

Q&A サイトにおける質問と回答の分析 (3) — 質問・回答履歴を用いたベストアンサー推定 —

栗山和子^{†1} 神門典子^{†2}

本研究では、Q&A サイトにおける利用者の質問・回答履歴が、質問者がベストアンサーを選択するときに何らかの影響を与えているかどうかを調べ、複数の回答群から自動的にベストアンサーを推定するときに利用可能であるかどうかについて考察する。

まず、Q&A サイトに登録している利用者の質問・回答履歴を分析することにより、頻繁にサイトを使用している利用者およびベストアンサーに数多く選ばれている回答者の質問・回答行動の特徴について考察した。また、複数の回答群から自動的にベストアンサーを推定するとき、質問・回答履歴のみを使用した推定が可能かどうか予備的な実験を行い、検証した。

結果として、利用者の質問・回答履歴にはカテゴリごと・利用者ごとの特徴があり、それを利用することによって、ある程度ベストアンサーを推定することが可能であることがわかった。

Analysis of Questions and Answers in Q&A Site (3) — Predicting Best-Answers Based on Users' Histories —

KAZUKO KURIYAMA^{†1} and NORIKO KANDO^{†2}

In this paper, we analyzed questions and answers in Q&A site to investigate whether users' questioning and answering histories affected questioners' selection of the best answer for their own questions, and to examine whether the histories are available to predict the best answers automatically.

First, we extracted some features from the histories of the users who often used the Q&A site and whose answers were frequently selected as the best answers. Then, we carried out a preliminary experiment to examine whether it is able to select the best answers from multiple candidates automatically only using the features of the users' questioning and answering histories.

Finally, we found that the users' histories differed from categories and types of users and the features of them could be used for predicting the best answers effectively.

1. はじめに

近年、Q&A サイトと呼ばれる質問回答サイトの利用者が増加している^{13),16)}。Q&A サイトの仕組みは、利用者が匿名（ユーザ ID）で質問を投稿し、その利用者とは別の利用者が回答を投稿するというものである。利用者が質問・回答を行う理由は様々であるが、その行動は利用者の自発的なものであり、必ずしも、質問に対して質問者の目的に沿った回答が得られるとは限らず、また、回答者が客観的に見て適切な回答をしたとしても質問者がその回答をベストアンサーとして選択するとは限らない。ベストアンサーとは、一般的に、質問者自身による選択、または他の利用者からの投票で選ばれるもので、質問者によって、あるいは投票結果によって、最も高く評価された回答のことである。

Kim ら^{7),8)}によれば、Q&A サイトの質問者は、他の利用者からの回答を評価するとき、これまで適合性 (relevance)¹⁴⁾ を評価する基準として重視されてきた「内容 (Content)」や「有用性 (Utility)」だけでなく、「社会的・感情的 (Socio-emotional)」な面も同様に重視している。すなわち、質問者は、ベストアンサーを内容の適切性や有用性だけで決めているわけではなく、回答の印象や回答者の態度もベストアンサーの選択に影響を与えていると言える。また、三浦ら¹²⁾によれば、質問者の6割以上は、質問する前に他の手段を用いて質問の内容についての調査を行っているが、回答者の半数以上は、回答する前に質問について調査を行っていない。

本稿では、「Yahoo!知恵袋」の研究機関提供用データ^{16),17)}を使用し、このサイトに登録している利用者の質問・回答履歴を分析することにより、質問・回答行動の特徴について調べ、過去の利用行動が質問者によるベストアンサーの選択や回答者による回答する質問の選択と関連があるかどうか考察し、その特徴を自動的なベストアンサーの推定に利用できるかどうかについて検証する。

2 節では、Q&A サイトの利用者間や質問・回答間の関係を質問・回答の推薦やベストアンサー推定に利用している関連文献について簡単に述べる。3 節では、質問・回答履歴を分析し、Q&A サイトを頻繁に利用している利用者やベストアンサーを多数獲得している利用

^{†1} 白百合女子大学
Shirayuri College

^{†2} 国立情報学研究所
National Institute of Informatics

者の質問・回答行動の特徴を示す。4節では、利用者の質問・回答履歴をベストアンサー推定に利用することが可能かどうか検証するため、予備的な実験を行い、その結果について考察する。5節では、本稿のまとめを述べ、本研究の今後の予定と課題を示す。

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。本節では、Q&A サイトの利用者間あるいは質問・回答間の関係を分析し、質問・回答の推薦やベストアンサー推定に利用している関連文献について簡単に述べる。

Jurczyk ら⁴⁾ は、質問、回答、質問者、回答者間のリンクを表すグラフ構造に着目し、質問者をハブ (hubs)、回答者をオーソリティ (authorities) として、Web ページのハブとオーソリティのスコアを計算する HITS アルゴリズム⁹⁾ を適用し、重要度の高い回答者を発見する方法を提案した。

