RD-003

コミュニティQAにおける良質な回答の選定タスク: 評価方法に関する考察
Selecting Good Answers for Community QA: A Note on Evaluation Methods

酒井 哲也† 石川 大介* 栗山 和子‡ 関 洋平♯ 神門 典子*
Tetsuya Sakai, Daisuke Ishikawa, Kazuko Kuriyama, Yohei Seki, Noriko Kando
†Microsoft Research Asia *国立情報学研究所 ‡白百合女子大学 ♯筑波大学

1. はじめに

近年，Web 上における User-Generated Content (UGC) の爆発的な増大により，膨大な情報の中から有用なものを選出する技術の重要性が急速に高まっている．特に UGC におけるテキストデータは，20 世紀までに主な研究対象とされていた論文や新聞記事のような「きれいな」データとは違い，不均質な著者層のインタラクションを経て構築された不均質なコンテンツであり，客観性・信頼性・文章の完成度や簡潔さのばらつきが大きい．このため，hedge detection [4]，credibility analysis [2]，opinion mining や sentiment analysis [13] のように信頼性・有用性の高い情報を選出し体系化するための研究が活性化している．

「Yahoo!知恵袋」*や「教えて!goo」† といったコミュニティQA サイトに蓄積される質問および回答群もまた膨大かつ玉石混淆な UGC であり，このようなサイトでは有用な情報が死蔵されるのを防ぐ必要がある．そこで本研究では，コミュニティQA において，与えられた質問に対する良質な回答を同サイトに投稿された回答群から自動選定するタスクを扱う．この応用としては，コミュニティQA サイトのページ内で，最良と思われる回答のみ，もしくは回答を良質だと思われる順にランクづけして提示することが考えられる．また，複数のコミュニティQA サイトを横断的にクロールし，上記の回答選出技術を応用して有用な QA データのみを収録したデータベースを構築することも考えられる．さらに，このタスクで確立された技術をコミュニティQA 以外の UGC に適用していくことも考えられる．

本研究では，コミュニティQA における良質な回答の自動推定タスクを評価する方法について検討する．研究の素材としては，国立情報学研究所主催の国際評価ワークショップ NTCIR-8‡ における community QA pilot task [7, 17] で用いられた Yahoo!知恵袋データ，これを元に作成されたテストコレクション，そしてタスク参加チームによるシステム出力結果 ("runs" と呼ぶ [7, 16]) を用いる．

本論文の構成は以下のとおりである．2 章では，関連研究に触れ，本研究の位置づけを明らかにする．3 章では，NTCIR-8 community QA pilot task と，そこで構築されたテストコレクションおよび提出された runs について概説する．4 章では，良質な回答の自動推定タスクのために我々が考案した評価方法を説明する．5 章では，NTCIR-8 の runs に対して提案する評価方法を実際に適用し，その妥当性について考察する．最後に 6 章において，本研究のまとめと今後の課題について述べる．

2. 関連研究

コミュニティQA データから良質な質問および回答を選出する研究は近年盛んである [1, 5, 6, 10, 19, 21]．例えば，石川ら [5, 6] や栗山ら [10] は，Yahoo!知恵袋データを対象にシステムに「ベストアンサー」(BA) を推定させる可能性について検討している．ここで，BA とは，Yahoo!知恵袋サイトに質問を投稿した質問者自身が，投稿された回答の中から最も良いと思われるものを 1 件だけ選定したものである§．しかし，BA は質問者の主観的判断により 1 件だけ選出されたデータであり，これをそのまま良質な回答を選出するシステムの評価に用いると，信頼性および網羅性に問題が生じる可能性がある．ここで，信頼性とは個人による判断の正誤や偏りの度合いを意味する．例えば石川ら [6] は，人手でさえ BA の推定が容易ではないことを示している．また，網羅性は，一般には複数ある良質な回答をどれくらいカバーできているかを意味する．従来，コミュニティQA データから良質なコンテンツを選定する研究においては，このようなシステム評価の問題について充分な検討がなされてはいなかった．

一方，コミュニティQA 以外の分野では，従来より，システムの適切な評価方法に着眼した多くの研究がある．このうち最も歴史が古いのは情報検索の分野であり，ランクつき文書検索結果 (近年では例えば Web 検索

*http://chiebukuro.yahoo.co.jp/
†http://oshiete.goo.ne.jp/
‡http://research.nii.ac.jp/ntcir/ntcir-ws8/ws-ja.html

§Yahoo!知恵袋サイトにおける BA は投票により選定される場合もあるが，従来研究 [5, 6, 10]，本研究ともに，質問者が選定した BA のみを扱っている．

結果) のための評価方法および指標が多数提案されている．特に，21 世紀に入り，多値適合性 [16] を扱う評価指標 nDCG (normalised discounted cumulative gain)[8] や Q-measure[14] が提案され，NTCIR などで活用されている [7, 9, 18]．本研究でも，回答 ID のランクつきリスト評価のためにこれらの情報検索評価指標を応用している．従来の検索評価方法と異なるのは，多値適合性の定義を 4 人の判定者の見解に基づいて行っている点である．詳細については 4 章で述べる．

複数の判定者の利用により評価実験の信頼性を向上させるという方法論は，テキスト要約および自動質問応答の評価で用いられている pyramid method[11, 12] の考え方に基づく．これは直感的には，複数の判定者が作成した正解データの揺れを扱うために，多くの判定者の見解が一致した部分を pyramid の頂点，少数の判定者の見解を pyramid の裾と位置づけて評価を行うものである．ただ，テキスト要約における summary content unit[12] や質問応答における nugget[11] が任意の文字列を扱うのに対し，本研究における評価はあくまで予め定められた回答 ID を扱うものであり，この点では問題が比較的単純である．

