CM-002

Q&A サイトにおける回答者クラスタリングに基づく質問マッチング
Question Matching in Q&A Site Based on Clustering of Answerers

小川 翔大‡　内田 真人‡
Shota Ogawa　Masato Uchida

1. はじめに

インターネットの普及に伴い，人々の情報伝達様式が変化してきている．インターネットを利用することで，素早く多くの情報が手に入るようになった．しかし，インターネット上で見つけることのできない情報も存在する．また，情報量が多くなった反面，必要な情報を探し出すのが困難になりつつある．そこで，こうした困難を解消するために，Yahoo!知恵袋 [1]や，OKWAVE[2]といった Q&A サイトが利用されている．Q&A サイトとはユーザーが知りたい情報を質問することで，他のユーザーが回答してくれるサービスである．そのため，このような Q&A サイトを利用することで，自分で見つけられなかった情報を得ることや，個人的な状況に対する意見を募ることが可能となる．しかし，必ず回答してもらえるとは限らず，適切でない回答がされる可能性もある．Q&A サイトを利用するユーザーは自分が投稿した質問に対して回答がされるまで待たなければならない．また，回答者は自分が回答できる質問を探すのに手間がかかる．したがって，Q&A サイトの利用者が効率的に情報を共有できるシステムの実現が必要である．

Q&A サイトでは，質問に対して適切な回答者を迅速に探し出すことで，利用者がより効率的に情報を共有することができる．質問と回答者を適切にマッチングすることができれば，回答者自身が質問を探す手間を省くことができ，質問者は素早く情報を共有することができる．ある質問が投稿されたとき，その質問を適切な回答者へとマッチングすることが本研究の目的である．質問者に対して過去に回答したことのある回答者，もしくは類似した回答者であれば回答できる可能性が高い．そこで本研究ではまず，回答者をクラスタリングする．そして，質問者の過去の質問に対する回答履歴に基づいて回答者をクラスタリングすることで質問を適切な回答者へとマッチングする手法を提案する．本研究では，Yahoo!知恵袋における実際のデータを用いて実験を行い，提案手法の有効性を示す．

2. 関連研究

Q&A サイトに関する研究は数多く行われている．Q&A サイトでは，ユーザーが質問をすることで他のユーザーからの回答を得ることができ，情報を共有することができる．このように Q&A サイトではユーザー同士が質問と回答のやり取りを行うことによってコミュニティを形成する．Q&A サイトのコミュニティを活性化させるシステムを考案するためには，ユーザーの行動的特徴を把握する必要がある．そこで佐藤ら[1]は，Q&A サイトにおけるコミュニティを分析するために，ユーザーのコミュニティにおける貢献度を定義することで，カテゴリの特性とユーザーの特性を表現できることを示した．

また，ユーザーが実際にどのような内容の質問・回答をしているかを把握する必要もある．栗山ら[2]は，Yahoo!知恵袋の質問と回答を分析した．その結果，質問は，客観的な正解が存在するタイプと，主観的な意見や嗜好を求めるタイプに大きく分けることができ，目的や意図に応じてさらに細かいタイプに分類できることを示した．また，質問のタイプごとの特徴的な表現を抽出して，質問のタイプが自動分類できる可能性を示した．

本研究と同様に，質問と回答のマッチングを目的とした研究も行われている．岩間ら[3]は，サポートセンターにおける質問に対する応答として，過去の投稿の中から類似した質問を探し出し，それに対する回答をマッチングさせる手法を提案している．これは，サポートセンターへの問い合わせ質問の大半が，類似した質問が過去に存在するものである特性を利用した手法である．しかし，Q&A サイトでは必ずしも過去に類似した質問や回答があるわけではないため，そのまま Q&A サイトへと適用することは難しい．

片山ら[4]は，過去に回答した質問とそれに対する回答をその回答者の知識として考え，知識と質問の類似度を定義した．そして，実際に回答した質問とそうでない質問とでそれぞれ類似度を比較した．その結果，実際に回答した質問の方が，類似度が高くなりやすいことを示し，質問マッチングへと応用できる可能性を示した．

甲谷ら[5]は，Q&A サイトにおける質問と回答の関係をネットワークとして捉え，その成長パターンに着目し，質問者と回答者のリンク予測を行うことで質問を回答者へとマッチングする手法を提案した．この研究ではネットワークの構造パターンに着目して質問をマッチングする．これに対し，本研究は回答者の類似性に着目した質問のマッチングについて検討する．