Agichtein ら²⁾ は、質問・回答のテキストおよび内容の質、グラフとして表した利用者間の関係、質問の閲覧回数 (ビュー) など様々な素性を用いた機械学習によって、質の高い質問、回答、利用者を識別する方法を提案した。また、この論文では、質問と回答の質には関連があり、質の高い質問には質の高い回答が付くことが示されている。この論文の継続研究として、Bian ら³⁾ は、利用者、質問、回答間の相互関係 (interaction) を合成 2 部グラフとみなし、半教師あり機械学習を用いて、質の高い質問、回答、利用者を識別する方法を提案した。

Liu ら¹¹⁾ は、回答に対する質問者の満足度 (satisfaction) を予測するために、質問と回答に関する複数の素性を抽出し、決定木、SVM、ブースティング、ナイーブベイズの 4 つの分類アルゴリズムを用いて、質問者が満足するかどうかを予測した。使用された素性は様々であるが、質問者・回答者の利用履歴としては、質問・回答回数、登録期間、質問と回答の割合、ベストアンサー獲得回数、ベストアンサー率、獲得投票数などを用いている。この論文では、予測の精度は、質問者の質問回数が増えるほど、つまり、質問者の利用履歴が多く蓄積されているほど上がるので、利用履歴を使用することは有効であると述べられている。

甲谷ら⁵⁾ は、Adamic らの研究¹⁾ に基づき、利用者をノード、スレッド内における各回答を回答者から質問者へのエッジとして、QA ネットワークと呼ばれる質問応答グラフを作成し、その成長パターンを分析することによって、各カテゴリのコミュニケーションタイプを分類した。また、甲谷ら⁶⁾ は、QA ネットワークにおける成長パターンに基づき、局所構造における成長に関する制約を用いてリンク予測を行い、ある時刻から時刻の間に新たに回答

が発生しそうな 2 人の利用者 (新しく発生しそうなエッジ) を発見する手法を提案し、この手法が、個々の回答者に回答するのに適切な質問者の質問を推薦するのに有効である可能性を示唆した。

Wang ら¹⁵⁾ は、質問と質の高い回答の間の関係を正リンク (positive links)、それ以外の質問と回答の間の関係を負リンク (negative links) とし、ベイジアン・ロジスティック回帰モデル (Bayesian logistic regression model) を用いて機械学習を行い、新しい質問をクエリとして正リンクを検索し、それを supporting set として使用して、新しい質問・回答ペアのスコアを計算し、ベストアンサーを推定する、Bayesian Analogical Reasoning の手法を提案した。

以上のように、Q&A サイトに関する研究において、の利用者間あるいは質問・回答間の関係は注目されているが、リンク関係やそのグラフ構造に着目して機械学習を用いてベストアンサーを総合的に推定するものが主であり、カテゴリや質問・回答のタイプの違いによる利用者ごとの特徴についてはほとんど検討されていない。そこで、本研究では、利用者をいくつかのグループに分けて質問・回答履歴を分析し、その特徴をベストアンサー推定に利用することができるかどうか検討する。

3. 質問・回答履歴の分析

3.1 使用データの概要

本研究では、Q&A サイトのデータとして「Yahoo!知恵袋 研究機関提供用データ 国立情報学研究所 (NII) 提供版 ver.1」¹⁷⁾ (以下、知恵袋データ) を使用する。知恵袋データは、2004 年 4 月 1 日から 2005 年 10 月 31 日に「Yahoo!知恵袋」¹⁶⁾ ベータ版に投稿された質問、回答、および、それらに付随するデータから抽出されたもので、解決済みの質問 3,116,009 件、質問者が選んだベストアンサー 3,116,008 件、その他の回答 10,361,777 件から成る。「Yahoo!知恵袋」におけるベストアンサー (以下、「BA」) とは、質問者によって選ばれた、あるいは他の利用者から投票によって選ばれた、「もっとも納得、満足した回答」¹⁶⁾ のことである。1 つの質問には、1 つの BA と 0 個以上のその他の回答が存在する。

利用者 ID は、本来の Yahoo!JAPAN ID とは異なる ID 番号に変換されているが、同じ Yahoo!JAPAN ID は同じ利用者 ID に変換されているため、変換後の利用者 ID により利用者ごととの質問・回答を識別することは可能である。ただし、一人の利用者が一つの Yahoo!JAPAN ID しか持っていないかどうかはこのデータからはわからないので、異なる利用者 ID が全て現実の異なる利用者に割り当てられたものであるかどうかは保証されない。

「Yahoo!知恵袋」では、質問はトピックのカテゴリごとに分けられ、質問者が質問投稿時にカテゴリを選択するようになっている。現在のカテゴリは、カテゴリ内の質問と回答を含んだまま、ベータ版以降何度か再構成されて数も増えており、知恵袋データが構築された時点のカテゴリから名称や階層構造が一部修正、追加、削除されている。知恵袋データでは、ベータ版当時のカテゴリ構造のデータは提供されていないため、本研究の分析では、基本的には、カテゴリ間の階層関係はないものとして扱う。