3. NTCIR-8 におけるパイロットタスク

本研究では，Yahoo!知恵袋データ第 1 弾 [6] と，これをもとに我々が NTCIR-8 community QA pilot task において作成したテストコレクション，そして同タスクに提出された参加チームの runs を用いて実験および考察を行う．タスクの仕様については文献 [7] で詳述しているので，ここでは概要のみ述べる．

Yahoo!知恵袋データ第 1 弾は，2004 年 4 月から 2005 年 10 月にかけて投稿された 3,116,009 件の「解決済みの質問」[7] と，これらに対する BA¶およびその他の回答 10,361,777 件を含む．各質問には，「エンターテインメントと趣味」「インターネット、PC と家電」など Yahoo!知恵袋サイトが定義した質問カテゴリが付与されている．我々は，NTCIR-8 において，質問カテゴリ毎の質問数の分布を考慮し，上記質問セットから 14 カテゴリをカバーする 1,500 件の質問をサンプリングした．

上記 1,500 件の質問には，1,500 件の BA を含め，合計 7,443 件の回答が付与されている．NTCIR-8 では，このデータを図 1 のような形式で各参加チームに配布し，各質問に対する全ての回答を良質と思われる順にランキングした run ファイルを各チームに複数個提出してもらった．参加したのは 4 チームで，本論文ではこれらをそれぞれ A, B, L, M と呼ぶ‖．ただし，B は BASELINE を意味し，回答をランダムにソートした run (B-1), 長い順にソートした run (B-2), 投稿日時の新しい順にソートした run (B-3) の 3 つに対応する．これらを含め，NTCIR-8 では 13 件の runs を評価したが，本研究ではこのうち「異常値」といえる M-5 を除いた 12 件を分析対象とする**．なお，参加者には，提出された run ファイルが最良の回答を選出するタスクと回答を良質な順にランキングするタスクの 2 つの観点から評価される旨を予め通知した．

```
<QUESTION NO="5-6">
<Q_ID> 125513 </Q_ID>
<DATE> 2004-06-20 02:59:16 </DATE>
<TOPCATEGORY_NAME> インターネット、PC と家電 </TOPCATEGORY_NAME>
<TOPCATEGORY_LABEL> internet </TOPCATEGORY_LABEL>
<CATEGORY_NAME> 家電、AV 機器 </CATEGORY_NAME>
<CATEGORY_ID> 2078297425 </CATEGORY_ID>
<NUM_ANSWERS> 2 </NUM_ANSWERS>
<QUESTION_TEXT>
MP3 と VoiceRecorder の違いを教えて下さい。 私の使用したい用途は、
自分の演奏（ピアノ）を録音して、 CD-R に保存したい、と思ってい
ます。お勧めはどっちですか？
</QUESTION_TEXT>
<ANSWER NO="1">
<DATE> 2004-06-20 03:01:56 </DATE>
<A_ID> 619943 </A_ID>
<USER_ID> 330554 </USER_ID>
<ANSWER_TEXT>
まったくちがうものです。VoiceRecorder はマイクなどで録音するとき
に使い、MP3 は音楽ファイルの形式です。
</ANSWER_TEXT>
</ANSWER>
<ANSWER NO="2">
<DATE> 2004-06-20 03:12:20 </DATE>
<A_ID> 620041 </A_ID>
<USER_ID> 624467 </USER_ID>
<ANSWER_TEXT>
ボイスレコーダーは音声を録音するための機械。 MP3 は音声をデジタル
データとして圧縮するときの形式
（MPEG Audio Layer-III)。
</ANSWER_TEXT>
</ANSWER>
</QUESTION>
```

図 1: NTCIR-8 community QA pilot task の質問と回答群の例.

前述の通り，コミュニティ QA における質問および回答は質が高いものばかりではない．さらに，BA には前述の信頼性・網羅性の問題が付随する．そこで我々は，参加チームに対する評価データ公開に先立ち，上記 1,500 件の質問および対応する回答群を対象に，4 名の判定者 (理系男子，文系男子，理系女子，文系女子の大学生) による質の判定を行った．具体的には，各質問の質を A, B の 2 段階で，各回答の質を A, B, C の 3 段階で絶対評価してもらった．NTCIR-8 では，このうち質問に対する

¶BA が欠落している質問が 1 件だけ存在する．

‖正式なチーム名は overview 論文 [7] を参照すれば容易に復元できる．

**今回のタスク設計では，学習データと評価データの分離をしなかったため，M-5 には評価データに付与された BA を直接参照した学習アルゴリズムが適用され，BA よる評価値が異常に高くなってしまった [7].

判定結果を用いて，4 名全員が A 判定をした 1,429 件の質問に絞ったシステム評価も併せて行ったが，その結果は全 1,500 件による評価結果とほぼ変わらなかった [7]. そこで本論文では，全 1,500 件の質問セットを用い，回答の判定結果のほうの利用方法に着目したシステム評価実験を行う．判定者に与えた回答の判定基準は以下の通りである．なお，A, B, C の件数に制限は設けなかった．