本研究では，回答者同士の類似度を定義してクラスタリングを行う．そして，質問者の過去の質問に対する回答履歴に基づいて，クラスタリングの結果を利用することで質問を適切な回答者へとマッチングするシステムを提案する．

3. 提案手法

Q&A サイトを利用するユーザーは自分が投稿した質問に対して回答がされるまで待たなければならない．また，回答者は自分が回答できる質問を探すのに手間がかかる．そこで，本研究では，質問者の過去の質問に対する回答履歴に基づいて回答者をクラスタリングした結果を利用することで，質問を適切な回答者へとマッチングするシステムを提案する．

提案手法の概要を図 3.1 に示す．提案手法では，まず，回答者がどのような質問に対して回答していたかを調べるため，過去のある期間内における回答の履歴を取得する．この期間を学習期間と呼ぶこととする．そしてその情報を元に，回答者ごとの特徴ベクトルを作成してクラスタリングを行う．その後，クラスタリングの結果を利用した回答者へのマッチングモデルを適用することで，学習期間以降

‡ 早稲田大学基幹理工学研究科情報理工・情報通信専攻，
Department of Computer Science and Communications Engineering, Waseda University, Tokyo, Japan
[1]https://chiebukuro.yahoo.co.jp/
[2]https://okwave.jp

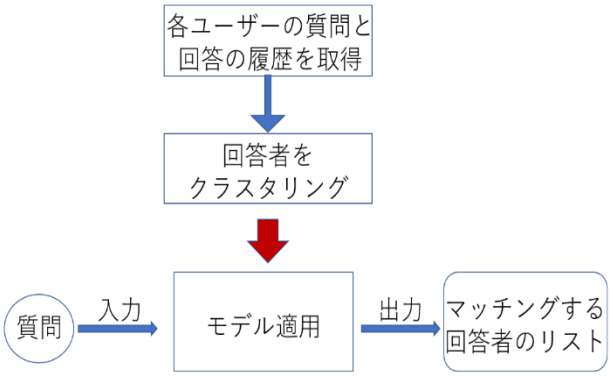

図 3.1　提案手法の概要

に投稿された質問を回答者に推薦する．以下では，提案手法の詳細について述べる．

3.1 特徴ベクトルの作成

提案手法では回答者をクラスタリングするために，各回答者の特徴ベクトルを作成する．以下でその詳細について説明する．

3.1.1 回答者の類似度

回答者を分類するにあたって，回答者同士の類似度を定義する必要がある．図 3.2 に示すような質問者と回答者の関係のモデルを考える．各ノードは質問者または回答者を表しており，回答者のノードから質問者のノードへと入るエッジは質問者が投稿したいずれかの質問に回答者が回答していることを表す．また，エッジの横の数字はその質問者に回答した回数を表す．図 3.2 の場合，回答者a_1とa_2は共通の質問者q_2とq_3に対して回答をしている．一方，回答者a_3は質問者q_4とq_5に対して回答をしているが，a_1，a_2と共通の質問者はいない．このとき回答者a_1とa_2は共通の質問者に対して回答をしているため，共通する知識を持っていると考えることができる．このことから，回答者a_1とa_2は類似しているといえる．逆に，回答者a_3はa_1，a_2のどちらとも類似していないといえる．すなわち，共通して回答している質問者が多いほどその回答者同士は類似しているといえる．この考えに基づき，本研究では 2 種類の方法で回答者の特徴ベクトルを定義した．

3.1.2 手法 1

一つ目は，ある二人の回答者が，共通の質問者に対して回答したかどうかで特徴づける方法である．回答者a_iが回答したことのある質問者の集合をQ_{a_i}と定義し，各要素が式 (3.1) となる対称行列$\boldsymbol{M}$を考える．

$$m_{ij} = \begin{cases} 1, & (|Q_{ai} \cap Q_{aj}| > 0) \\ 0, & (|Q_{ai} \cup Q_{aj}| > 0) \end{cases} \tag{3.1}$$

