Vol.2010-IFAT-100 No.5
2010/11/4

Q&A サイトにおける質問と回答の分析 (4) – 質問タイプ分類の一致度について –

栗 山 和 子[†1]　神 門 典 子[†2]

著者らは，以前の論文で Q&A サイトにおける質問のタイプ分類を提案した．本研究では，NTCIR-8 ワークショップのコミュニティQA パイロットタスクにおいて使用した 1500 件の質問データを，4 人の判定者によって，提案した 13 種類の質問タイプに分類し，判定者間の一致度を計算することによって，分類の信頼性を検証した．また，質問タイプと判定の一致・不一致との関連について考察し，質問タイプの妥当性について検討した．結果として，判定者間の一致度は高く，質問タイプ分類の信頼性と妥当性が確かめられた．

Analysis of Questions and Answers in Q&A Site(4) – Inter-assessors Agreement for Question Types Classification –

Kazuko Kuriyama[†1] and Noriko Kando[†2]

We had proposed classification of question types in community Q&A sites. In this paper, we classified 1500 questions, which were used for NTCIR-8 Community QA pilot task, into 15 types manually by four assessors. Then, we computed κ coefficient for inter-assessor agreement to examine reliability of the classification. Furthermore, we analysed questions with incomplete agreement to examine validity of the question types. The result showed validity of the question types and reliability of the classification.

†1 白百合女子大学
Shirayuri College
†2 国立情報学研究所
National Institute of Informatics

1. はじめに

近年では，コミュニティQA(CQA) サイトと呼ばれる Wolrd Wide Web 上のサービスの利用者が増えてきている[12),15)]．CQA サイトの仕組みは，登録した利用者が質問を投稿し，また，別の利用者が回答を投稿するというものである．CQA サイトは，WWW 上の掲示板のように匿名（ユーザ ID）で自由に質問・回答を投稿できることから，サーチエンジンのような情報アクセス支援ツールとしてだけではではなく，コミュニケーションや知識の共有・交換の場として捉えられている．CQA サイトにおける質問・回答行動は利用者の自発的なものであるが，利用者の年齢・性別・職業・利用目的等は様々であり，質問・回答で記述される情報・知識の内容と表現も多様であるため，ある質問に対して質問者の意図に沿った回答や信頼性が高い回答が得られるとは限らない．

そこで，著者らは，以前の論文[11)] で，質問者の目的や意図によって質問文の表現が異なるかどうか，また，質問文の表現が回答の内容や表現にどのように影響を与えているかを調べるため，質問のタイプについての分析を行った．本稿では，その質問タイプの妥当性を検証するため，NTCIR-8 ワークショップのコミュニティQA パイロットタスク（以下，CQA タスク）においてテストコレクションとして使用した質問データを用いて，複数の判定者による質問タイプの分類を行い，判定者間の分類の一致度について考察する．また，質問タイプの定義の類似が判定の一致・不一致と関連していかどうか考察し，質問タイプの妥当性について検討する．

2 節では CQA サイトについての関連研究について簡単に述べる．3 節では，複数の判定者の分類の一致度を κ 係数を用いて計算し，質問タイプ分類の信頼性を検証する．4 節では，質問タイプの定義の類似と分類の一致・不一致の関連について調べ，質問タイプの分け方が妥当かどうか考察する．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

CQA サイトに関する研究は数多く行われている．本節では，質問と回答をデータとして内容的あるいは数量的に分析し，質問に対して適切な回答を得るために利用することを目的としている研究の主要なものについて述べる．

海外の CQA サイトとしては Yahoo! Answers が代表的なものであり，CQA サイトに関する研究として，Yahoo! Answers の質問・回答の質の評価や質問に対する適切な回答の抽出・選択などに関する研究が数多く行われている．Jurczyk ら[8)] は，HITS アルゴリズムに

Vol.2010-IFAT-100 No.5
2010/11/4

基づき，利用者・回答間グラフを用いて，オーソリティを探し，利用者の専門性を評価している．Jeon ら[7] は，回答の質を予測するために，クリックされた回数など，テキスト以外の特徴を利用する方法を提案している．Agichtein ら[2] は，質問と回答の組の様々な特徴について分析し，教師付き機械学習によって，質問者が投稿された回答に満足するかどうかを予測することを試みている．Bian ら[3] は，訓練用データの件数を減らすため，教師付き他者評価強化学習を用いて，質問・回答の質と利用者の評価の推定を行っている．Suryanto ら[14] は，既存の回答を，新たに投稿された質問に対する適合性と質問自体の質という 2 点を総合して評価し，質問に適した回答を選択する方法を提案している．

CQA サイトのコミュニケーションのタイプに関する研究として，Adamic ら[1] は，利用者をノードとし，回答を回答から質問者へのエッジとしてグラフ化した QA ネットワークを用いて，CQA サイトで行われているコミュニケーションには知識交換，相談，議論の 3 つのタイプがあり，CQA サイトで使用されている既存のカテゴリがそのいずれかのタイプに分類できることを示している．Adamic らの研究に基づき，甲谷ら[9] は，「教えて!goo」[12] について，手作業で各カテゴリの質問 10 個ずつを知識交換，相談，議論の 3 つのタイプに分類し，同様の素性がカテゴリ分類に有効であるかどうかを検証している．その結果，「教えて!goo」に関しては，Adamic らの用いた素性では，カテゴリを 3 つのコミュニケーションタイプに適切に分類できないことが示されている．