A 質問内容を十分満たす答えが含まれている

B 質問内容に部分的に適合している，もしくは部分的に不適合

C 質問内容と全く関係がない

判定結果の一致度については文献 [7] を参照されたい．

4. 評価方法

4.1 NTCIR-8 で用いた評価方法

NTCIR-8 では，BA に基づく評価と，前述の 4 名の判定者による回答判定結果に基づく評価を併用した．BA に基づく評価は，各質問につき，システムが第 1 位に出力した回答が BA である場合は 1, そうでない場合は 0 とする単純な評価指標 “hit at rank 1 (BA-Hit@1)” を用いた．BA-Hit@1 の質問セットに関する平均は，システムが何割の質問に対して BA を的確に推定できるかを意味する．

一方，4 名の判定者の回答判定結果に基づく評価は，前述の pyramid method [11, 12] を参考に以下のように行った．まず，表 1(a) のように，4 名の判定結果を「適合パターン」で表し，A もしくは B 判定が多い順に並べた．(異なる並べ方について次節で述べる．) ここで，例えば AAB という適合パターンは，2 名が A 判定，1 名が B 判定，残りの 1 名が C 判定を下したことを表す．次に，このように配置した適合パターンを，表 1(c) のように情報検索評価における「適合レベル」[16] に対応させた．ここで，$L3, L2, L1$ はそれぞれ高適合，適合，部分適合を意味する．様々な対応付けが考えられるが，NTCIR-8 では，$L3, L2, L1$ の文書数が同程度になるように区分けを行った．このようにして得た多値適合性データを GA (good answers) データと呼ぶことにする．

なお，原理的には，例えば適合パターン AAAA と AAAB に異なる適合レベルを付与するなど，より細分化された適合レベルを設けることも可能である．しかし今回は，TREC robust track [20] や NTCIR IR4QA [18] の文書検索タスクにおける適合レベルが (不適合も含めて)3 レベル，NTCIR CLIR [9] の文書タスクにおける適合レベルが 4 レベルであったことを参考に，$L3, L2, L1, L0$ の 4 レベルを考えることにした．

GA データの $L3, L2, L1$ を一律に二値適合性データとして扱えば，BA の場合と同様に hit at rank 1 が計算できる．これを GA-Hit@1 と表記する．しかし，この指標は部分適合回答でも正解と見なすものであり，全正解が $7443 - 50 = 7393$ 件もあることを考えると (表 1 参照)，非常に甘い指標である．本当に質の高い回答を選出するシステムの評価には不十分であると考えられる．そこで，GA データの多値適合性を活用した以下の 3 つの評価指標を併せて用いる．

まず，$L3, L2, L1$ の回答に対する利得をそれぞれ 3, 2, 1 と定義する ($L0$ の回答に対する利得は 0 とする)[16]. これは例えば，$L3$ の回答の価値が $L1$ の回答の価値の 3 倍であると見なすことを意味する．ある質問に対してシステムが出力した回答リストの第 r 位における利得を $g(r)$ で表し，同様に，理想的な回答リストの第 r 位における利得を $g^*(r)$ で表す．ここで，理想的な回答リストとは，$L3, L2, L1$ の回答をこの順番で全て列挙したものをいう．さらに，$cg(r) = \sum_{i=1}^{r} g(i), cg^*(r) = \sum_{i=1}^{r} g^*(i), C(r) = \sum_{i=1}^{r} I(i)$ と定義する．ここで $I(r)$ は第 r 位の回答が正解 ($L3, L2, L1$) であるか否かを表すフラグである††．このとき，検索評価指標 nDCG[8] および Q-measure[14] を応用した指標が以下のように定義できる．

$$GA\text{-}nG@1 = g(1)/g^*(1) \tag{1}$$

$$GA\text{-}nDCG = \frac{\sum_{r=1}^{l} g(r)/\log(r+1)}{\sum_{r=1}^{l} g^*(r)/\log(r+1)} \tag{2}$$

$$GA\text{-}Q = \frac{1}{R}\sum_{r} I(r)\frac{C(r)+\beta cg(r)}{r+\beta cg^*(r)} \tag{3}$$

ここで，l は回答の順位の cut-off 値，β はユーザの「忍耐力」を表す定数である [15].

GA-nG@1 は，第 1 位における nDCG (normalised discounted cumulative gain)[8] と等価である．例えば $L3$ の回答をもつ質問に対してシステムが第 1 位に $L1$ の回答を出力した場合，$GA\text{-}nG@1 = 1/3$ となる．(一方，$GA\text{-}Hit@1 = 1$ となる．また，$L3, L2, L1$ に対する利得を全て同じ値に設定すれば，GA-nG@1 は GA-Hit@1 に帰着する．) 今回の質問セットは 2〜19 件の回答を有していることから，GA-nDCG の cut-off 値は $l = 20$ とした．これは回答が 2 件の質問に対しては第 2 位における nDCG を，回答が 19 件の質問に対しては第 19 位にお

††従って $GA\text{-}Hit@1 = I(1)$ と定義できる．

表 1: GA データ作成のための適合パターンから適合レベルへの対応表.

(a) パターン	(b) 回答数	(c) レベル	(d) 回答数
AAAA	1301	$L3$	2806
AAAB	1505		
AABB	1525	$L2$	2910
ABBB	1385		
BBBB	1241	$L1$	1677
AAA	2		
AAB	14		
ABB	76		
BBB	231		
AA	1		
AB	7		
BB	105		
A	1	$L0$	50
B	32		
(C's only)	17		
total	7443	total	7443

表 2: GA_A データ作成のための適合パターンから適合レベルへの対応表. GA と比べた適合レベルの上下を ⇑ (2段階) および ↑ / ↓ (1 段階) で示す.