ある 2 人の回答者a_iとa_jに着目したときに，式 (3.1) は共通の質問者に 1 人でも回答していた場合は1となり，共通した質問者が 1 人もいなかった場合は0となる．また，$\boldsymbol{M}$の対角成分は同一の回答者同士の比較になるため，すべて1になる．$\boldsymbol{M}$の各行が回答者一人の特徴ベクトルとなり，次元数は回答者数と等しくなる．例として，図 3.2 における対称行列$\boldsymbol{M}$は次のようになる．

$$M = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.2}$$

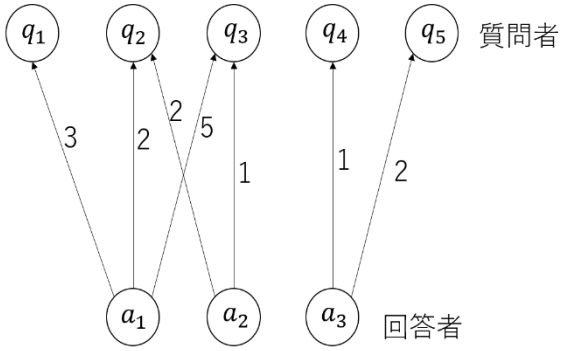

図 3.2　質問者と回答者の例

3.1.3 手法 2

二つ目は，自然言語処理で用いられることが多い指標である TF-IDF を利用した方法である．これは，その文書内でのある単語の出現度である TF (Term Frequency)と，文書間でのある単語の出現度に応じた数値である IDF (Inverse Document Frequency)の積で定義される指標である．片山ら[4]は，質問と回答履歴の類似度を計算する際にこの指標を利用している．本研究では文書を回答者，単語を質問者とみなしてこの指標を適用した．

回答者a_iの質問者q_jへの回答頻度$TF_{a_iq_j}$は式 (3.3) のように表される．

$$TF_{a_iq_j} = \frac{n_{a_iq_j}}{N_{a_i}} \tag{3.3}$$

$n_{a_iq_j}$は回答者a_iが質問者q_jに回答した回数，N_{a_i}は回答者a_iの総回答数である．つまり，回答者a_iが回答したすべての質問のうち，質問者q_jが投稿したものがどれだけ占めているかを表す数値である．$TF_{a_iq_j}$は$0 \leq TF_{a_iq_j} \leq 1$の範囲の値をとる．回答者$a_i$のベクトル$\boldsymbol{TF}_{a_i}$は，$TF_{a_iq_j}$を$j$について順に並べたものとなり，ベクトルの次元数は質問者数と等しくなる．例えば，図 3.2 における$\boldsymbol{TF}_{a_i}$は式 (3.4) のようになる．

$$\boldsymbol{TF}_{a_i} = \begin{pmatrix} TF_{a_1q_1} \\ TF_{a_1q_2} \\ TF_{a_1q_3} \\ TF_{a_1q_4} \\ TF_{a_1q_5} \end{pmatrix} = \begin{pmatrix} 0.3 \\ 0.2 \\ 0.5 \\ 0.0 \\ 0.0 \end{pmatrix} \tag{3.4}$$

質問者の回答されてない度合いを出現稀少度とすると，質問者q_jの出現希少度IDF_{q_j}は式 (3.5) のように表される．

$$IDF_{q_j} = -\log\left(\frac{d_{q_j}}{D}\right) \tag{3.5}$$

d_{q_j}は質問者q_jに回答をしたことのある回答者の数，Dは回答者数である．$d_{q_j} \leq D$であるから，$IDF_{q_j} \geq 0$となる．例えば，質問者q_jがすべての回答者に回答されていた場合，$d_{q_j} = D$ となり，質問者q_jの出現希少度IDF_{q_j}は0となる．ある質問者に回答している回答者が少ないほど，その質問者は回答者をより特徴付けるものと考えることができ，出現希少度は大きくなる．ベクトル$\boldsymbol{IDF}$はIDF_{q_j}をjについて順に並べたものとなり，次元数は TF と同様に質問者数と等しくなる．例として，図 3.2 における$\boldsymbol{IDF}$は式 (3.6) のようになる．

$$\boldsymbol{IDF} = \begin{pmatrix} IDF_{q_1} \\ IDF_{q_2} \\ IDF_{q_3} \\ IDF_{q_4} \\ IDF_{q_5} \end{pmatrix} = \begin{pmatrix} 1.10 \\ 0.41 \\ 0.41 \\ 1.10 \\ 1.10 \end{pmatrix} \tag{3.6}$$

質問者q_1，q_4，q_5は一人の回答者からのみ回答されているため出現稀少度が高くなっている．一方，質問者q_2，q_3は二人の回答者から回答されているため出現稀少度が低くなっている．手法 2 では，以上のように TF と IDF を導出し，$TF_{a_i q_j} IDF_{q_j}$をjについて順に並べたものを回答者a_iの特徴ベクトルとする．

3.2 クラスタリング

3.1 節の方法で作成した 2 種類の特徴ベクトルそれぞれに対してクラスタリングを行う．クラスタリングの手法としてk-means 法を用いた．k-means 法とはn個のデータを既定の個数k個のクラスタに分類する手法である[6].