質問の分類に関する研究としては，Kim ら[10] は，Yahoo! Answers のベストアンサーに付けられた質問者のコメントを手作業で分析することにより，質問者がベストアンサーを選んだ理由を適合性の基準 (relevance criteria) として分類し，ベストアンサーの選択理由の分布について考察している．また，Rodrigues ら[13] らは，Yahoo! Answers の質問を 3 つの次元，General/Personal, Community/Individual, Social/Non-social で分析し，8 つの質問タイプ，Factual Information, General Advice, General Opinion, Personal Opinion, Chatting, Entertainment, Other に分類している．

以上のように，CQA サイトに関して，質問・回答の質の評価，ベストアンサーや類似の質問・回答の抽出・選択，コミュニケーションの分析，質問タイプの分類など，様々な研究が行われている．

以前の論文[11] では，上記のような研究への応用も念頭に置いて，実際の Q&A サイトの質問を分析し，CQA サイトに適したタイプの分類を提案した．本稿では，複数の判定者による質問タイプの一致度を用いて，質問タイプ分類の信頼性と妥当性について検証する．

3. 判定者間一致度

3.1 使用データ・判定者・質問タイプ

本研究では，質問データとして，NTCIR-8 ワークショップのコミュニティQA パイロットタスクのテストコレクションの質問データを使用する．CQA タスクの詳細や質問データの形式は，オーガナイザによる overview[5] を参照されたい．

CQA タスクのテストコレクションの質問・回答データは，「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」[16]（以下，知恵袋データ）から，各カテゴリの質問件数の比率に応じてランダムに抽出した 1500 件の質問とそれに付随する回答から成る．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」ベータ版に投稿された質問と回答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 件から構成されている．1 つの質問には，1 つのベストアンサーと 0 個以上のその他の回答が存在する．

Yahoo!知恵袋のカテゴリは，利用者の動向に合わせて随時変更されており，知恵袋データにおいても，元の質問・回答の投稿期間によってカテゴリ構成・名称が統一されていない部分があるため，CQA タスクでは，テストコレクションを作成した時点での Yahoo!知恵袋のカテゴリ構成に合わせて，カテゴリを再構成している．また，CQA タスクでは，知恵袋データ全体におけるカテゴリごとの質問件数の割合を反映するように，各カテゴリから抽出する質問件数を設定しているため，質問件数はカテゴリごとに異なっている．

本稿では，テストコレクションの再構成されカテゴリ構成を用いて分析を行うので，テストコレクションのトップカテゴリ分類と各カテゴリに属する質問の件数を表 1 に示す．トップカテゴリは，さらに詳細な 118 のカテゴリに分かれている[5] が，テストコレクションの質問件数の総数が 1500 件と少ないので，本稿では，14 のトップカテゴリのみを使用した．

質問タイプの分類作業は，CQA タスクの適合判定システムを利用して，同タスクの適合判定作業を行ったのと同じ 4 人の判定者が行った．判定者の属性を表 2 に示す．

分類作業は，CQA タスクのメインタスクの正解データ作成のための，各質問の質の判定（A：質問である，B：質問ではない，の 2 値），および，質問に付随する回答の質の判定（A：質問に適合，B：質問に部分的に適合または不適合，C：質問には無関係（不適合））の 3 値）と同時に同じ作業ウィンドウ上で行っているため，質問タイプの判定時には質問の内容だけでなく，付随する回答の内容も参照できるようになっている．

質問タイプは，以前の論文[11] で提案したものをそのまま使用した．次項以降で参照する

表 1 NTCIR-8 CQA タスク テストコレクションにおけるカテゴリと質問件数

カテゴリ番号	トップカテゴリ（日本語）		質問件数
1	yahoo	（Yahoo! JAPAN）	222
2	entertainment	（エンターテインメントと趣味）	167
3	health	（健康、美容とファッション）	154
4	lifeguide	（暮らしと生活ガイド）	136
5	internet	（インターネット、PC と家電）	135
6	sports	（スポーツ、アウトドア、車）	120
7	love	（生き方と恋愛、人間関係の悩み）	120
8	education	（教養と学問、サイエンス）	120
9	school	（子育てと学校）	89
10	news	（ニュース、政治、国際情勢）	63
11	travel	（地域、旅行、お出かけ）	58
12	business	（ビジネス、経済とお金）	42
13	career	（職業とキャリア）	38
14	manners	（マナー、冠婚葬祭）	36
合計			1,500

表 2 判定者の属性

判定者	文系/理系	性別	学年	年齢	専門分野	知恵袋利用頻度	質問・回答歴
J1	理系	男性	学部 3 年	20 代	情報	2 週間に 1 度	なし
J2	文系	男性	学部 4 年	20 代	文芸思想	週に 2-3 回	なし
J3	理系	女性	学部 3 年	20 代	生化学	月に 2-3 回	なし
J4	文系	女性	学部 3 年	30 代	アジア史	月に 1 回	なし