(a) パターン	(b) 回答数	(c) レベル	(d) 回答数
AAAA	1301	$L3$	2808
AAAB	1505		
AAA⇑	2		
AABB	1525	$L2$	1540
AAB	14		
AA↑	1		
ABBB↓	1385	$L1$	3046
ABB	76		
AB	7		
A↑	1		
BBBB	1241		
BBB	231		
BB	105		
B	32	$L0$	49
(C's only)	17		
total	7443	total	7443

ける nDCG を計算することと等価である．また，GA-Q のパラメタ β は文献 [15] に従い 1 とする．

なお，我々は，NTCIR-8 で用いた 1,500 件の質問における BA と GA の関係を調査した．その結果，970 件について BA に $L3$ が，399 件について BA に $L2$ が，130 件について BA に $L1$ が，残る 1 件については BA に $L0$ が付与されていた‡‡．従って，NTCIR-8 の BA1,500 件に関しては，信頼性の問題はそれほど深刻ではないと思われる．また，BA をシステムの出力と見なした場合の GA-Hit@1 は定義により $1499/1500 = 0.9993$ となる．

4.2 対応表を差し替えた評価

表 1 は，個々の判定が A であるか B であるかよりも，A もしくは B 判定の個数のほうが重要であるという考え方に基づいている．このため，例えば，適合パターン ABBB のほうが AAA よりも高い適合レベルに写像されている．このような対応付けの恣意性がシステム評価結果に与える影響を調べるため，我々は表 2 に示す第二の対応表に基づく評価実験も行った．これは，「A もしくは B の個数」ではなく，「A のみの個数」に応じて適合パターンを配置したもので，表 1 に比べ適合レベルが上下したものについては矢印を表示している．大きく変わったのは適合パターン ABBB の回答 1,385 件が $L2$ から $L1$ に降格となった点である．なお，$L1$ へ写像される回答の件数が多くなっているのは，今回の質問セットにおいて適合パターン BB を不正解 ($L0$) 扱いにしてしまうと，正解をひとつも含まない質問が出てきてしまうためである．表 2 を元に作成された多値適合性データを GA_A と呼ぶこととし，対応する評価指標を GA_A-nG@1 のように表記する．なお，表 1，2 の比較から，$L3, L2, L1$ を全て正解とみなす評価指標 Hit@1 を用いた場合，GA_A による結果は GA によるものとほぼ変わらないことがわかる．すなわち，多値適合性に基づく指標 GA_A-nG@1, GA_A-nDCG, GA_A-Q のみ議論すればよい．

‡‡$L0$ が付与された BA の内容は「あんたウザいよ。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。」であった．

表 3: BA に対する個々の判定者の判定結果.

	J1	J2	J3	J4
A	1240	843	970	1110
B	246	649	515	379
C	14	8	15	11

4.3 利得の値を差し替えた評価

本研究では，GA を利用する際の利得の与え方がシステム評価に与える影響についても考慮する．NTCIR-8 公式結果は，$L3, L2, L1$ の回答にそれぞれ 3, 2, 1 の利得を与えることにより算出したが，例えばこれを 2, 2, 1 に変えることは，表 1 において AAAA と AAAB を $L3$ から $L2$ に降格することと等価である．本研究では，比較的極端な利得の設定例として，GA に基づく $L3, L2, L1$ の回答にそれぞれ 10, 5, 1 の利得を与えた場合のシステム評価結果を検証する．また，この場合の評価指標を $GA_{10:5:1}$-nG@1 のように表記する．この場合も，多値適合性に基づく指標のみ議論すればよい．

4.4 判定者 1 名に基づく評価

上記の各種評価実験では，評価の信頼性向上のため，4 名の判定者による判定結果を利用することを前提にしている．このように複数判定者を用いること自体の効果を検証するため，我々は 1 名の判定者の結果に基づく評価実験も行った．表 3 に，判定者 J1, J2, J3, J4 が 1,500 件の BA に対して下した判定結果を示す．今回は A, B, C 判定を単純に適合レベル $L2, L1, L0$ にそれぞれ対応させることにしたが，この結果，J2 と J4 のデータについては，正解 ($L2, L1$) が 1 件もない質問が生じた．そこで本研究では，J1 と J3 の判定結果からそれぞれ作成

表 4: GA による NTCIR-8 community QA task 公式結果 (質問 1,500 件の平均)．各指標により runs をソートし，隣接する対にのみ両側符号検定を実施した．その差が統計的に有意である場合，上位の値に $**(\alpha = 0.01)$ もしくは $*(\alpha = 0.05)$ を付与した．ただし，検定の結果が順位と食い違う場合には $\dagger(\alpha = 0.05)$ を付与した．

run	BA-Hit@1	run	GA-Hit@1	run	GA-nG@1	run	GA-nDCG	run	GA-Q
M-2	0.4980	M-4	0.9973	M-2	0.9211	M-2	0.9747	M-2	0.9690†
M-1	0.4980	M-2	0.9967	M-1	0.9203	M-4	0.9745	A-2	0.9689
M-4	0.4847	M-1	0.9967	M-4	0.9202	A-2	0.9742**	M-4	0.9688**
B-2	0.4847	M-3	0.9960	B-2	0.9170	M-1	0.9741*	M-1	0.9682*
A-2	0.4840	B-2	0.9953	A-2	0.9166	B-2	0.9735*	B-2	0.9680
M-3	0.4813	A-2	0.9953	A-1	0.9140**	A-1	0.9734**	A-1	0.9680**
A-1	0.4813**	B-3	0.9940	M-3	0.8956**	M-3	0.9679**	M-3	0.9609**
B-3	0.3820**	A-1	0.9940	B-3	0.8213**	B-3	0.9460**	B-3	0.9359**
B-1	0.2713**	B-1	0.9920	B-1	0.7751**	B-1	0.9311**	B-1	0.9169
L-3	0.1767	L-3	0.9887	L-3	0.6883	L-2	0.9191**	L-2	0.9081**
L-2	0.1767	L-2	0.9887	L-2	0.6883	L-3	0.9142**	L-3	0.9002**
L-1	0.1767	L-1	0.9887	L-1	0.6883	L-1	0.9096	L-1	0.8927