k-means 法では，データとなるベクトル間の距離を用いてクラスタリングを行うため，距離を計算する関数が必要となる．手法 1 ではベクトル間の距離の定義を式(3.7)にあるユークリッド距離とし，手法 2 では式(3.8)に表されるコサイン類似度を用いて$1 - \cos(\boldsymbol{p}, \boldsymbol{q})$ をベクトル間の距離とした．ただし，$\boldsymbol{p}$と$\boldsymbol{q}$は二つのデータ$\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$，$\boldsymbol{q} = (q_1, q_2, \ldots, q_n)$ を表す．

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{3.7}$$

$$\cos(\boldsymbol{p}, \boldsymbol{q}) = \frac{< \boldsymbol{p}, \boldsymbol{q} >}{||\boldsymbol{p}|| ||\boldsymbol{q}||} \tag{3.8}$$

3.3 マッチングモデル

最後に，回答者へと質問をマッチングするモデルについて説明する．ある質問が投稿されたとき，その質問に対する適切な回答者を探すのが本研究の目的である．質問者に対して過去に回答したことのある回答者，もしくは類似した回答者であれば回答できる可能性が高い．そこで提案手法では，質問者の過去の質問に対する回答履歴に基づいて，クラスタリングの結果を利用することで質問を適切な回答者へとマッチングする．

回答者をk個のクラスタに分割した場合を考える．このとき図 3.3 のように，学習期間に回答を行った回答者全体の集合Aと，クラスタリングによって得られた，類似した回答者から成る部分集合$C_i\ (i = 1,2, \ldots, k)$ がある．また，学習期間以降に質問者q_jが質問を投稿したとする．このとき，その質問者が学習期間に投稿した質問に対する回答者に応じて，各クラスタ$C_i\ (i = 1,2, \ldots, k)$ からそれぞれ適切な人数の回答者を選んでマッチングする．これにより，回答者全体Aから選ぶ場合よりも適切な回答者へと質問をマッチングすることができ，Q&A サイトの利便性の向上が見込める．

そこで本研究では，質問者の過去の履歴に基づいて各クラスタでマッチングする人数の分配を決定する．いま，学習期間以降に質問者q_jが質問を投稿したとする．このとき，q_jが学習期間に投稿した質問に対して回答した回答者を考えると，各回答者はいずれかのクラスタに属する．クラスタ数をk，マッチングする合計回答者数をn人，学習期間におけるクラスタ$C_i\ (i = 1,2, \ldots, k)$ に属する回答者からの回答数の総和を$f(C_i)$ とする．

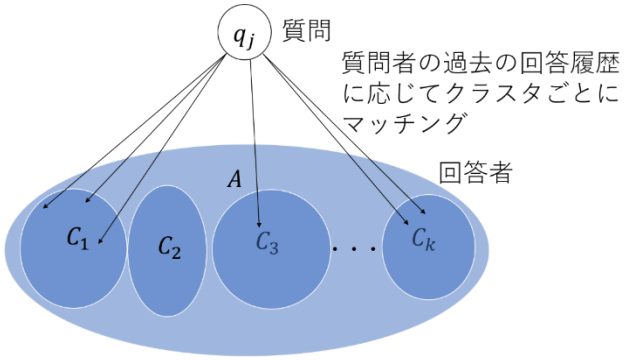

図 3.3　質問マッチングの概要図

Algorithm 1 Calculate $N(C_i)$

Ensure: $N(C_i) = 0$ for all i
  if $n \geq$ sum of $|C_i|$ such i that $f(C_i) > 0$ **then**
    for each cluster **do**
      if $f(C_i) > 0$ **then**
        $N(C_i) \Leftarrow |C_i|$
        $f(C_i) \Leftarrow 0$
      else
        $f(C_i) \Leftarrow 1$
      end if
    end for
  end if
  $start \Leftarrow$ sum of $N(C_i)$ for all i
  for $counter = start$ to $n - 1$ **do**
    select one cluster proportional to the ratio of $f(C_i)$
    N(selected cluster) $\Leftarrow N$(selected cluster) $+ 1$
  end for