ため，その定義を表 3 に再掲する．

3.2 分類結果

各判定者による質問タイプの分類結果をカテゴリごとに表 4,5 と図 1（積み上げ図）に示す．1 つの質問データは 1 つの質問タイプに分類される．1 つの質問データの中に複数の質問文が含まれる場合もあるが，判定者には，そのような場合には主要な質問文の質問タイプに分類するように，すなわち，必ず，1 つの質問データは 1 つの質問タイプに分類するように指示した．

大きく，客観的な情報や事実を求める A タイプ（情報検索型），個人的な助言・意見・経験などを求める B タイプ（社会調査型），実際は質問ではない C タイプ（非質問型）と分けて考えると，表 4,5，図 1 からわかるように，判定者によらず，A タイプが多いカテゴリ，B タイプが多いカテゴリ，両方のタイプが含まれているカテゴリがあることがわかる．例えば，5(internet) や 6(education) は A タイプが多く，7(love) や 9(school) は B タイプが多

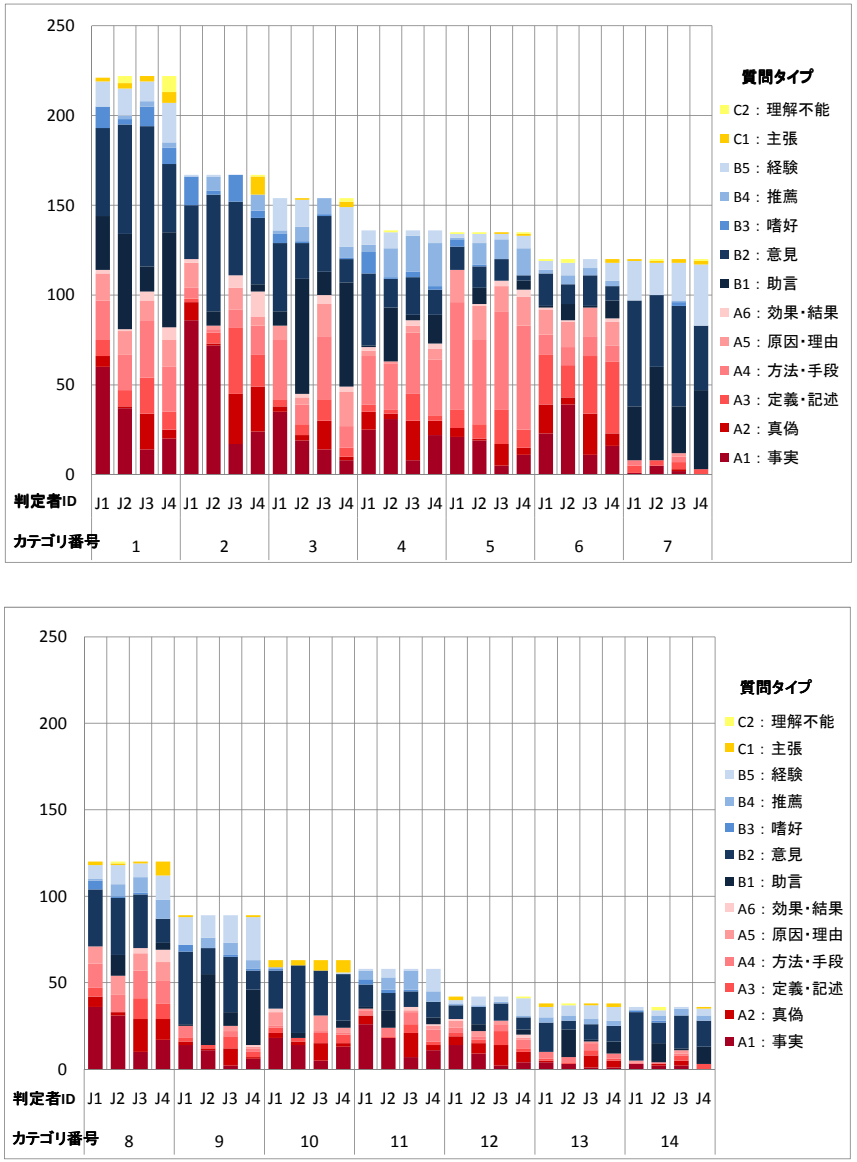

図 1 質問タイプ分類結果

表 3 質問タイプ

質問タイプ ID	質問タイプ	定義
A. 情報検索型：サーチエンジンや図書館のレファレンス・サービスを利用して回答を探すことが可能な質問		
A1	事実	事実としての名称（人・組織の名称，場所・位置等）や数的表現（金額，日付，大きさ等）を尋ねる質問
A2	真偽	伝聞や推測の真偽や可能・不可能を尋ねる質問
A3	定義・記述	ある事物の定義・証明・説明・属性・事例・歴史的経緯などを尋ねる質問
A4	方法・手段	あることを行う方法や手段を具体的に尋ねる質問
A5	原因・理由	ある物事の客観的な原因や理由を尋ねる質問
A6	効果・結果	ある物事の客観的な結果・効果・過程・現象を尋ねる質問
B. 社会調査型：客観的な正解はなく，特定の個人・集団に対してアンケート調査を行うことで回答を得るような質問		
B1	助言	質問者の意見・行動について主観的な価値判断を伴う助言を求める質問
B2	意見	ある物事について回答者の意見を広く求める質問
B3	嗜好	ある物事について回答者個人の好みを尋ねる質問
B4	推薦	ある物事について回答者の推薦するものや一般に人気・評価が高いものを尋ねる質問
B5	経験	ある物事について回答者の経験・体験の有無あるいは経験・体験の具体的内容・実例を尋ねる質問
C. 非質問型：情報検索やアンケート調査によって客観的あるいは主観的な回答を得ることが目的ではなく，質問者が自分の主張に対する反響・反応を求めている記述表現		
C1	主張	質問ではなく、ある物事について質問者の意見・嗜好・推測などを述べているもの
C2	理解不能	記述として何が書かれているのか分析者には理解できなかったもの