した正解データによる評価を行う．これらをもとに算出した評価指標を J1-Hit@1, J3-Hit@1 のように表記する．$L2, L1$ の利得はそれぞれ 2, 1 とした．

BA1,500 件を信頼性があまり高くない網羅性の低いデータとみなし，7,393 件の正解をもつ GA(表 1 参照) を信頼性・網羅性ともに高いデータとみなせば，それぞれ 7,299 件，7,156 件の正解をもつ J1, J3 データは，信頼性は BA と同様にあまり高くないが，網羅性はある程度高いデータと位置づけることができる．

5. 評価結果と考察

5.1 GA による NTCIR-8 公式結果

表 4 に，BA および GA に基づく NTCIR-8 community QA pilot task の評価結果を示す．各カラム内で，runs は評価指標の質問 1,500 件に関する平均値によりソートされている．同程度の成績の runs の「クラスタ」を大まかに把握するため，各カラム内で隣り合う runs について両側符号検定を実施し，その結果を $**(\alpha = 0.01)$ もしくは $*(\alpha = 0.05)$ で示している．例えば，BA-Hit@1 による評価では，A-1 と B-3 の差，B-3 と B-1 の差，B-1 と L-3 の差は統計的に有意である ($\alpha = 0.01$)．(もちろん，これらの統計的検定結果の間に推移率は成り立たない．) なお，GA-Q のカラムにある † は，平均値の上では M-2 のほうが A-2 よりも僅差で高いが，A-2 は M-2 を 327 件の質問について上回っており，274 件について下回っているため，検定の結果が順位とは逆転していることを示している．

表 5 はシステムのランキング間の類似性を Kendall 順位相関係数 [18] により数値化したものである．ただし，表 4 における BA-Hit@1 のカラムにおいて，M-3 と A-1 の成績は同じであり，BA-Hit@1 と GA-nG@1 の順位相関は本質的には 1 である．

これらの結果より，以下の知見が得られる．

表 5: 異なる指標によるランキング間の Kendall 順位相関 (GA).

	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q
BA-Hit@1	0.848	0.970	0.848	0.818
GA-Hit@1	1	0.818	0.758	0.727
GA-nG@1	-	1	0.879	0.848
GA-nDCG	-	-	1	0.970

(1) 部分適合正解でも正解と見なす GA-Hit@1 は，システム間の差異をうまく検出できず，評価指標としてあまり役に立たない．表 4 では統計的有意差がひとつも得られておらず，また NTCIR-8 における質問カテゴリ別評価では，全ての runs の値が 1 となってしまうカテゴリが複数あった [7].

(2) GA-nG@1 は，BA-Hit@1 とランキングが一致しており，かつ BA では検出できなかった差異を見出している．(BA-Hit@1 によれば A-1 と M-3 は同じ成績であるが，GA-nG@1 によれば両者の差は統計的に有意である．A-1 は M-3 に対して 200 件の質問について上回っており，138 件の質問について下回っている．) これは，BA の信頼性と網羅性を補う GA による評価により，新たな発見が得られることを示唆する．

(3) GA-nDCG, GA-Q によるランキングは，互いに非常に似ているが，他の評価指標によるランキングとは若干異なる．さらに，ランキング全体を見渡して評価するこれらの指標のほうが，他の評価指標よりも多くの統計的有意差を検出できている．従って，最適な回答を 1 つ出力させるタスクと回答を良質な順にランク付けさせるタスクでは，各々に適した評価尺度を用いるべきであろう．

知見 (2) は，BA の信頼性と網羅性を補うアプローチの有効性を示唆してはいるが，必ずしも我々の用いた GA 構築方法の最適性を保証するものではない．そこで，以

表 6: GA_A による NTCIR-8 community QA task 評価結果 (質問 1,500 件の平均)．各指標により runs をソートし，隣接する対にのみ両側符号検定を実施した．その差が統計的に有意である場合，上位の値に $**(\alpha = 0.01)$ もしくは $*(\alpha = 0.05)$ を付与した．また，GA による結果と比較したときの順位の変化を矢印で示している．

run	GA_A-nG@1	run	GA_A-nDCG	run	GA_A-Q
M-2	0.8998	M-2	0.9676	M-2	0.9637
M-1	0.8984	M-4	0.9672	A-2	0.9636
M-4	0.8979	B-2 ↑ 2	0.9669	B-2 ↑ 2	0.9634
B-2	0.8963	A-2 ↓ 1	0.9669*	M-4 ↓ 1	0.9633*
A-2	0.8941	M-1 ↓ 1	0.9667	M-1 ↓ 1	0.9626
A-1	0.8932**	A-1	0.9663**	A-1	0.9626**
M-3	0.8684**	M-3	0.9588**	M-3	0.9540**
B-3	0.7834**	B-3	0.9326**	B-3	0.9257**
B-1	0.7233**	B-1	0.9123**	B-1	0.9011
L-3	0.6229	L-2	0.8984**	L-2	0.8919**
L-2	0.6229	L-3	0.8917**	L-3	0.8816**
L-1	0.6229	L-1	0.8859	L-1	0.8724