本研究では、カテゴリごとの差異を見るため、以前の論文¹⁰⁾で質問タイプの分類に使用した5つのカテゴリ、「数学、サイエンス」、「恋愛、人間関係の悩み」、「パソコン、周辺機器」、「料理、グルメ、レシピ」「健康、病気、ダイエット」のうち、「情報検索型」（サーチエンジンや質問応答システムによる情報検索や図書館のレファレンス・サービスなどを利用すれば適切で客観的な回答が得られるタイプ）の質問が8割以上であった「パソコン、周辺機器」、「社会調査型」（適切な回答の基準がなくQ&Aサイトに参加している人に個人的な意見・嗜好・経験などを尋ねることによって質問者が主観的に適切な回答や結論を決定するアンケート調査のようなタイプ）の質問が9割以上であった「恋愛、人間関係の悩み」、「情報検索型」と「社会調査型」の質問がほぼ半々であった「料理、グルメ、レシピ」の3つのカテゴリの質問と回答のデータを使用する。

3.2 質問・回答履歴についての統計

まず、利用者数、質問者数、回答者数、BA獲得者数を表1に示す。利用者数は異なる利用者IDの個数である。質問者数、回答者数、BA獲得者数は、それぞれ、総数（1回以上質問/回答/BAを獲得した利用者的人数）、質問回数/回答回数/BA獲得回数が0回、1回、2～9回、10～79回、80回以上である利用者的人数である。総数の後ろの括弧内は利用者数に対する割合（%）である。

ここでは、1回の利用者をミニマムユーザ、2～9回以下の利用者をライトユーザ、10回～79回以上の利用者をミドルユーザ、80回以上の利用者をヘビーユーザとし、利用回数ごとに5つのグループに分けて、人数を数えた。ヘビーユーザを80回以上とした理由は、データの収録期間が79週間（78週間と2日）であるので、1週間に1回よりも多く利用した利用者を頻繁に利用したものと考えたからである。

表1の0回の列の人数から、利用者の中には、質問だけをしている利用者と回答だけをしている利用者が存在することがわかる。それぞれの利用者全体に対する割合は、質問だけ24.4%、回答だけ31.9%であるが、このうち、質問だけをしていて1回しか質問をしていない利用者は35,497人（60.1%）、回答だけをしていて1回しか回答をしていない利用者は

表1 利用者数、質問者数、回答者数、BA獲得者数

	総数	0回	1回	2～9回	10～79回	80回以上
利用者数	242,333					
質問者数	165,064 (68.1)	77,269	54,784	65,416	37,070	7,794
回答者数	183,242 (75.6)	59,091	53,671	63,824	42,353	23,394
BA獲得者数	119,263 (49.2)	123,070	39,599	45,098	26,590	7,976

は38,504人（49.8%）である（括弧内は、それぞれ、質問だけ、回答だけの人数に対する割合）。すなわち、質問または回答のどちらかしかしていない利用者であっても、質問だけの利用者の約4割、回答だけの利用者の約5割が複数回の利用をしている。このことから、継続的な利用者の中にも、質問するためだけ、または、回答するためだけにQ&Aサイトを利用している利用者がある程度存在することがわかる。

1回以上質問している質問者の質問回数、1回以上回答している回答者の回答回数、回答者のBA獲得回数、回答者のBA率、それぞれの平均値(ave)、中央値(med)、最小値(min)、最大値(max)を表2に示す。BA率は、利用者ごとの回答回数に対するBAの割合（BA獲得回数/回答回数）である。また、質問回数と回答回数の関連を調べるために、質問回数と回答回数についてのクロス表（人数）を表3に示す。後ろの括弧内は、利用者数の総数に対する割合（%）である。5%以上の人数を太字で表す。

表2 質問回数、回答回数、BA獲得回数、BA率の統計

回数	ave	med	min	max
質問回数	18	3	1	6,169
回答回数	73	4	1	30,979
BA獲得回数	17	1	0	7,780
BA率	0.27	0.19	0	100

表2、表3からわかるように、質問あるいは回答の一般的な回数（中央値）は3～4回であり、質問・回答とも10回未満である利用者が多く、利用者全体の68.3%を占める。一方、質問・回答とも10回以上である利用者は13.9%と少ないが、質問を10回以上している利用者の割合は27.1%、回答だけを10回以上している利用者の割合は18.5%であり、全体の利用回数の平均値は、そのようなミドル/ヘビーユーザによって引き上げられている。つまり、ミドル/ヘビーユーザの中には、主に質問または回答のどちらかだけの利用を行っている利用者が存在していると考えられる。

表 3 質問回数と回答回数のクロス表

質問回数 \ 回答回数	0 回	1 回	2~9 回	10~79 回	80 回以上
0 回	0 (0.0)	38,504 (15.9)	29,242 (12.1)	7,797 (3.2)	1,726 (0.7)
1 回	35,497 (14.6)	5,954 (2.5)	8,863 (3.7)	3,686 (1.5)	784 (0.3)
2~9 回	21,177 (8.7)	7,699 (3.2)	18,457 (7.6)	13,938 (5.8)	4,145 (1.7)
10~79 回	2,362 (1.0)	1,470 (0.6)	7,034 (2.9)	15,295 (6.3)	10,909 (4.5)
80 回以上	55 (0.0)	44 (0.0)	228 (0.1)	1,637 (0.7)	5,830 (2.4)