く，3(health) や 10(news) は判定者ごとに多いタイプが異なっている．C タイプは，質問としてはその他のようなものなので，どのカテゴリにも非常に少数しか含まれていない．

詳細な 13 の質問タイプ別で見ると，全体的には，A1,A4,A5,B1,B2,B5 が多いが，カテゴリごと，判定者ごとに，多いタイプが異なっていることがわかる．カテゴリ 5,6 は A タイプが多いが，5 は A4（方法・手段）が多く，6 は A1（事実），A3（定義・記述），A5（原因・理由）が多い．これは，5（internet：インターネット、PC と家電）がパソコンやインターネットなどの具体的な操作法などを尋ねる質問が多いのに対し，6（education：教養と学問、サイエンス）が学術的・科学的な事実や問題についての質問が多いからであると考えられる．B タイプが多いカテゴリについても同様であり，7（love：生き方と恋愛、人間関係の悩み）では B2（意見）が圧倒的に多いのに対し，9（school：子育てと学校）では B1（助言）と B2（意見）に分散している．

9 では，属性が理系の判定者 J1 と J3 が B2 に分類している質問が多いのに対し，文系の判定者 J2 と J4 は B1 に分類している質問が多い．B1 と B2 の違いは，質問者が自分の意見・行動に対して助言を求めているか否かであるが，助言を求めているかどうかの解釈が判

表 4 質問タイプ分類結果（カテゴリ 1–10）

		質問タイプ												
カテゴリ番号	判定者	A1	A2	A3	A4	A5	A6	B1	B2	B3	B4	B5	C1	C2
1	J1	60	6	9	22	15	2	30	49	12	0	14	2	1
	J2	37	1	9	20	13	1	53	61	3	2	15	3	4
	J3	14	20	20	32	11	5	14	78	11	3	11	3	0
	J4	20	5	10	25	15	7	53	38	9	3	22	6	9
2	J1	86	10	2	6	14	2	0	30	16	0	1	0	0
	J2	72	1	6	2	2	0	8	65	2	8	1	0	0
	J3	17	28	37	10	12	7	0	41	15	0	0	0	0
	J4	24	25	18	16	5	14	4	37	4	9	0	10	1
3	J1	35	3	4	33	8	0	8	38	5	2	18	0	0
	J2	19	3	6	11	4	2	64	20	1	8	15	1	0
	J3	14	16	12	35	18	5	13	31	1	9	0	0	0
	J4	8	2	5	12	19	3	58	13	1	6	22	3	2
4	J1	25	10	4	27	3	2	1	40	12	4	8	0	0
	J2	31	3	2	26	1	0	30	16	1	16	9	0	1
	J3	8	22	15	34	4	3	3	21	3	20	3	0	0
	J4	22	8	3	31	6	3	16	14	2	24	7	0	0
5	J1	21	5	10	60	18	0	0	13	4	1	2	0	1
	J2	19	1	8	47	19	1	9	12	1	12	5	0	1
	J3	5	12	19	55	14	3	0	12	0	11	3	1	0
	J4	11	4	10	58	16	4	5	3	0	15	7	1	1
6	J1	23	16	28	11	14	1	1	18	0	2	5	0	1
	J2	39	4	18	10	14	1	9	11	0	5	7	0	2
	J3	11	23	32	11	16	0	1	17	0	4	5	0	0
	J4	16	7	40	9	13	2	10	8	0	3	10	2	0
7	J1	1	0	4	3	0	0	30	59	0	0	22	1	0
	J2	5	0	3	0	0	0	52	40	0	0	18	1	1
	J3	2	1	4	3	2	0	26	56	2	1	21	2	0
	J4	0	0	3	0	0	0	44	36	0	0	34	2	1
8	J1	36	6	5	14	10	0	0	33	5	1	8	2	0
	J2	31	2	0	10	11	0	12	33	1	7	11	1	1
	J3	10	19	12	16	10	3	0	31	1	9	8	1	0
	J4	17	12	9	13	11	7	4	14	0	11	14	8	0
9	J1	14	2	2	7	0	0	1	42	4	0	16	1	0
	J2	11	1	2	0	0	0	41	15	0	6	13	0	0
	J3	2	10	7	3	3	0	8	32	1	7	16	0	0
	J4	6	1	3	2	1	1	32	11	1	5	25	1	0
10	J1	18	3	3	1	8	2	0	22	1	1	0	4	0
	J2	14	2	2	0	0	0	3	39	0	0	0	3	0
	J3	5	10	6	1	9	0	0	26	0	0	0	6	0
	J4	13	2	5	1	3	0	4	27	0	0	1	7	0