表 7: $GA_{10:5:1}$ による NTCIR-8 community QA task 評価結果 (質問 1,500 件の平均)．見方は 6 と同様である．

run	$GA_{10:5:1}$-nG@1	run	$GA_{10:5:1}$-nDCG	run	$GA_{10:5:1}$-Q
M-2	0.8861	M-2	0.9586	A-2 ↑ 1	0.9460*
M-1	0.8849	M-4	0.9582	M-2 ↓ 1	0.9455
M-4	0.8846	A-2	0.9580**	M-4	0.9452
B-2	0.8808	M-1	0.9577*	B-2 ↑ 1	0.9444
A-2	0.8802	B-2	0.9569	A-1 ↑ 1	0.9444
A-1	0.8771**	A-1	0.9568**	M-1 ↓ 2	0.9442**
M-3	0.8498**	M-3	0.9476**	M-3	0.9316**
B-3	0.7452**	B-3	0.9111**	B-3	0.8866**
B-1	0.6813**	B-1	0.8871**	B-1	0.8528
L-3	0.5609	L-2	0.8688**	L-2	0.8414**
L-2	0.5609	L-3	0.8606**	L-3	0.8260**
L-1	0.5609	L-1	0.8527	L-1	0.8108

降，適合パターンから適合レベルへの対応の決め方と，各適合レベルに与える利得の値が評価結果に与える影響について検証する．最後に，信頼性向上のために複数の判定者を用いること自体の効果を検証する．

5.2 対応表の影響

表 6 に，GA_A に基づく NTCIR-8 community QA pilot task の評価結果を示す．前述の通り，この実験は，適合パターンから適合レベルへの対応付けの評価への影響を検証するため，表 1 を表 2 で差し替えて行ったものである．この差し替えにより影響を受ける多値適合性に基づく評価指標の結果のみ示しており，表 4 と比べたときの順位の変化を矢印で表している．これらの結果より，以下の知見が得られる．

(4) GA_A-nG@1 によるランキングは，表 4GA-nG@1 によるものと同一である．これは，評価指標 nG@1 が，適合パターンから適合レベルへの対応のさせ方に影響を受けにくいことを示唆している．

(5) 一方， GA_A-nDCG, GA_A-Q は対応表の差し替えの影響を若干受けている．(表 4 において第 5 位であった B-2 が第 3 位に浮上している．) 従って，回答をランク付けするタスクの評価においては，対応表の選択が評価結果に影響する可能性があることに注意する必要がある．

5.3 利得の影響

表 7 に，$GA_{10:5:1}$ に基づく NTCIR-8 community QA pilot task の評価結果を示す．前述の通り，この実験は，各適合レベルに対する利得の与え方の影響を検証するため，$L3, L2, L1$ にそれぞれ 3, 2, 1 点与える代わりに 10, 5, 1 点与えたものである．これにより影響を受ける多値適合性に基づく評価指標の結果のみ示しており，表 4 と比べたときの順位の変化を矢印で表している．

これらの結果より，以下の知見が得られる．

(6) $GA_{10:5:1}$-nG@1, $GA_{10:5:1}$-nDCG によるランキングは，表 4 の GA-nG@1 によるものと同一である．これは，評価指標 nG@1, nDCG が利得の与え方に影響を受けにくいことを示唆しており，文書検索タスクにおける従来研究の結果とも合致する [14].

(7) 一方， $GA_{10:5:1}$-Q は利得の値の差し替えの影響を若干受けている．(僅差ではあるが，順位を上げたものが 3 件，下げたものが 2 件．) 従って，回答をランク付けするタスクの評価においては，nDCG よりも Q のほうが利得の与え方に左右されやすい可能性がある．(ただし，この利得の与え方は比較的極端である．)

5.4 複数判定者利用の効果

表 8 および 9 に，判定者 J1, J3 のデータに基づく NTCIR-8 community QA pilot task の評価結果を示す．(4.4 節で述べた理由により，J2, J4 のデータは用いなかった．) 前述の通り，この実験は，複数の判定者を利用することの効果を検証するためのものである．4 名の判定者を利用した表 4 と比べたときの順位の変化を矢印で表している．

表 10 に，評価指標を固定し，GA, J1, J3 に基づくランキングをそれぞれ比較した場合の Kendall 順位相関係数を示す．

これらの結果により，以下の知見が得られる．

(8) 各評価指標について，J1, J3 に基づく結果を表 4 の GA に基づく結果と比較すると，ランキングはかなり変化している．また，表 10 より，特に nG@1, nDCG, Q については，J1 と J3 の順位相関よりも，

表 8: J1 による NTCIR-8 community QA task 評価結果 (質問 1,500 件の平均). 見方は 6 と同様である.