クラスタ$C_i\ (i = 1,2, \ldots, k)$ でマッチングする人数を$N(C_i)$とする．$f(C_i)$ に比例した確率でクラスタを一つ決定し，選ばれたクラスタがC_jの場合$N(C_j)$ に 1 を加算する．この操作をn回繰り返し各クラスタでマッチングする人数を決定する．このとき，$f(C_i) > 0$であるクラスタに属する回答者の合計人数がn人未満の場合が起こりうる．この場合，まず，$f(C_i) > 0$であるクラスタに属する回答者全員をマッチング対象とする．そして，$f(C_i) = 0$であるクラスタすべてから等確率で残りの人数分を決定する．この一連の操作を Algorithm 1 に示す．

各クラスタでマッチングする人数が決定したのち，クラスタ$C_i\ (i = 1,2, \ldots, k)$ 内のどの回答者へとマッチングするかは各回答者の回答数に比例した確率で決定する．

4. 実験データ

本研究では実験データとして，Yahoo!知恵袋にて 2008 年 10 月～2009 年 3 月の間に投稿された質問とそれらに対する回答を用いた．2008 年 10 月～2008 年 12 月の期間を学習期間とし，2009 年 1 月～2009 年 3 月のデータをマッチン

表 4.1　学習期間におけるカテゴリごとのユーザー数の構成

カテゴリ	質問者数	分析対象質問者数	回答者数	分析対象回答者数
恋愛相談	36,889	31,311	57,370	6,911
政治、社会問題	12,001	9,768	22,645	2,102

グモデルに適用して評価実験を行った．本研究で用いるデータは情報学研究データリポジトリ[1]から提供されている Yahoo!知恵袋データ（第 2 版）の一部である．このデータは 2004 年から 2009 年までの期間に Yahoo!知恵袋に投稿された質問と回答で構成されている．解決済み（ベストアンサーが決定された）の質問 16,257,413 件と，それらに対する回答 50,053,894 件が含まれている．ベストアンサーとは，一つの質問に対する複数の回答の中から最も適していると判断された一つの回答である．

各質問はその内容に応じて，カテゴリが割り当てられている．本研究では，カテゴリ間の性質の違いがどのように結果に影響するかを調べるために，全 446 個のカテゴリの中から「恋愛相談」，「政治、社会問題」の 2 個のカテゴリを選択した．これは，「恋愛相談」は多くの意見を募ることができるような専門知識を必要としない質問が多いと考えられるのに対し，「政治、社会問題」は専門知識がないと答えられないような質問が多いと考えられるためである．

質問と回答にはそれぞれ投稿者を識別できるユニークな番号が付与されており，同一のユーザーによる投稿を抽出抽出できる．よって，回答者ごとに，どの質問者にどれだけ回答したかのリストを作成することができる．そのリストをもとに，学習期間に投稿された質問と回答について，手法 1 と手法 2 のそれぞれの場合で回答者の特徴ベクトルを作成する．また，手法 2 においては一人の回答者のみから回答されている質問者は除いた．その理由は以下の通りである．例えば，質問者q_jが一人の回答者のみから回答されていた場合を考えると，特徴ベクトルのj番目の要素が0より大きい回答者は一人しかいないことになる．このとき，式 (3.8) に示したコサイン類似度を計算する際に，j番目の要素がすべての回答者の組において0となり類似度を測ることができないため，次元数を削減する目的でそのような質問者を除いた．

各ユーザーは複数のカテゴリで質問または回答を投稿しているが，本研究ではカテゴリごとでの質問と回答に着目して特徴ベクトルを作成し，クラスタリングを行う．また，回答者を絞るために 10 人以上の質問者に対して回答したことのある回答者を対象とした．これは，提案手法では回答が少ない回答者の特徴を抽出するのが難しいからである．学習期間におけるカテゴリごとのユーザー数の構成を表 4.1 に示す．

5. 評価実験

本研究の提案手法が実際に有効であるかどうかを検証するため，第 4 章で説明した実験データを提案手法に適用した結果と，その考察を説明する．

5.1 実験概要

分析対象とした「恋愛相談」，「政治、社会問題」の 2 個のカテゴリにおいて，手法 1 と手法 2 で作成した 2 種類の特徴ベクトルに対してそれぞれk-means 法でクラスタリングを行った．クラスタ数は$2 \leq k \leq 61$のすべての自然数とした．また，マッチングした質問が実際に回答されるかどうかを検証するために，学習期間以降に投稿された質問に，上記のクラスタリングの結果を利用した 3.3 節のマッチングモデルを適用し，評価実験を行った．以下にその手順を説明する．