以上のことから、利用者には、サイト利用が少ないミニマム/ライトユーザと日常的にサイトを利用しているミドル/ヘビーユーザがおり、利用者全体に占める割合は、ミニマム/ライトユーザが高いが、ミドル/ヘビーユーザの中には、主に、質問と回答のどちらかだけを頻繁に行う者がいると考えられる。参考のために、質問回数と回答回数の相関係数を計算したが、0.26 と低く、質問回数と回答回数そのものには相関はない。

BA 率については、回答者間の BA 率のばらつきは大きく、ミニマムユーザがたまたま BA を獲得して BA 率が 1 になることもあれば、ヘビーユーザであるにも関わらず、非常に BA 率が低い場合もある。極端な例では、391 回も回答しているのに、1 回も BA に選ばれていない利用者もいる。

次に、利用者ごとの回答回数と BA 率との関係を調べるため、回答回数ごとの BA 率の平均値、中央値、最小値、最大値を表 4 に、回答回数と BA 率のクロス表（人数と質問者全体に対する割合 (%)）を表 5 に、回答回数を x 軸、BA 率を y 軸としたグラフを図 1 に示す。表 5 では、5%以上の人数を太字で表す。

表 4 回答回数ごとの BA 率

回答回数	BA 率			
	ave	med	min	max
1 回	0.29	0	0	1
2~9 回	0.27	0.25	0.01	1
10~79 回	0.26	0.23	0	1
80 回以上	0.24	0.22	0	0.92

表 4、表 5、図 1 から、平均的な BA 率は 0.2 程度だが、回答回数が増えるにしたがって、BA 率は下がっていくことがわかる。このことから、回答経験が豊富であったとしても、必ずしも、それに応じて BA 率が高くなるわけではないことがわかるが、回答回数が多いにも関わらず、BA 率を高く保っているヘビーユーザも少数ながら存在する。次項では、その

表 5 回答回数と BA 率のクロス表

BA 率 (r)	0	0 < r ≤ 0.2	0.2 < r ≤ 0.4	0.4 < r ≤ 0.6	0.6 < r ≤ 0.8	0.8 < r ≤ 1.0
1 回	38,217 (20.9)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	15,454 (8.4)
2~9 回	24,420 (13.3)	6,382 (3.5)	13,960 (7.6)	12,456 (6.8)	3,682 (2.0)	2,924 (1.6)
10~79 回	1,340 (0.7)	16,571 (9.0)	18,219 (9.9)	5,052 (2.8)	1,016 (0.6)	155 (0.1)
80 回以上	2 (0.0)	10,039 (5.5)	11,232 (6.1)	1,885 (1.1)	223 (0.1)	13 (0.0)

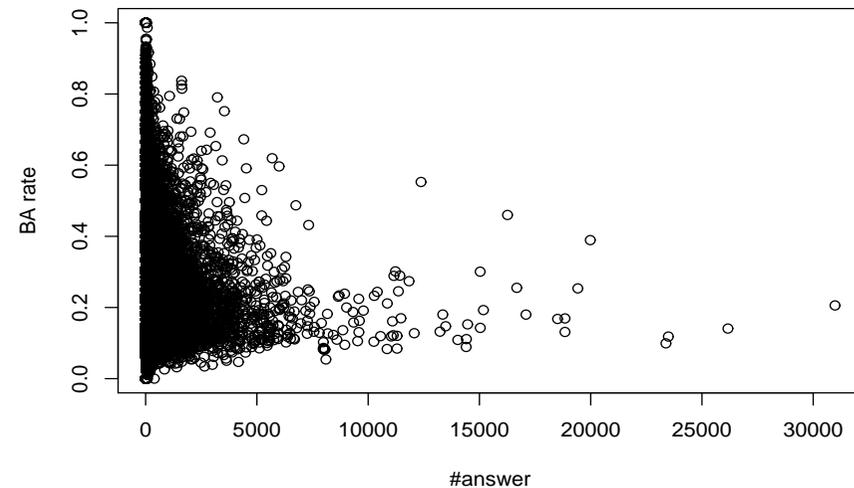


図 1 回答回数と BA 率

ようなヘビーユーザの特徴について考える。

3.3 質問者・回答者の専門性

質問者・回答者の分野専門性について検討するため、「恋愛、人間関係の悩み」(以下, LH), 「パソコン、周辺機器」(以下, PE), 「料理、グルメ、レシピ」(以下, FG) の3つのカテゴリに質問・回答を投稿している利用者の質問・回答履歴について考える。まず, カテゴリごとの利用者数, 質問者数, 回答者数, BA 獲得者数を表 6 に示す。総数の後ろの括弧内のイタリックはカテゴリごとの利用者数に対する割合(%), 質問者数, 回答者数, BA 獲得者数の回数ごとの人数の後ろの括弧内は, それぞれ, 質問者数, 回答者数, BA 獲得者数に対する割合(%)である。