run	J1-Hit@1	run	J1-nG@1	run	J1-nDCG	run	J1-Q
M-3 ↑ 3	0.9893	M-2	0.9400	M-2	0.9791	A-2 ↑ 1	0.9762
M-4 ↓ 1	0.9880	M-4 ↑ 1	0.9393	A-2 ↑ 1	0.9788	M-2 ↓ 1	0.9760
M-1	0.9860	M-1 ↓ 1	0.9393	M-4 ↓ 1	0.9787	M-4	0.9755
B-2 ↑ 1	0.9860	A-2 ↑ 1	0.9387	M-1	0.9787	M-1	0.9755
A-2 ↑ 1	0.9853	A-1 ↑ 1	0.9357	A-1 ↑ 1	0.9781	A-1 ↑ 1	0.9754
M-2 ↓ 4	0.9847	B-2 ↓ 2	0.9333**	B-2 ↓ 1	0.9772**	B-2 ↓ 1	0.9743**
A-1 ↑ 1	0.9847	M-3	0.9170**	M-3	0.9730**	M-3	0.9698**
B-3 ↓ 1	0.9840	B-3	0.8710**	B-3	0.9591**	B-3	0.9542**
B-1	0.9800	B-1	0.8400**	B-1	0.9494**	B-1	0.9432**
L-3	0.9760	L-3	0.7723	L-2	0.9390**	L-2	0.9345**
L-2	0.9760	L-2	0.7723	L-3	0.9359**	L-3	0.9301**
L-1	0.9760	L-1	0.7723	L-1	0.9325	L-1	0.9252

表 9: J3 による NTCIR-8 community QA task 評価結果 (質問 1,500 件の平均). 見方は 6 と同様である.

run	J3-Hit@1	run	J3-nG@1	run	J3-nDCG	run	J3-Q
M-3 ↑ 3	0.9920	M-4 ↑ 2	0.8887	M-4 ↑ 1	0.9675	M-4 ↑ 2	0.9656
M-2	0.9920	M-2 ↓ 1	0.8877*	M-2 ↓ 1	0.9672	M-2 ↓ 1	0.9650*
M-4 ↓ 2	0.9913	M-1 ↓ 1	0.8847	M-1 ↑ 1	0.9662*	B-2 ↑ 2	0.9643*
M-1 ↓ 1	0.9913	B-2	0.8840	B-2 ↑ 1	0.9661	M-1	0.9639*
A-1 ↑ 3	0.9893	A-1 ↑ 1	0.8797	A-2 ↓ 2	0.9651	A-2 ↓ 3	0.9637
A-2	0.9887	A-2 ↓ 1	0.8787**	A-1	0.9650**	A-1	0.9634**
B-2 ↓ 2	0.9880**	M-3	0.8587**	M-3	0.9590**	M-3	0.9569**
B-3 ↓ 1	0.9740	B-3	0.7817**	B-3	0.9362**	B-3	0.9338**
B-1	0.9640**	B-1	0.7330**	B-1	0.9197	B-1	0.9150
L-3	0.9393	L-3	0.6327	L-2	0.9056**	L-2	0.9059**
L-2	0.9393	L-2	0.6327	L-3	0.8996**	L-3	0.8974**
L-1	0.9393	L-1	0.6327	L-1	0.8948	L-1	0.8903

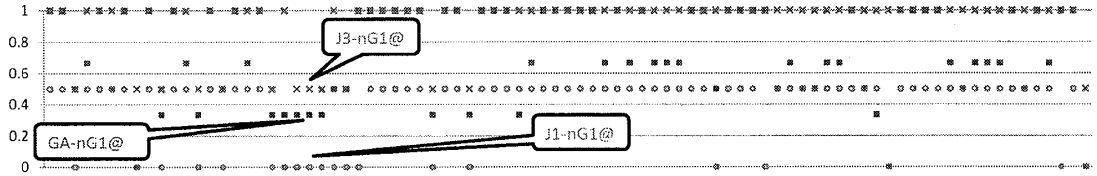

図 2: *J3-nG*@1 > *J1-nG*@1 となった質問 85 件に関する GA-nG@1, J1-nG@1, J3-nG@1 の比較.

GA と J1 の順位相関および GA と J3 の順位相関のほうが高い. (例えば Q については, それぞれ 0.788, 0.939, 0.848 である.) これらのことから, 複数判定者の利用により, 複数の異なる個人の見解を反映した総括的な評価が可能になることがわかる.

図 2 に, GA-nG@1 が複数の個人の異なる見解を反映した総括的な評価を行う様子を一部視覚化した. ここで, 縦軸は最も成績の良かった run である M-2 の GA-nG@1(四角), J1-nG@1(×),J3-nG@1(○) の値, 横軸は 1,500 件の質問のうち J3-nG@1 > J1-nG@1 となった質問 85 件の ID を表している. (一方, J3-nG@1 < J1-nG@1 となった質問は 243 件もあり, 本論文中での視覚化には適さなかった.) J1 と J3 の評価が食い違った場合でも, GA は (J2 と J4 の結果も考慮に入れた上で) その間をとった評価を下す様子がわかる.

6. まとめと今後の課題

本研究では, Yahoo!知恵袋の QA データとこれに基づく NTCIR-8 community QA pilot task のテストコレクションおよび runs を題材に, 良質な回答を自動選定するタスクのための評価方法の検討を行った. 質問者自身が回答群から選んだ BA データの信頼性と網羅性を補うために, 4 名の判定者の判定結果に基づく多値適合性データ GA を構築し, これに基づく複数の評価指標によるシステム評価を行った. また, この方法における適合パターンから適合レベルへの対応付けと, 各適合レベルに対する利得の与え方が評価結果に与える影響について考察を行った. さらに, GA による評価結果を, 1 人の判定者に基づく評価結果と比較した. 得られた主な知見は以下のとおりである.