学習期間である 2008 年 10 月～2008 年 12 月の期間をT_1，マッチングモデルを適用する 2009 年 1 月～2009 年 3 月の期間をT_2とする．このとき，T_1で回答した回答者の集合をA_{T_1}，そのうちクラスタリングの対象となった回答者の集合を$\hat{A}_{T_1}$とする．4.2 節で説明したように，クラスタリングの対象とした回答者を絞ったため，$\hat{A}_{T_1} \subset A_{T_1}$となる．評価期間$T_2$に投稿された質問のうち，$T_1$で質問したことのある質問者が投稿したものについて考える．このとき，$\hat{A}_{T_1} \subset A_{T_1}$であるから，いずれのクラスタにも属さない回答者からのみ回答されていた質問者も存在する．このような質問者からの質問は 3.4 節で説明した過去の回答履歴が存在せず，マッチングモデルを適用できないため実験対象から除いた．回答履歴が存在する質問者の質問であっても，T_2の期間でいずれのクラスタにも属さない回答者からのみ回答されている質問が存在する．このような質問に対しても同様にマッチングモデルを適用することができないため，実験対象から除いた．これらの除外対象以外の質問を回答者へとマッチングし，実際にその回答者が回答しているかどうかを確認することで提案手法の妥当性を検証する．また，以下に示すマッチング手法と比較することで提案手法の性能を評価する．

提案手法と比較する質問のマッチング手法として二つの方法を使用した．一つ目は「ランダム」であり，$\hat{A}_{T_1}$からランダムに回答者を選び，質問をマッチングする．二つ目は「回答数比例」であり，回答数に比例した確率で$\hat{A}_{T_1}$から回答者を選び，マッチングする．これはクラスタ数を 1 とした場合の提案手法と同様の手法となる．なお，T_2ではT_1に現れなかった回答者も回答をしており，提案手法と同様に，比較手法においても，そのような回答者に対して質問をマッチングすることはできない．

5.2 実験結果と考察

クラスタリングの対象となった「恋愛相談」，「政治、社会問題」の 2 個のカテゴリに対して，ランダム，回答数比例，手法 1，手法 2 の 4 つのマッチング手法を適用して実験を行った．

各手法について以下の 2 項目を求めた．

- 通算正解人数：実際に回答していた回答者へマッチングした回数
- 実質正解人数：通算正解人数のうち重複を除いた人数

また，質問ごとのマッチング人数nは 100 に設定した．それぞれの手法ごとに 10 回繰り返して実験を行い，その平均を結果とした．評価期間T_2に投稿された全質問数と実験対象になった質問数を表 5.1 に，実験結果を図 5.1～図 5.4 にそれぞれ示す．