表 6 カテゴリごとの利用者数, 質問者数, 回答者数, BA 獲得者数

カテゴリ		総数	1 回	2~9 回	10~79 回	80 回以上
LH	利用者数	68,199				
	質問者数	24,093 (35.3)	9,859 (40.9)	10,672 (44.3)	3,161 (13.1)	401 (1.7)
	回答者数	62,717 (92.0)	19,905 (31.7)	25,158 (40.1)	14,247 (22.7)	3,407 (5.4)
	BA 獲得者数	30,867 (45.3)	13,038 (42.2)	13,435 (43.5)	4,079 (13.2)	315 (1.0)
PE	利用者数	58,102				
	質問者数	43,493 (74.9)	21,114 (48.5)	18,940 (43.5)	3,343 (7.7)	96 (0.2)
	回答者数	27,420 (47.2)	12,335 (48.5)	10,429 (41.0)	1,708 (6.7)	948 (3.7)
	BA 獲得者数	14,550 (25.0)	7,127 (49.0)	5,350 (36.8)	1,721 (11.8)	352 (2.4)
FG	利用者数	29,264				
	質問者数	11,959 (40.9)	6,871 (57.5)	4,407 (36.9)	661 (5.5)	20 (0.2)
	回答者数	23,286 (79.6)	10,764 (46.2)	9,544 (41.0)	2,708 (11.6)	270 (1.2)
	BA 獲得者数	10,125 (34.6)	5,783 (57.1)	3,673 (36.3)	639 (6.3)	30 (0.3)

表 6 から, カテゴリごとに質問者と回答者の割合が異なることがわかる。社会調査型の質問が多く, 内容が比較的身近で回答しやすい「恋愛、人間関係の悩み」では, そのカテゴリの利用者の 9 割以上が回答者として回答したことがあるのに対し, 情報検索型の質問が多く, 回答に対してある程度の専門知識を必要とするような「パソコン、周辺機器」では, 利用者の約半数しか回答をしていない。また, BA 獲得者についても, 「恋愛、人間関係の悩み」では, 半数近くの利用者が回答を BA として選択されたことがあるのに対し, 「パソコン、周辺機器」では, 1/4 しか BA として選択されたことがない。「料理、グルメ、レシピ」は, 回答者についても BA 獲得者についても両方のカテゴリの中間くらいである。いずれのカテゴリの質問回数・回答回数でも, ミニマム/ライトユーザが 7~9 割を占めるが, 「恋

愛、人間関係の悩み」では, 回答回数が 10 回以上であるミドル/ヘビーユーザの割合が約 3 割を占めているのに対し, 他の 2 つのカテゴリでは, 1 割程度である。以上のことから, 質問についての専門知識を持たなくても答えられるようなカテゴリでは, 利用者の回答回数は増えるが, 専門知識や情報を持っていないと答えにくい質問が多いようなカテゴリでは, 回答回数が比較的少なくなり, BA 獲得者の全体の利用者に対する割合も低くなると考えられる。

各カテゴリの回答者の分野専門性を調べるため, 各カテゴリの利用者のうち, 回答回数 2~9 回のライトユーザと 80 回以上のヘビーユーザで, BA 率が (a)0~0.4 以下と (b)0.6 以上 1.0 以下の回答者の利用カテゴリの上位 5 つを表 7 に示す。ただし, 集計は回答者ごとに各カテゴリの割合を計算し, 各グループの回答者の割合の合計で順位の計算を行った。空欄になっているところは, 条件に該当する回答者が存在しなかったところである。比較のために, 質問件数の多い順にカテゴリ, 質問件数, 質問件数全体に占める割合(%)を表 8 に示す。

表 7 カテゴリごとの利用者の主要な利用カテゴリ

カテゴリ	ライトユーザ (2~9 回)		ヘビーユーザ (80 回以上)	
	(a) $0 \leq r \leq 0.4$	(b) $0.6 \leq r \leq 1.0$	(a) $0 \leq r \leq 0.4$	(b) $0.6 \leq r \leq 1.0$
LH	1	恋愛相談、人間関係の悩み	恋愛相談、人間関係の悩み	恋愛相談、人間関係の悩み
	2	Yahoo!オークション	Yahoo!オークション	Yahoo!知恵袋
	3	Yahoo!知恵袋	Yahoo!知恵袋	Yahoo!オークション
	4	アダルト	パソコン、周辺機器	政治、社会問題
	5	政治、社会問題	病気、症状、ヘルスケア	パソコン、周辺機器
PE	1	パソコン、周辺機器	パソコン、周辺機器	パソコン、周辺機器
	2	恋愛相談、人間関係の悩み	恋愛相談、人間関係の悩み	Yahoo!オークション
	3	Yahoo!知恵袋	Yahoo!オークション	Yahoo!知恵袋
	4	Yahoo!オークション	Yahoo!知恵袋	恋愛相談、人間関係の悩み
	5	政治、社会問題	インターネット	インターネット
FG	1	恋愛相談、人間関係の悩み	恋愛相談、人間関係の悩み	レシピ、調理法
	2	Yahoo!知恵袋	レシピ、調理法	Yahoo!知恵袋
	3	Yahoo!オークション	Yahoo!オークション	恋愛相談、人間関係の悩み
	4	レシピ、調理法	Yahoo!知恵袋	Yahoo!オークション
	5	料理、グルメ、レシピ	動物、植物、ペット	国内

表 7 から, 「恋愛、人間関係の悩み」と「パソコン、周辺機器」では, 回答回数や BA 率にかかわらず, そのカテゴリが回答者の主要回答カテゴリの一番目に来ている。「恋愛、人