表 10: GA, J1, J3 に基づくランキング間の Kendall 順位相関 (指標を固定).

	GA vs. J1	GA vs. J3	J1 vs. J3
Hit@1	0.788	0.758	0.788
nG@1	0.909	0.909	0.879
nDCG	0.939	0.909	0.848
Q	0.939	0.848	0.788

- GA の多値適合性を活用した評価指標の利用により, BA では得られないシステム間の差異を見出せる可能性がある.

- GA-nG@1 による回答をひとつ出力するタスクの評価は，適合パターンから適合レベルへの対応付けや，各適合レベルへの利得の与え方に大きな影響を受けないものと思われる．
- 一方，回答をランク付けするタスクの評価は，適合パターンから適合レベルへの対応付けや，各適合レベルへの利得の与え方に若干の影響を受ける可能性があるので，評価結果の解釈には注意が必要である．
- GA により，複数の個人の異なる見解を反映した総括的な評価結果を得ることができる．

実用上の指針としては，本研究で扱った評価方法の範囲では，回答をひとつ出力するタスクの評価には GA-nG@1，回答をランク付けするタスクの評価には GA-nDCG を用いれば，対応表や利得の選択にあまり依存しない結果が得られると思われる．

本研究は，コミュニティQA を対象とした良質な回答を選定するタスクの評価のための第一歩であり，多くの課題を抱えている．まず，BA では得られない知見が GA により得られる感触はつかんだものの，そのコストパフォーマンスが明らかではない．即ち，質問者が既に提供してくれている BA に加えて，何人の判定者にどれだけの労力を払ってもらえれば新たな知見が得られるか，目下のところ明らかではない．さらに，今回の知見が他の質問セット，さらには Yahoo!知恵袋以外のコミュニティQA データに当てはまるか否かの検証も必要である．また，人間によるシステムの相対評価と評価指標との関係も明らかにしたい．

謝辞

本研究の実施にあたって，ヤフー株式会社が国立情報学研究所に提供した「Yahoo! 知恵袋データ (第 1 弾)」を利用いたしました．

参考文献

[1] Agichtein, E., Liu, Y. and Bian, J.: Modeling Information-Seeker Satisfaction in Community Question Answering, *ACM TKDD*, Volume 3, Issue 2, Article No.10 (2009).

[2] Akamine, S. *et al.*: WISDOM: A Web Information Credibility Analysis System, *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pp. 1-4 (2009).

[3] Cong, G. *et al.*: Finding Question-Answer Pairs from Online Forums, *ACM SIGIR 2008 Proceedings*, pp. 467-474 (2008).

[4] Ganter, V. and Strube, M.: Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia and Shallow Linguistic Features, *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 173-176 (2009).

[5] 石川, 栗山, 関, 神門: Q&A サイトにおけるベストアンサー推定可能性の検証, 情報処理学会研究報告 2009-FI-97 (2009).

[6] 石川, 栗山, 酒井, 関, 神門: Q&A サイトにおけるベストアンサー推定の分析とその機械学習への応用, 情報知識学会第 18 回年次大会 (2010).

[7] Ishikawa, D., Sakai, T. and Kando, N.: Overview of the NTCIR-8 Community QA Pilot Task (Part I): The Test Collection and the Task, *NTCIR-8 Online Proceedings* (2010).

[8] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM TOIS*, Vol. 20, No. 4, pp. 422-446 (2002).

[9] Kishida, K. *et al.*: Overview of CLIR Task at the Sixth NTCIR Workshop, *NTCIR-7 Proceedings* (2007).

[10] 栗山, 神門: 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-, 情報処理学会研究報告 2009-FI-97 (2009).

[11] Lin, J. and Demner-Fushman, D.: Will Pyramids Built of Nuggets Topple Over? *HLT/NAACL 2006 Proceedings*, pp. 383-390 (2006).

[12] Nenkova, A., Passonneau, R. and McKeown, K.: The Pyramid Method: Incorporating Human Content Selection Variation in Sumarization Evaluation, *ACM Transactions on Speech and Language Processing*, Volume 4, Number 2, Article 4 (2007).

[13] Pang, B. and Lee, L.: Opinion Mining and Sentiment Analysis, *Foundations and Trends in Information Retrieval*, 2(1-2), pp. 1-135 (2008).

[14] Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, Volume 43, Issue 2, pp.531-548 (2007).

[15] Sakai, T.: On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, *Proceedings of the First Workshop on Evaluating Information Access (EVIA 2007)*, pp.32-43 (2007).

[16] 酒井: チュートリアル：情報検索テストコレクションと評価指標, 情報処理学会研究報告 2008-FI-89 / 2008-NL-183, pp.1-8 (2008).

[17] Sakai, T., Ishikawa, D. and Kando, N.: Overview of the NTCIR8 Community QA Pilot Task (Part II): System Evaluation, *NTCIR-8 Online Proceedings* (2010).

[18] Sakai, T. *et al.* Overview of NTCIR-8 ACLIA IR4QA, *NTCIR-8 Online Proceedings* (2010).

[19] Sun, K. *et al.*: Learning to Recommend Questions based on User Ratings, *ACM CIKM 2009 Proceedings*, (2009).

[20] Voorhees E. M.: Overview of the TREC 2004 Robust Retrieval Track, *TREC 2004 Proceedings*, (2005).

[21] Wang, X.-J. *et al.*: Ranking Community Answers by Modeling Question-Answer Relationships via Analogical Reasoning, *ACM SIGIR 2009 Proceedings*, pp. 179-186 (2009).