[1] https://www.nii.ac.jp/dsc/idr

表 5.1　質問数の構成

カテゴリ	全質問数	実験対象質問数
恋愛相談	98,417	23,980
政治、社会問題	39,250	10,331

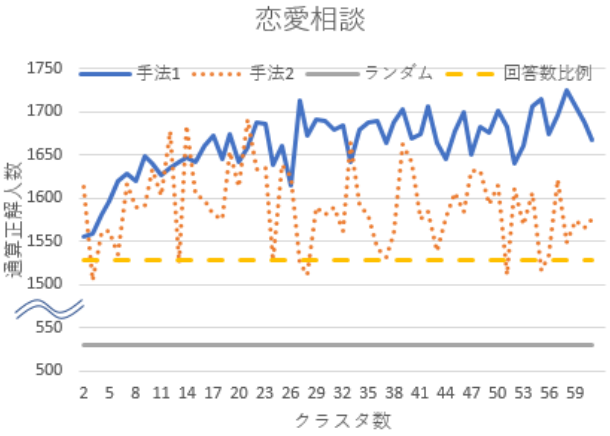

図 5.1　恋愛相談（通算正解人数）

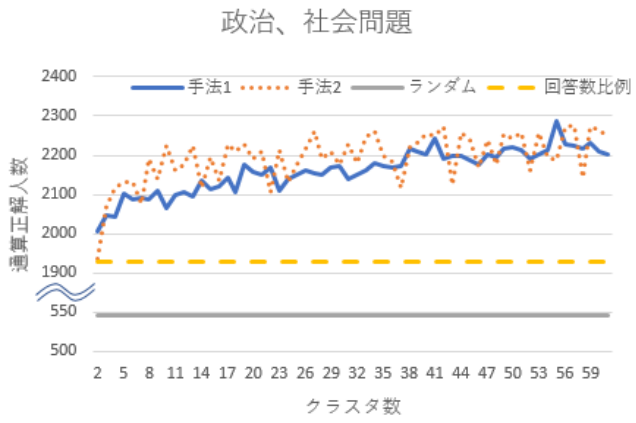

図 5.3　政治、社会問題（通算正解人数）

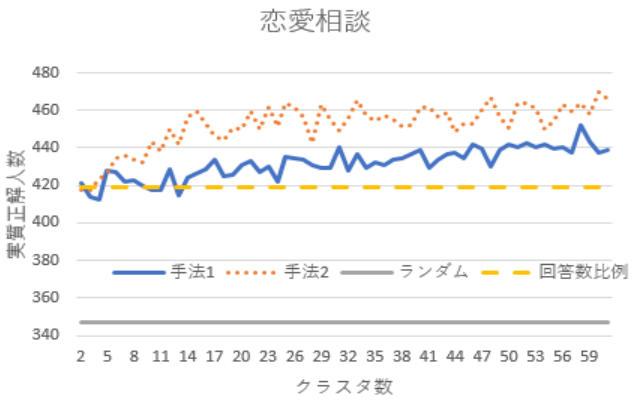

図 5.2　恋愛相談（実質正解人数）

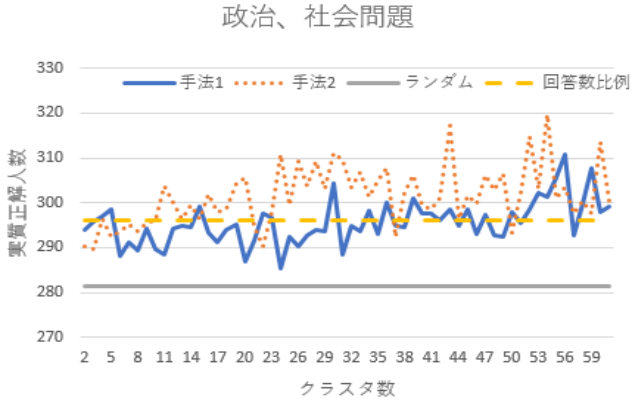

図 5.4　政治、社会問題（実質正解人数）

「恋愛相談」と「政治、社会問題」の両方のカテゴリにおいてランダムはほかの手法と比べて通算正解人数，実質正解人数共に低くなっている．よって，回答数が多い回答者へ優先的にマッチングすることで実際に回答していた回答者を見つける確率が大きくなることがわかる．しかし，実質正解人数は通算正解人数と比べると差が大きくない．そのため，実際に行われた回答が一部の回答者に偏っていることがわかる．

回答数比例と提案手法を比較したとき，概ね提案手法が上回っているが，図 5.4 に示した「政治、社会問題」の実質正解人数はほとんど差がない．しかし，図 5.3 に示した通算正解人数は提案手法が大きく上回っている．「恋愛相談」の方を見てみると，図 5.1，5.2 からわかる通り通算正解人数・実質正解人数共に提案手法が回答数比例を上回っている．「恋愛相談」においては特別な専門知識がなくとも多くの人が回答可能であるような一般的な内容の質問が多いのに対し，「政治、社会問題」は専門知識を必要とする質問が多くなる．したがって「政治、社会問題」では回答できる回答者が限られてくるため，実質正解人数が増えなかったのだと考えられる．つまり，回答数比例では専門知識のない回答者へ多くマッチングしてしまったのに対し，提案手法では一部の専門知識を持った回答者へ優先的にマッチングを行えていたといえる．

また「恋愛相談」においては，前述の通り，専門知識のない回答者でも回答ができる質問が多い．つまり，ほとんどの質問に対して一部の同じ回答者が回答している「政治、社会問題」に比べ，質問ごとに回答者が異なる．そのため，単に回答数の多い回答者に優先的にマッチングしている回答数比例では実質正解人数が少ない．一方，提案手法では質問者ごとに適した回答者を選ぶことにより，効率よく回答者にマッチングすることができる．そのため実質正解人数が多くなったと考えられる．これらのことから，提案手法はクラスタを利用することで，母集団の性質に対応したマッチングを行えているといえる．

図 5.1 に示す通り，「恋愛相談」では手法 1 の方が手法 2 よりも通算正解人数が大きい．式 (3.1) より，手法 1 では一度でも共通の質問者に回答をしていれば値を 1 にするため，回答数の多い回答者同士がクラスタを形成する傾向にある．これにより手法 1 では回答数の多い回答者へマッチングすることが多くなるため，通算正解人数が大きくな

ったと考えられる．しかし図 5.2 に示したように，手法 2 は実質正解人数では手法 1 を上回っている．式 (3.4) に示すように，回答数の割合でベクトルの値を表現したため手法 1 のような回答数の多い回答者同士のクラスタにはなりにくい．そのためマッチングする回答者に偏りが少なくなり，実質正解人数が大きくなったのだと考えられる．