表 8 質問件数の多いカテゴリ

順位	カテゴリ	質問件数	%	順位	カテゴリ	質問件数	%
1	恋愛相談、人間関係の悩み	210,105	6.7	11	国内	61,380	2.0
2	Yahoo!知恵袋	208,793	6.7	12	アダルト	61,369	2.0
3	Yahoo!オークション	190,432	6.1	13	自動車	57,627	1.8
4	パソコン、周辺機器	171,848	5.5	14	芸能人、タレント	55,783	1.8
5	病気、症状、ヘルスケア	84,394	2.7	15	野球	53,341	1.7
6	政治、社会問題	78,777	2.5	16	レシピ、調理法	52,430	1.7
7	テレビ、ラジオ	75,194	2.4	17	一般教養	51,940	1.7
8	インターネット	70,529	2.3	18	ファッション	48,984	1.6
9	言葉、語学	68,513	2.2	19	メンタルヘルス	47,807	1.5
10	動物、植物、ペット	61,541	2.0	20	健康、病気、ダイエット	41,978	1.3

間関係の悩み」,「Yahoo!オークション」,「Yahoo!知恵袋」は、回答回数や BA 率によらず、どのカテゴリにおいても回答者の回答が多いが、表 8 からわかるように、この 3 つのカテゴリは質問件数が多い順に 1~3 位のカテゴリであり、その影響もあるのではないかとと思われる。この 3 カテゴリを除くと、「パソコン、周辺機器」では、ライトユーザの (b)、ヘビーユーザの (a),(b) のいずれでも「インターネット」という「パソコン、周辺機器」と類似するカテゴリが含まれている。また、「料理、グルメ、レシピ」では、ライトユーザの (a) にしか、当該カテゴリが含まれていないが、ライトユーザ (b) の 2 位とヘビーユーザ (a),(b) の 1 位にその下位カテゴリに当たる「レシピ、調理法」が入っている。これらのことから、「パソコン、周辺機器」や「料理、グルメ、レシピ」のように、回答するためにその分野の知識を必要とするカテゴリの回答者は、そのカテゴリで回答した回数が多く、特に、BA 率が高い回答者は、関連カテゴリでも回答することが多く、そのカテゴリおよびその関連カテゴリについて関心や専門知識を持っているのではないかと考えられる。一方、「恋愛、人間関係の悩み」のような専門知識があまり必要ない一般的なカテゴリの利用者は、一般的に人気のある (利用者の多い) カテゴリをよく利用していると思われる。

4. ベストアンサーの推定

4.1 使用データと特徴

利用者の質問・回答履歴を BA の推定に用いるためには、ある程度の期間、蓄積された履歴が必要である。本研究で使用する知恵袋データの期間は 19ヶ月 (2004年4月1日から2005年10月31日) と短いので、履歴を抽出する期間を 2004年4月1日から2005年9月30日までの18ヶ月とし、BAの推定を行う質問の期間を2005年10月の1ヶ月間とする。

ただし、9月末に投稿された質問の中には、9月から10月にまたがって回答が投稿されているものがあるので、知恵袋データの2005年10月分の質問データの一番目の質問の質問番号「639760」より前の番号の質問とその回答をそれぞれ「質問データ1」,「回答データ1」とし、利用者の履歴抽出に用いる。そして、質問番号「639760」以降の質問を「質問データ2」とし、そのベストアンサーを推定する。「質問データ1」の質問件数は2923826件、「質問データ2」の質問件数は192183件であり、それぞれ、全質問件数の93.8%, 6.2%にあたる。

ある1件の質問を除いて、全ての質問データには対応するBAが存在するので、質問に対する回答件数が1の場合にはBAを推定する必要はない。質問1件あたりの平均回答件数は4.3件である¹⁰⁾であるので、本稿では、平均的な回答件数の質問を実験対象とするために、回答件数が2~5件の質問のみをBA推定の対象とする。

推定には、以下の値(重み)をそれぞれ計算し、いくつかのパターンで単純に合計したものをスコアとして使用する。

- (w1) 回答順序：質問に対する回答の中での回答順序
(回答順序ごとのBAである確率)
- (w2) カテゴリ重み：回答者のカテゴリごとの回答率
(そのカテゴリでの回答回数/全回答回数)
- (w3) BA率：回答者のカテゴリごとのBA率
(そのカテゴリでのBA獲得回数/そのカテゴリでの回答回数)
- (w4) 相性：回答者が質問者の質問に回答してBAを獲得した割合
(質問者からBAとして評価された回数/質問者に対する回答回数)

w1~w4のそれぞれの値の範囲は0~1である。

回答順序は回答者の履歴そのものではないが、質問者あるいは他の回答者との関係として重要であると考えるので、特徴として使用する。回答順序は、回答日付ではなく、回答番号から決定するものとし、回答番号が小さいほど回答の順序(投稿時刻)が早いものとみなす。

質問データ1, 回答データ1から、データ全体、カテゴリ「恋愛、人間関係の悩み」,「パソコン、周辺機器」,「料理、グルメ、レシピ」の回答件数ごとの質問件数の割合を表9に、回答件数2~5件の質問における回答件数ごとのBAの回答順序の割合を表10に示す。

