手法を提案した．

(1) 候補ベクトルから質問の特徴を除いたベクトルを生成し，クラスタリングを行う

(2) 候補ベクトルから質問および回答に共通する特徴を除いたベクトルを生成し，クラスタリングを行う

この 2 通りに加え，候補ベクトルをそのままクラスタリングした場合の 3 通りについて実験を行った．ファジー c 平均法におけるファジー度は著者らが適当な値を選んだ．

生成された各クラスタについての考察を述べる．ただ候補ベクトルをクラスタリングした場合および，**(1)** 手法では解の特徴毎にクラスタリングされているとは言い難い結果であった．**(1)** 手法にあまり効果がなかった理由としては，解に共通する語の影響があったと考えられる．例えば，「二日酔いの解消法」について，解に現れやすい語としては“ アセトアルデヒド ”や“ アルコール ”といった語がある．これらは解の内容に関わる語というよりは解を詳細に説明する際に用いられる語である．このような語が高い値をもつベクトルをそのままクラスタリングしたために解の内容毎のクラスタリングにならなかったと考えられる．**(2)** 手法では，簡潔に表すことのできる解情報については解の内容に基づくクラスタリングが出来た事例もあった．例えば，「二日酔いの解消法」についての“ 風呂に入る ”という解や，「切り花を長持ちさせる方法」についての“ 砂糖を入れる ”という解など，名詞＋動詞で表せる単純な解については適切なクラスタが生成されていた．しかし，全体的な結果としては，どちらの手法も解の内容毎の適切なクラスタを生成する事ができておらず，クラスタリングについては今後も効果的な手法を考案する必要があることがわかった．

また，生成したクラスタから補完情報（別解情報および追加情報）を抽出する手法についても考えていかなければならない．

6.　ま　と　め

ある質問意図に対する解情報を含む Web ページおよび QA コンテンツが与えられた際に，その解情報に対する別解情報を Web ページおよび QA コンテンツから取得して相互的に補完

することを提案した．本稿では *conversational* question であり，かつ *task-oriented* question であるような質問意図を対象とし，それに対して得られた解情報に対する別解情報の取得手法について述べた．我々は，与えられた入力から解情報を出力する手法について複数の段階に分けて提案した．提案手法ではまず，入力されたクエリを用いて解を含む Web ページおよび QA ペアの候補集合を取得する．次に取得した各候補について，質問意図に対する解情報を含むかを判定するためのスコアを計算する．最後に，解情報を含む Web ページおよび QA ペアを，その解情報の内容に基づいて分類を行うというものである．

提案手法について，2 つの実験を行った．まず，解を含む可能性の高さを表すスコアの妥当性評価を行った．次に，解を含む Web ページおよび QA ペアのクラスタリングを行った．解を含む Web ページおよび QA ペア取得については，Baseline を上回る提案手法があることを確認した．しかし，実験数が少ないこともあり，提案手法の有効性についてはまだ示すに至っていない．クラスタリングの実験についても，提案手法が有効であったとはいえず，その評価方法を含め，今後の課題として取り組んでいかなければならない．

謝　　辞

本研究の一部は，京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」，および，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」，計画研究「情報爆発時代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，A01-00-02，課題番号：18049041），および，文部科学省科学研究費補助金若手研究（B）「オンデマンド利用を目的とする Web からの知識発見に関する研究」（研究代表者：大島裕明，課題番号：21700105），および，NICT 高度通信・放送研究開発委託研究「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 課題ア Web コンテンツ分析技術」（研究代表者：田中克己）によるものです．ここに記して謝意を表します．

文　　献

[1] F. M. Harper, D. Moy and J. A. Konstan: “Facts or friends?: distinguishing informational and conversational questions in social q&a sites”, pp. 759–768 (2009).

[2] D. Metzler and W. B. Croft: “Analysis of statistical question classification for fact-based questions”, Inf. Retr., **8**, 3, pp. 481–504 (2005).

[3] 大島裕明, 小山聡, 田中克己：“文書群を問合せとした兄弟カテゴリー文書の検索”, 電子情報通信学会論文誌 D, **J90-D**, 2, pp. 196–208 (2007).

[4] 高田夏希, 大島裕明, 田中克己：“Web と QA コンテンツの相互補完に基づくソーシャルサーチ”, WebDB (2010).

[5] J. Carbonell and J. Goldstein: “The use of mmr, diversity-based reranking for reordering documents and producing summaries”, pp. 335–336 (1998).

[6] X. Xue, J. Jeon and W. B. Croft: “Retrieval models for question and answer archives”, pp. 475–482 (2008).

[7] J. Jeon, W. B. Croft and J. H. Lee: “Finding similar questions in large question and answer archives”, pp. 84–90 (2005).

[8] K. Wang, Z. Ming and T.-S. Chua: “A syntactic tree matching approach to finding similar questions in community-based qa services”, pp. 187–194 (2009).

[9] F. M. Harper, D. Raban, S. Rafaeli and J. A. Konstan: “Predictors of answer quality in online q&a sites”, pp. 865–874 (2008).

[10] E. Agichtein, C. Castillo, D. Donato, A. Gionis and G. Mishne: “Finding high-quality content in social media”, pp. 183–194 (2008).

[11] J. Bian, Y. Liu, E. Agichtein and H. Zha: “Finding the right facts in the crowd: factoid question answering over social media”, pp. 467–476 (2008).

[12] 近藤光正, 中辻真, 田中明通, 内山匡：“重要語抽出を用いた外部 API からの関連コンテンツ推薦”, 第 24 回人工知能学会全国大会 JSAI2010, 1D2-1 (2010).

[13] 小谷彬, 小山聡, 田中克己：“複数 Web サイトからの共通側面の抽出と類似サイト検索”, 電子情報通信学会第 17 回データ工学ワークショップ DEWS2006 論文集, 2C-i9 (2006).