表 5 質問タイプ分類結果（カテゴリ 11-14）

カテゴリ番号	判定者	質問タイプ												
		A1	A2	A3	A4	A5	A6	B1	B2	B3	B4	B5	C1	C2
11	J1	26	5	0	3	1	0	1	13	3	5	1	0	0
	J2	18	0	1	5	0	0	10	10	2	7	5	0	0
	J3	7	14	5	7	1	2	0	9	1	11	1	0	0
	J4	11	3	2	7	2	1	4	9	0	6	13	0	0
12	J1	14	5	2	3	4	1	0	8	0	1	2	2	0
	J2	9	6	2	2	3	0	4	10	0	1	5	0	0
	J3	2	12	8	4	2	0	0	10	0	1	3	0	0
	J4	4	6	2	5	1	2	3	7	0	1	10	0	1
13	J1	4	1	1	4	0	0	0	17	0	3	6	2	0
	J2	3	0	1	3	0	0	16	5	0	3	6	0	1
	J3	1	7	3	4	1	0	1	9	0	3	8	1	0
	J4	1	4	1	3	0	0	7	9	0	3	8	2	0
14	J1	3	0	0	1	1	0	0	28	1	0	2	0	0
	J2	2	1	0	1	0	0	11	12	1	3	3	0	2
	J3	2	3	3	1	2	0	1	19	0	4	1	0	0
	J4	0	0	3	0	0	0	10	15	0	3	4	1	0

定者の属性によって異なる可能性がある．

3.3 一致度（κ 係数）の計算結果

複数の判定者による分類の信頼性を検証するため，判定者間一致度を測る尺度である κ 係数を計算した．κ 係数は Cohen[4] によって提案された統計的評価尺度であり，2 人の判定者による複数のカテゴリへの分類の信頼性を評価するために使用される．Cohen の κ 係数を 3 人以上の判定者に拡張したものが Fleiss の κ 係数である[6]．本稿では，2 人の判定者間の一致度を中心に考察するが，参考として 4 人の判定者と 3 人の判定者の一致度を Fleiss の κ で計算した結果も合わせて示す．どちらの κ 係数も 1 に近いほど一致度が高く，一致度の強さは値の範囲によって表 6 のように解釈される．

表 7,8 に 4 人全員，3 人ずつの組合せ，2 人ずつの組合せで計算した κ 係数を示す．4 人と 3 人の組合せについては Fleiss の κ 係数を用いた．表 7 は，13 の質問タイプへの分類の一致度，表 8 は，分類を大きく A,B,C の 3 タイプにまとめたときの一致度である．比較を容易にするため，それぞれを，図 2,3 に折れ線グラフとして示す．

表 7 からわかるように，13 の質問タイプについては，2 人ずつの組合せでは，カテゴリ 14 を除いて κ 係数は 0.2〜0.6 の範囲に入っており，全体的な平均は 0.39 となるので，一致度の判定は fair から moderate となる．カテゴリ 14(manners) は 0〜0.45 とバラつきが

表 6 κ 係数による一致度

κ 係数	一致度の判定
< 0	poor
0.00–0.20	slight
0.21–0.40	fair
0.41–0.60	moderate
0.61–0.80	substantial
0.81–1.00	almost perfect

表 7 判定者間一致度（κ 係数）（13 タイプ）

	全員	J1-J2-J3	J1-J2-J4	J2-J3-J4	J1-J2	J1-J3	J1-J4	J2-J3	J2-J4	J3-J4
All	0.41	0.41	0.40	0.42	0.39	0.45	0.37	0.39	0.46	0.43
1	0.40	0.45	0.42	0.38	0.50	0.43	0.33	0.44	0.43	0.30
2	0.30	0.30	0.26	0.28	0.30	0.42	0.23	0.25	0.30	0.35
3	0.28	0.26	0.29	0.29	0.29	0.29	0.25	0.26	0.39	0.28
4	0.35	0.31	0.31	0.42	0.27	0.32	0.25	0.37	0.43	0.47
5	0.47	0.46	0.45	0.47	0.48	0.47	0.44	0.45	0.45	0.52
6	0.38	0.35	0.35	0.40	0.33	0.36	0.38	0.36	0.36	0.48
7	0.48	0.46	0.48	0.48	0.40	0.58	0.48	0.40	0.56	0.50
8	0.44	0.49	0.39	0.41	0.43	0.59	0.37	0.47	0.40	0.38
9	0.35	0.29	0.35	0.41	0.24	0.39	0.32	0.32	0.60	0.37
10	0.44	0.39	0.41	0.43	0.33	0.54	0.47	0.31	0.46	0.53
11	0.31	0.27	0.31	0.32	0.25	0.34	0.33	0.27	0.38	0.35
12	0.44	0.47	0.40	0.45	0.41	0.51	0.39	0.51	0.43	0.43
13	0.53	0.46	0.51	0.54	0.47	0.56	0.56	0.42	0.56	0.66
14	0.20	0.17	0.15	0.30	0.09	0.24	0.00	0.30	0.45	0.19