「政治、社会問題」では手法 1・手法 2 の間に大きな差が見られない．これは前述した通り回答者の偏りが顕著なため，どちらの手法でも概ね同じ回答者群をクラスタリングしたからであると考えられる．

また，クラスタ数による影響を考える．手法 2 の正解人数を除いて，クラスタ数の増加に伴い数値が大きくなっている．これはクラスタ数が増えるにつれて質問者に対応する回答者をより精細にマッチングすることができるからだと考えられる．しかし，クラスタ数が増えすぎた場合には，ランダムに近づいてしまうため性能が落ちる．図 5.1 においてクラスタ数 20 以降は手法 2 の通算正解人数が小さくなっていっているのはそのためであると考えられる．

6. まとめと今後の課題

本研究では，Q&A サイトに質問が投稿されたときに，その質問者の履歴に基づいて適切な回答者を選定し，質問をマッチングする手法を提案した．はじめに，回答者の類似度を測る手法を 2 種類定義し，回答者のクラスタリングを行った．その後，クラスタリングの結果を利用した質問マッチングモデルを適用した． Yahoo!知恵袋における実際データを用いたマッチングモデルの評価実験では，提案手法を用いて回答者へマッチングした質問が実際に回答されていることを確認した．また，ランダムに回答者へマッチングする手法と，回答数の多い回答者へ優先的にマッチングする手法の二つと比較し，クラスタを利用した提案手法の優位性を示した．また，回答者の類似度を測る手法の違いにより，通算正解人数と実質正解人数の大小関係が影響を受けることを確認した．

今後の課題としては，ベストアンサーの活用が挙げられる．本研究ではベストアンサーを考慮せず，すべての回答の価値を等しく扱っている．しかし，実際には回答の中には優劣が存在し，質問に対して不適切な回答もされる場合がある．したがって，ベストアンサーに重みを与え，ほかの回答よりも重要な要素として扱うことで，より適切な回答者へと質問をマッチングできる可能性がある．

また，質問内容を考慮することも今後の課題である．本研究では回答者と質問者の繋がりに着目したが，実際には同一質問者からの質問でも内容が大きく異なる場合がある．そのため，質問で求められている知識を持っていない回答者に対してマッチングしてしまう恐れがある．これを解決するには，あらたな回答者の類似度を定義する必要がある．

さらに，本研究ではカテゴリごとでの質問と回答に着目して特徴ベクトルを作成したが，これを複数のカテゴリにおける質問と回答を統合して利用できるように拡張することも考えられる．

謝辞

本研究では，国立情報学研究所がヤフー株式会社から提供を受けて研究者に提供しているデータセット「Yahoo!知恵袋データ (第 2 版)」を利用した．また，本研究の一部は，日本学術振興会における科学研究費補助金基盤研究(B)(課題番号 17H01742)による支援を受けている．ここに記し謝意を表す．

参考文献

[1] 佐藤弘樹，島田論，伏見卓恭， 福原知宏， 斉藤和巳， 佐藤哲司. ”知識共有サイトにおける参加者の貢献度に着目したコミュニティ分析法”. 2010

[2] 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”. In: 研究報告データベー スシステム（DBS）2009.19 (July 2009), pp. 1–8. issn: 09196072. url: https://ci.nii. ac.jp/naid/10026776651/.

[3] 岩間雄太，伊藤孝行，佐藤元紀. “参照構造を利用した質問応答システムの実装と評価”. In: 人工知能学会全国大会論文集 28 (2014), pp. 1–4. issn: 1347-9881. url: https://ci.nii.ac.jp/naid/40020077006/.

[4] 片山亮，鈴木恵，二村秀憲. “QA サイトにおける質問推薦へ向けた履歴データの分析”. In: 電子情報通信学会技術研究報告. AI, 人工知能と知識処理 109.439 (Feb. 2010), pp. 11–16. issn: 09135685. url: https://ci.nii.ac.jp/naid/110008000250/.

[5] 甲谷優，川島晴美，藤村考. “QA コミュニティの成長パターンに基づく回答者への質問推 薦”. In: 日本データベース学会論文誌 8.1 (June 2009), pp. 89–94. issn: 18831060. url: https://ci.nii.ac.jp/naid/40016752604/.

[6] 石井健一郎，上田修功. 続・わかりやすいパターン認識 教師なし学習入門. オーム社, 2014. isbn: 9784274062568