表9から、カテゴリによって質問1件あたりの回答件数の傾向が異なることがわかる。「恋愛、人間関係の悩み」の質問の半数以上は回答件数が6件以上であるのに対し、「パソコン、周辺機器」の質問の9割以上は、回答件数5件以下である。しかし、表10からわかるように、回答件数および回答順序ごとのBAの割合は、カテゴリによらず、ほぼ同程度である。

表 9 回答件数ごとの質問件数

回答件数	全体		LH		PE		FG	
	件数	(割合)	件数	(割合)	件数	(割合)	件数	(割合)
	2,923,826		195,996		160,983		35,287	
1	499,945	(17.1)	7,169	(3.7)	44,039	(27.4)	5,663	(16.0)
2	530,689	(18.2)	13,970	(7.1)	42,923	(26.7)	6,268	(17.8)
3	470,348	(16.1)	20,044	(10.2)	31,274	(19.4)	5,795	(16.4)
4	368,805	(12.6)	23,253	(11.9)	19,214	(11.9)	4,628	(13.1)
5	277,813	(9.5)	23,382	(11.9)	10,586	(6.6)	3,540	(10.0)
6 以上	776,226	(26.5)	108,178	(55.2)	12,947	(8.0)	9,393	(26.6)

表 10 BA の回答順序の割合

回答件数	回答順序	全体		LH		PE		FG		推定
2	1	205,253	(0.387)	5,709	(0.409)	16,722	(0.39)	2,517	(0.402)	0.4
	2	325,436	(0.613)	8,261	(0.591)	26,201	(0.61)	3,751	(0.598)	0.6
3	1	107,243	(0.228)	4,770	(0.238)	7,181	(0.23)	1,348	(0.233)	0.23
	2	143,711	(0.306)	6,195	(0.309)	9,551	(0.305)	1,872	(0.323)	0.31
	3	219,394	(0.466)	9,079	(0.453)	14,542	(0.465)	2,575	(0.444)	0.46
4	1	58,746	(0.159)	3,752	(0.161)	3,056	(0.159)	787	(0.17)	0.16
	2	72,961	(0.198)	4,775	(0.205)	3,681	(0.192)	946	(0.204)	0.2
	3	97,746	(0.265)	6,151	(0.265)	4,991	(0.26)	1,245	(0.269)	0.265
	4	139,352	(0.378)	8,575	(0.369)	7,486	(0.39)	1,650	(0.357)	0.375
5	1	33,781	(0.122)	2,720	(0.116)	1,298	(0.123)	473	(0.134)	0.12
	2	40,314	(0.145)	3,369	(0.144)	1,478	(0.14)	520	(0.147)	0.145
	3	50,839	(0.183)	4,366	(0.187)	1,900	(0.179)	647	(0.183)	0.185
	4	63,990	(0.23)	5,358	(0.229)	2,519	(0.238)	833	(0.235)	0.235
	5	88,888	(0.32)	7,569	(0.324)	3,391	(0.32)	1,067	(0.301)	0.315

このことから、(w1) 回答順序の重みについては、表 10 の右端の列のように設定する。その他の重みについても、質問データ 1 と回答データ 1 を利用して計算する。

4.2 実験と結果

質問データ 2 の全質問について、質問データ 1 と回答データ 1 から重みを計算し、いくつかのパターンで重みを合計したスコアを用いて、各パターンでスコアが一番高いものを BA と推定したときの精度を表 11 に示す。精度は、BA 推定が正解した質問件数/全質問件数として計算した。比較のためのベースラインは、ランダムに BA を選択した場合とし、質問ごとにランダムな選択を 30 回試行したときの精度の平均を用いる。質問データ 2 に含まれるカテゴリごとの質問数は、「恋愛、人間関係の悩み」6777 件、「パソコン、周辺機器」6746 件、「料理、グルメ、レシピ」1465 件である。

表 11 スコアごとの精度

カテゴリ	w1	w2	w3	w4	w1+w2+w3+w4	w1+w3+w4	w1+w3	ランダム
LH	0.404	0.323	0.385	0.301	0.349	0.402	0.406	0.262
PE	0.504	0.458	0.503	0.423	0.483	0.510	0.514	0.308
FG	0.416	0.401	0.431	0.371	0.437	0.435	0.435	0.299

表 11 からわかるように、どのスコアを使用しても、ランダムに選択するよりは、精度が高い。最も精度が高いのは w1+w3 で、回答順序と利用者の BA 率が、BA の推定に有効であることがわかる。カテゴリごとに見たとき、「パソコン、周辺機器」、「料理、グルメ、レシピ」、「恋愛、人間関係の悩み」の順に精度が高くなっているが、この順序は、1 質問あたりの回答件数が少ない順と一致している（この順に、各カテゴリの回答件数の平均は 2.8 件、4.3 件、6.7 件となっている¹⁰⁾）。これは、回答件数が多くなるにしたがって、質問者のベストアンサーの選択基準が回答の内容を重視したものになり、回答順序や回答者のそのカテゴリでの回答回数（カテゴリ重み）やカテゴリごとの BA 率などのカテゴリ専門性のような直接的でない要素の影響が弱くなるからではないかと考えられるが、回答の内容を用いた BA 推定と比較しないと、明確なことは言えない。