表 8 判定者間一致度（κ 係数）（3 タイプ：A/B/C）

	全員	J1-J2-J3	J1-J2-J4	J2-J3-J4	J1-J2	J1-J3	J1-J4	J2-J3	J2-J4	J3-J4
All	0.62	0.62	0.59	0.60	0.58	0.70	0.61	0.59	0.59	0.64
1	0.53	0.58	0.52	0.55	0.59	0.54	0.44	0.61	0.54	0.48
2	0.63	0.62	0.55	0.63	0.49	0.82	0.63	0.59	0.57	0.74
3	0.35	0.35	0.35	0.29	0.34	0.53	0.39	0.28	0.38	0.33
4	0.55	0.52	0.52	0.52	0.53	0.60	0.59	0.44	0.45	0.69
5	0.58	0.61	0.52	0.61	0.57	0.62	0.47	0.65	0.54	0.66
6	0.54	0.53	0.48	0.55	0.43	0.60	0.58	0.57	0.45	0.65
7	0.50	0.47	0.49	0.50	0.38	0.57	0.51	0.45	0.61	0.47
8	0.70	0.75	0.65	0.72	0.68	0.79	0.56	0.78	0.70	0.69
9	0.55	0.58	0.50	0.51	0.50	0.75	0.48	0.45	0.55	0.56
10	0.60	0.56	0.51	0.60	0.40	0.80	0.62	0.51	0.54	0.76
11	0.58	0.58	0.52	0.56	0.43	0.75	0.63	0.60	0.51	0.60
12	0.57	0.67	0.47	0.49	0.58	0.84	0.56	0.61	0.30	0.58
13	0.69	0.62	0.83	0.61	0.74	0.64	0.94	0.51	0.80	0.58
14	0.30	0.39	0.25	0.25	0.48	0.38	0.13	0.35	0.11	0.30

Vol.2010-IFAT-100 No.5
2010/11/4

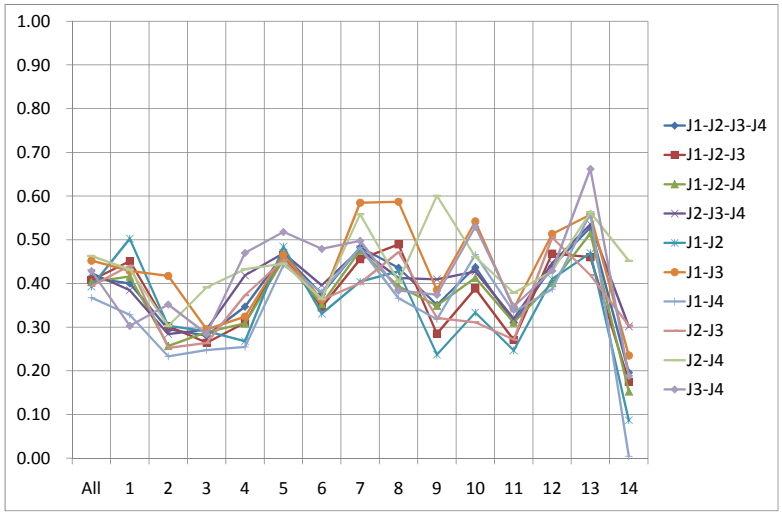

図 2 判定者間一致度（κ 係数）（13 タイプ）

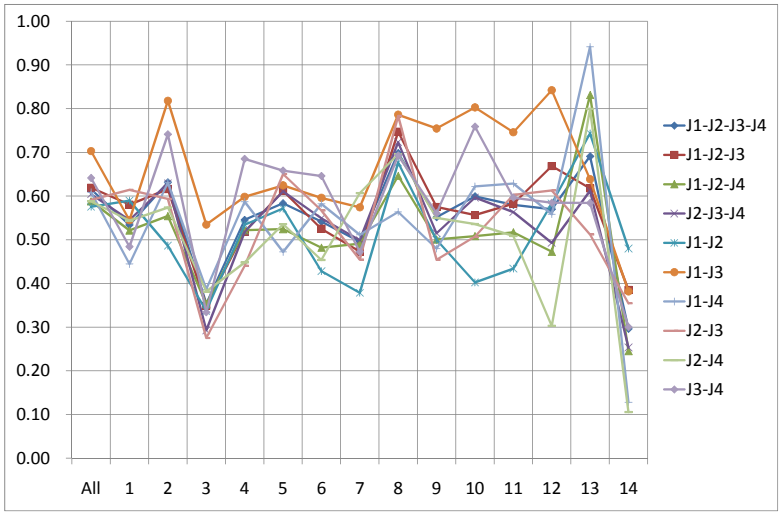

図 3 判定者間一致度（κ 係数）（3 タイプ：A/B/C）

大きいが，これが質問件数が十分でない（38 件）ことの影響であるか否かついては，別途，サンプル数を考慮に入れた検討が必要である．また，表 8 からわかるように，A,B,C の 3 タイプにまとめた場合は，ほとんどの値が 0.4〜0.8 の範囲に入り，全体的な平均は 0.56 となるので，一致度の判定は moderate から substantial になる．

以上のことから，詳細な質問タイプについても，それを大きく 3 つのタイプにまとめた場合でも，判定者間の一致度は高く，判定（分類）の信頼性と質問タイプの妥当性が確かめられた．

カテゴリごとの一致度を見てみると，図 2,3 からわかるように，13 タイプと 3 タイプの質問タイプへの分類の一致度の傾向は似ているが，5(internet)，6(education)，7(love)，9(school) のようにどちらの分類でも一致度が比較的高いカテゴリもあれば，3(health) や 9(school) のように一致度が高くないカテゴリもある．前項で述べたように，5,6 は A タイプが多く，7,9 は B タイプが多く，3,9 は判定者によって多い質問タイプが異なる．すなわち，A,B いずれかのタイプに質問が偏っているカテゴリでは比較的に判定者間の一致度が高くなる（安定する）と考えられる．

判定者の組合せについては，13 タイプでは J1–J3, J2–J4 の一致度が高く，3 タイプでは J1–J3, J3–J4 が高い．前項でも触れたように，表 2 の通り，判定者 J1 と J3 は理系，J2 と J4 は文系であるので，理系・文系という属性によって質問あるいは質問タイプの解釈が似ている可能性があり，それが詳細な質問タイプに分類する場合の一致度の高さとして現れたと考えられる．

4. 一致・不一致の分析

質問タイプの妥当性について考察するため，判定者が混同しやすい質問タイプとその理由について検討する．質問タイプの分布として，表 9 に全判定者の分類が一致した質問タイプごとの質問件数，表 10 に 3 人の判定者の分類が一致した件数と 2 人ずつの組が一致した件

数を示す．例えば，表 10 の「A1/A2」について，列の「3:1」は 3 人が A1 に分類し 1 人が A2 に分類した,「1:3」は 1 人が A1 に分類し 3 人が A2 に分類した,「2:2」は 2 人が A1 に分類し 2 人が A2 に分類したという意味である．

表 9 全判定者一致の件数

タイプ	件数	タイプ	件数	タイプ	件数
A1	23	B1	32	C1	3
A2	6	B2	108	C2	0
A3	35	B3	4		
A4	74	B4	12		
A5	24	B5	40		

表 10 不一致の組合せの件数

タイプ組合せ	3:1	1:3	2:2	合計	タイプ組合せ	3:1	1:3	2:2	合計
A1 / A2	23	18	24	65	A4 / B4	5	2	0	7
A1 / A3	33	25	20	78	A4 / B5	1	1	1	3
A1 / A4	5	20	8	33	A4 / C2	0	0	1	1
A1 / A5	3	11	1	15	A5 / A6	1	0	0	1
A1 / A6	3	1	4	8	A5 / B1	4	0	1	5
A1 / B1	2	3	1	6	A5 / B2	12	7	10	29
A1 / B2	6	16	6	28	A5 / B5	1	1	0	2
A1 / B4	1	3	2	6	A5 / C1	1	0	0	1
A1 / B5	3	3	0	6	A5 / C2	0	0	1	1
A2 / A3	2	2	0	4	A6 / B2	1	0	1	2
A2 / A4	0	0	3	3	A6 / B5	3	0	0	3
A2 / A5	1	1	1	3	B1 / B2	28	30	46	104
A2 / B1	2	3	2	7	B1 / B4	1	0	1	2
A2 / B2	3	1	3	7	B1 / B5	1	5	3	9
A2 / B5	1	1	0	2	B2 / B3	13	5	6	24
A3 / A4	0	2	1	3	B2 / B4	2	10	4	16
A3 / A5	0	5	1	6	B2 / B5	23	20	18	61
A3 / B2	2	1	1	4	B2 / C1	15	3	1	19
A3 / C2	1	0	0	1	B3 / B4	1	15	7	23
A4 / A5	6	3	3	12	B4 / B5	0	3	0	3
A4 / B1	29	7	16	52	C1 / C2	3	0	0	3
A4 / B2	5	2	1	8					

表 10 からわかるように，不一致の組合せでは，B1/B2 が 104 件と最も多いが，表 9 から全判定者が一致している件数も，B2 が 108 件と最も多いので，件数が多いものが混同しやすい組合せだとは必ずしも言えない．以下に，具体例として，不一致の合計件数が 50 件以上である，A1/A2, A1/A3, A4/B1, B1/B2, B2/B5 の質問を挙げる．

[A1/A2] entertainment（邦楽）：稲森いずみが出ているドラマの主題歌はもう販売されてますか？

[A1/A3] news（話題のことば）：今朝のニュースで皇太子の記者会見が映ってまして、「私が感銘を受けた詩」みたいな感じで、ある詩が朗読されてたのですが、（子育てに関する詩） その全文が見れるＨＰなどありましたら、教えてください。

[A4/B1] education（動物、植物、ペット）：同居している猫に猫パンチされまくりたいんですけどどうしたらいいですか？

[B1/B2] travel（海外）：レゲエが好きでもっと深く勉強するために メキシコに行こうと思っています。 滞在期間は７日くらいを目処に考えています。 行ってはいけない場所などありますか。 気をつけておくべきことなどありましたら教えてください。

[B2/B5] business（企業と経営）：サラリーマンっていそがしいのですか？

上の質問は，各組合せの不一致で代表的なものであるが，不一致の理由は，(a) 質問の内容・形式の解釈の違い，あるいは，(b) 質問タイプの解釈の違いのいずれかであると考えられる．上の例では，A1/A2 の質問は,「発売されていますか？」という質問を事実を聞いていると考えれば A1（事実）に分類できるが,「発売されているか否か」という真偽を尋ねていると形式的に解釈すれば A2（真偽）に分類されるとも考えられる．また，B2/B5 の質問は，一般的な意見を聞いていると考えれば B2（意見）に，サラリーマンとして勤めたことがある人に聞いていると考えれば B5（経験）を尋ねているとも考えられる．これらは，(a) 質問の内容・形式の解釈の違いによるものである．それに対して，A4/B1 の質問は，猫にパンチをされることに対する解決法を尋ねる質問であり，内容・形式としては方法を尋ねる文である．方法・手段についての質問であれば A4 に分類するのが適切であるが，B1（質問者の意見・行動に対する主観的価値判断）に分類した判定者は，A4 あるいは B1 の定義の解釈が他の判定者と異なっているため，B1 に分類したと考えられる．つまり，質問タイプの定義の解釈が異なるので，B1 に分類したと考えられる．