5. おわりに

本研究では、Q&A サイトにおける利用者の質問・回答履歴が、質問者がベストアンサーを選択するとき何らかの影響を与えているかどうか、また、複数の回答群から自動的にベ

ストアンサーを推定するとき利用可能であるかどうかについて調べるため、Q&A サイトに登録している利用者の質問・回答履歴を分析し、頻りにサイトを使用している利用者およびベストアンサーに数多く選ばれている回答者の質問・回答行動の特徴について考察した。また、複数の回答群から自動的にベストアンサーを推定するとき、質問・回答履歴を使用した推定が可能かどうか予備的な実験を行い、検証した。その結果、利用者の質問・回答履歴には、カテゴリごと、利用者ごとの特徴があり、それを利用することによって、ある程度ベストアンサーを推定することが可能であることがわかった。

今後の課題として、以下のようなことを行いたい。

- 以前の論文¹⁰⁾ で提案した質問タイプの分類と利用者の質問・回答行動の特徴との関連についての分析
- BA 推定に有効な、内容以外の要素についての検討
- 内容以外の要素を素性として用いた機械学習による BA 推定

謝辞 本研究の実施にあたっては、ヤフー株式会社が国立情報学研究所に提供した「Yahoo!知恵袋 研究機関提供用データ 国立情報学研究所 (NII) 提供版 ver.1」¹⁷⁾ を利用した。

本研究は、平成 21 年度国立情報学研究所共同研究「コミュニティ情報における有用性・信頼性」による。

参 考 文 献

- 1) Adamic, L. et al.: Knowledge Sharing and Yahoo Answers: Everyone Knows Something, *Proc. of the 17th International Conference on World Wide Web*, Beijing, WWW2008 (2008).
- 2) Agichtein, E. et al.: Finding High-Quality Content in Social Media, *WSDM '08: Proc. of the International Conference on Web Search and Web Data Mining*, ACM, pp.183-194 (2008).
- 3) Bian, J. et al.: Learning to Recognize Reliable Users and Content in Social Media with Coupled Mutual Reinforcement, *WWW '09: Proc. of the 18th International Conference on World Wide Web*, ACM, pp.51-60 (2009).
- 4) Jurczyk, P. and Agichtein, E.: Discovering Authorities in Question Answer Communities by Using Link Analysis, *CIKM '07: Proc. of the 16th ACM Conference on Information and Knowledge Management*, ACM, pp.919-922 (2007).
- 5) 甲谷優ほか: QA サイトにおける質問応答グラフの成長パターン分析, 日本データベース学会論文誌, Vol.7, No.3, pp.61-66 (2008).
- 6) 甲谷優, 川島晴美, 藤村考: QA サイト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, 日本データベース学会論文誌, Vol.8, No.1, pp.89-94 (2009).
- 7) Kim, S., Oh, J. and Oh, S.: Best-Answer Selection Criteria in a Social Q&A site from the User-Oriented Relevance Perspective, *American Society for Information Science and Technology (ASIS&T) 2007 Annual Meeting*, ASIS&T (2007). (http://curric.dlib.vt.edu/papers/ASIST2007_0525_Yahoo_Answers_Final_version.pdf).
 - 8) Kim, S. and Oh, S.: User's Relevance Criteria for Evaluating Answers in Social Q&A Site, *Journal of the American Society for Information Science and Technology*, Vol.60, No.4, pp.716-727 (2009).
 - 9) Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment, *J. ACM*, Vol.46, No.5, pp.604-632 (1999).
 - 10) 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. 情報学基礎研究会報告, Vol.2009-FI-95, pp.1-8 (2009).
 - 11) Liu, Y., Bian, J. and Agichtein, E.: Predicting Information Seeker Satisfaction in Community Question Answering, *SIGIR '08: Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp.483-490 (2008).
 - 12) 三浦麻子, 川浦康至: 人はなぜ知識共有コミュニティに参加するのか: 質問行動と回答行動の分析, 社会心理学研究, Vol.23, No.3, pp.233-245 (2008).
 - 13) OKWave: 教えて!goo. <http://oshiete.goo.ne.jp/> (参照 2010-01-03).
 - 14) Saracevic, T.: Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II, Part III, *Journal of the American Society for Information Science and Technology*, Vol.58, No.13, pp.1915-1933, 2126-2144 (2009).
 - 15) Wang, X.-J. et al.: Ranking Community Answers by Modeling Question-Answer Relationships via Analogical Reasoning, *SIGIR '09: Proc. of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp.179-186 (2009).
 - 16) Yahoo!JAPAN: Yahoo!知恵袋. <http://chiebukuro.yahoo.co.jp/> (参照 2010-01-03).
 - 17) Yahoo!JAPAN: 「Yahoo!知恵袋」データの提供について. <http://research.nii.ac.jp/tdc/chiebukuro.html> (研究機関用データは国立情報学研究所 IDR 事務局から配布されている) (参照 2010-01-03).