前者は質問文の形式によるものであるので，機械学習等を用いて自動で分類を行う場合は，質問文の形式的な特徴以外の性質も利用する必要があると考えられる．後者の不一致の理由は，判定者に対する説明不足によるものである．今回の分類作業の説明・指示は文書で

行い，口頭での補足説明はしていない．そのため，判定者に質問タイプの一部の定義が十分理解されていなかった可能性がある．前項で示したように，判定者間一致度によって分類（判定）の信頼性そのものは確認できたが，質問タイプの定義と判定者への説明の方法については検討と改善の必要がある．

5. おわりに

本稿では，以前の論文[11]で提案した質問タイプの妥当性を検証するため，NTCIR-8 ワークショップのコミュニティQA パイロットタスクの質問データを用いて，複数の判定者による質問タイプの分類を行い，判定者間の一致度について考察した．その結果，複数の判定者間の一致度の信頼性は高く，質問タイプ分類の妥当性が確かめられた．

詳細な質問タイプの分類ではカテゴリによっては一致度が低いものがあったが，質問タイプを A,B,C の 3 つにまとめた分類では，一致度は改善した．したがって，実際に，CQA サイトにおいて質問者に質問タイプを選択・付与してもらう場合や投稿された質問を自動で質問タイプに分類する場合には，大きい 3 つのタイプに分ける方が容易であり実用的であると考えられる．

また，判定者の属性（理系・文系）によって，同じ質問を分類する質問タイプが異なる傾向があり，属性が同じ判定者間の一致度は属性が異なる判定者間の一致度より高くなることがわかった．

今後の課題として，以下のようなことが挙げられる．

- 判定者の属性が質問タイプ分類にどういう影響を与えているか検討する．
- 判定者が質問タイプを混同しやすい理由を，質問文の内容・形式と質問タイプの定義に分けて考察し，質問タイプの再構成について検討する．
- 質問を「情報検索型」と「社会調査型」に自動的に分けるため，質問の記述形式や表現パターンなどについて分析し，自動分類に使用できる特徴とその取得方法について検討する．

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を通して提供している「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」[16]，および，NTCIR-8 コミュニティQA パイロットタスク参加者用テストコレクション[5] を使用した．本研究は，平成 22 年度国立情報学研究所公募型共同研究の一部として遂行した．

参 考 文 献

1) Adamic, L. et al.: Knowledge Sharing and Yahpp 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, *Procddeings of WSDM '08*, ACM, pp.183–194 (2008).
3) Bian, J. et al.: Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media, *Proc. of WWW '08*, ACM, pp.467–476 (2008).
4) Cohen, J.: A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, Vol.20, No.1, pp.37–46 (1960).
5) Diasuke, I., Tetsuya, S. and Noriko, K.: Overview of the NTCIR-8 Community QA Pilot Task (Part I): The Test Collection and the Task, *Proc. of NTCIR-8 Workshop Meeting*, Tokyo, National Institute of Infomatics, pp.421–432 (2010).
6) Fleiss, J., Cohen, J. and Everitt, B.: Large sample standard errors of kappa and weighted kappa, *Psychological Bulletin*, Vol.72, No.5, pp.323–327 (1969).
7) Jeon, J., Croft, W.B., Lee, J.H. and Park, S.: A framework to Predict the Quality of Answers with Non-Textual Features, *Proc. of SIGIR '06*, ACM, pp.228–235 (2006).
8) Jurczyk, P. and Agichtein, E.: Discovering AUTHORities in Question Answer Communities by Using Link Analysis, *Proc. of CIKM '07*, ACM, pp.919–922 (2007).
9) 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，*DBSJ Journal*, Vol.8, No.1, pp.89–94 (2009).
10) 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 Techonology*, Vol.60, No.4, pp.716–727 (2009).
11) 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).
12) OKWave: 教えて!goo. http://oshiete.goo.ne.jp/（参照 2010-09-30).
13) Rodrigues, E.M. and Milic-Frayling, N.: Socializing or Knowledge Sharing? Characering Social Intent in Community Question Answering, *Proc. of CIKM'09*, ACM (2009).
14) Suryanto, M.A., Lim, E.-P., Sun, A. and Chiang, R.H.: Quality-Aware Collaborative Queation Answering: Methods and Evaluation, *Proc. of WSDM'09*, ACM (2009).
15) Yahoo!JAPAN: Yahoo!知恵袋. http://chiebukuro.yahoo.co.jp/（参照 2010-09-30).
16) Yahoo!JAPAN: 「Yahoo!知恵袋」データの提供について. http://research.nii.ac.jp/tdc/chiebukuro.html （参照 2010-09-30).